Cyrill Stachniss
Kerstin Schill
David Uttal (Eds.)

LNAI 7463

Spatial Cognition VIII

International Conference, Spatial Cognition 2012
Kloster Seeon, Germany, August/September 2012
Proceedings

Springer

Cyrill Stachniss Kerstin Schill
David Uttal (Eds.)

Spatial Cognition VIII

International Conference, Spatial Cognition 2012
Kloster Seeon, Germany, August 31 – September 3, 2012
Proceedings

Springer

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editors

Cyrill Stachniss
Albert-Ludwigs-University
Institute of Computer Science
Georges-Koehler-Allee 79, 79110 Freiburg, Germany
E-mail: stachnis@informatik.uni-freiburg.de

Kerstin Schill
University of Bremen
Cognitive Neuroinformatics
Enrique-Schmidt-Str. 5, 28359 Bremen, Germany
E-mail: kschill@informatik.uni-bremen.de

David Uttal
Northwestern University
Department of Psychology and
School of Education and Social Policy
Sheridan Road 2029, Evanston, IL 60208-2710, USA
E-mail: duttal@northwestern.edu

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-32731-5 e-ISBN 978-3-642-32732-2
DOI 10.1007/978-3-642-32732-2
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012944707

CR Subject Classification (1998): I.2, H.2.8, I.2.10, H.3-5, I.4

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains the papers presented at the Spatial Cognition 2012 (SC 2012) conference held in Kloster Seeon, Germany, from August 31 to September 3. The aim of SC 2012 was to bring together researchers interested in spatial cognition from different fields such as cognitive science, psychology as well as computer science and robotics. This goal is also reflected by the diversity of papers in this volume.

We received 59 submissions and selected 31 papers for oral presentation at SC 2012. The Program Chairs made the acceptance decisions, which were based on the recommendation of at least one area chair and two to three reviewers. Every submission was discussed at the Program Committee meeting in mid-April 2012. In addition to the oral presentations, we had a poster session in which researchers could present preliminary results and have informal discussions. We furthermore had three invited talks given by Bastian Leibe from RWTH Aachen, David Lubinski from Vanderbilt University, and J. Kevin O'Regan from the Université Paris Descartes.

We as the organizers would especially like to thank a number of people that supported SC 2012, especially Christian Freksa, Thomas Barkowski, Carsten Rachuy, Ingrid Friedrichs from the University of Bremen, and last but not least, Gian Diego Tipaldi from the University of Freiburg. In addition to that, we would like to thank the DFG and the SFB/TR-8 as well as the Spatial Intelligence and Learning Center (SILC) for supporting SC 2012 financially. Their support is gratefully acknowledged.

August 2012

Cyrill Stachniss
Kerstin Schill
David Uttal

Organization

SC 2012 was organized by the Computer Science Departments of the University of Freiburg and the University of Bremen in cooperation with the DFG through the SFB/TR8 and the NSF through the Spatial Intelligence and Learning Center (SILC).

Program Chairs

Cyrill Stachniss	University of Freiburg, Germany
Kerstin Schill	University of Bremen, Germany
David Uttal	Northwestern University, USA

Local Arrangments Chair

Carsten Rachuy	University of Bremen, Germany

Publication Chair

Gian Diego Tipaldi	University of Freiburg, Germany

Publicity Chairs

Carsten Rachuy	University of Bremen, Germany
Cyrill Stachniss	University of Freiburg, Germany

Workshop Chair

Udo Frese	University of Bremen, Germany

Area Chairs

Achim Lilienthal	Örebro University, Sweden
Alexander Klippel	Pennsylvania State University, USA
Alycia Hund	Illinois State University, USA
Giorgio Grisetti	La Sapienza Rome, Italy
Hanspeter Mallot	University of Tübingen, Germany
John Bateman	University of Bremen, Germany
Mary Hegarty	UC Santa Barbara, Germany
Stefan Woelfl	University of Freiburg, Germany
Terry Regier	UC Berkeley, USA

Reviewers

Henrik Andreasson
Alper Aydemir
Leanne Beaudoin-Ryan
Andreas Birk
Mark Blades
Mathias Broxvall
Beth Casey
Susan Wagner Cook
Christian Dornhege
Frank Dylla
Carola Eschenbach
Sara Irina Fabrikant
Caitlin Fausey
Cipriano Galindo
Antony Galton
Merideth Gattis
Nicholas Giudice
Tilbe Goksun
Klaus Gramann
Slawomir Grzonka
Stephen Hirtle
Armin Hornung
Petra Jansen
Gabriele Janzen
Kim Kastens
Maddy Keehner
Jonathan Kelly
Peter Kiefer
Antonio Krueger
Benjamin Kuipers
Rainer Kümmerle
Kevin Lai
Sander Lestrade
Rui Li
Andrew Lovett
Zoltan-Csaba Marton
Tobias Meilinger
Daniel R. Montello
Oscar Martinez Mozos
Stefan Münzer
Philippe Muller
Lauren Myers
Nora Newcombe
Penney Nichols-Whitehead
Matthijs Noordzij
Seyda Ozcaliskan
Kaustubh Pathak
Patrick Pfaff
Pedro Pinies
Laurent Prevot
Shannon Pruden
Martin Raubal
Florian Raudies
Kai-Florian Richter
Dario Lodi Rizzini
Roy Ruddle
Stefan Schiffer
Angela Schwering
Somayajulu Sripada
Bastian Steder
Martin Stommel
Todor Stoyanov
Hauke Strasdat
Andrew Stull
Sabine Timpf
Barbara Tversky
Andrew Vardy
Jan Oliver Wallgrün
Thomas Wolbers
Diedrich Wolter
Desislava Zhekova
Joost Zwarts

Sponsoring Institutions

The conference was sponsored by the University of Freiburg and the University of Bremen in collaboration with the NSF through the Spatial Intelligence and Learning Center (SILC) and the DFG through the SFB/TR8.

Table of Contents

Predicting What Lies Ahead in the Topology of Indoor Environments

Alper Aydemir[1], Erik Järleberg[1], Samuel Prentice[2], and Patric Jensfelt[1]

[1] CVAP, Royal Institute of Technology, KTH, Sweden
{aydemir,erikjar,patric}@kth.se
[2] CSAIL, Massachusetts Institute of Technology, USA
prentice@mit.edu

Abstract. A significant amount of research in robotics is aimed towards building robots that operate indoors yet there exists little analysis of how human spaces are organized. In this work we analyze the properties of indoor environments from a large annotated floorplan dataset. We analyze a corpus of 567 floors, 6426 spaces with 91 room types and 8446 connections between rooms corresponding to real places. We present a system that, given a partial graph, predicts the rest of the topology by building a model from this dataset. Our hypothesis is that indoor topologies consists of multiple smaller functional parts. We demonstrate the applicability of our approach with experimental results. We expect that our analysis paves the way for more data driven research on indoor environments.

1 Introduction

Imagine a mobile robot tasked with finding an object on an unexplored office building floor. The robot needs to plan its actions to complete the task of object search and the search performance depends on the accuracy of the robot's expectations. As an example, having found a corridor and an office, its expectation of finding another room by exploring the corridor should be higher than exploring the office as corridors act as connectors in most indoor environments.

In most systems where this type of structural information can be beneficial, the models of indoor environments are hard-coded and not learned from data. Indoor environments are generally organised in interconnected spaces each fulfilling a certain function. A natural way of modeling these environments is by building a graph where each vertex represents a room in the environment and an edge between two vertices indicates a direct, traversable path. Each vertex can have several attributes such as a room category (kitchen, office, restroom etc.), area size and perimeter length. This type of representation is often called a topological map in the literature. More recently, researchers became interested in augmenting topological maps with semantic information by extracting the aforementioned attributes from sensory data [1, 2, 3]. Although there exists a large body of work on building topological maps, little consideration is given to the analysis and prediction in these maps. One reason for this is building data

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 1–16, 2012.

driven models of topological maps requires collecting data from a high number of actual buildings, recording the floorplan layout including the rooms and adding each room's attributes. This is much harder than an image annotation task.

We leverage on the MIT floorplan database [4], and we assume unique space categories, meaning that each floorplan can contain only one of each space category. This is a limitation that we plan to overcome in future work. This leads to a dataset containing 567 floors, 6426 unique spaces with 91 space categories and 8446 connections between the spaces in total. An example partial topology from the dataset is shown in figure 1. To the best of our knowledge, no previous work exists on the analysis and usage of a dataset of this type and scale. First, we provide an analysis of the topological properties of a large indoor floorplan dataset. Second, we develop a method to predict both the structure (i.e. which type of rooms are connected to each other) and the vertex labelings (i.e. which type of rooms are most commonly found) from a large real-world annotated semantic indoor topology database. We do this on basis of the hypothesis is that indoor environments are topologically arranged in small functional units, e.g. $\{corridor - bathroom - office\}$ or $\{corridor - mailbox - office\}$. Therefore by extracting these frequently occurring topological patterns we can make accurate predictions even though the specific input graph at hand contains rooms of previously unknown categories. Rooms with unknown categories in the input graph corresponds to a real world problem where a robot's classifier may be largely uncertain about a room's category or that the robot has no model for that specific room. Even in this case, the system should still provide reasonable predictions.

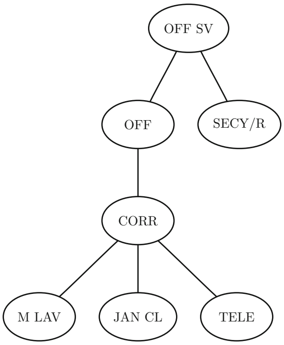


Fig. 1. An example graph from the dataset

2 Problem Formulation

2.1 Preliminaries

We represent each floor as an undirected graph. Each vertex in a graph is assigned a label from an ordered, finite alphabet such that no two vertices share the same label [5]. A graph is then a three-tuple $G = (V, E, \alpha)$ where V is a finite vertex set, $E \subset V \times V$ is a finite edge set and $\alpha : V \to \mathcal{L}$ is a vertex label mapping. Let $\mathcal{G}$ be the set of all formable graphs using the label alphabet $\mathcal{L}$.

The *graph edit distance* is a notion used to measure how similar two graphs are to each other. It is based upon what are called *edit operations* on a graph. An edit operation is a change performed upon a graph to transform it into a new graph. Normally one considers: vertex substitutions, vertex additions, edge additions, and vertex deletions as possible graph edit operations. We will restrict these operations to two specific types: *edge addition* between two existing vertices in the graph; and, *vertex addition*, which creates a new labeled vertex connected to one of the existing vertices. This is to ensure that we get no disconnected parts and the resulting graphs are connected. With this restriction upon the set of possible edit operations, one cannot always expect to be able to transform an arbitrary graph g_1 into g_2. However if we restrict the domain so that $g_1 \subseteq g_2$ or vice-versa, it is always possible to transform one into the other without considering vertex deletions for example.

We will also denote by $\phi(g_1, g_2)$ the set of possible edit operation sequences transforming g_1 into g_2. Using this we define the distance between two graphs g_1 and g_2 as the minimal cost of transforming one graph into the other: $d(g_1, g_2) = \min\limits_{s \in \phi(g_1,g_2)} c(s)$. It can be shown that this function satisfies the four properties of a metric [6]. We define the ball of a certain radius r to be the set of all graphs which are at most r edit operations away from the graph. That is, $B(G, r) = \{G' \in \mathcal{G} | d(G, G') \leq r\}$.

A *graph database* $\mathcal{D} = \{G_1, ..., G_n\}$ is a set of graphs. Given a graph $G \in \mathcal{G}$ and a graph database $\mathcal{D}$, we define the *projected database* as the set of supergraphs of G. We denote this set as $\mathcal{D}_G = \{G' \in \mathcal{D} | G \subseteq G'\}$. The cardinality of the projected database is called the *frequency* of the graph G in the graph database $\mathcal{D}$ and is denoted by $freq(G) = |\mathcal{D}_G|$.

We may now define the *support* of the graph G as:

$$supp(G) = \frac{freq(G)}{|\mathcal{D}|} \tag{1}$$

A graph G will be called a *frequent subgraph* in $\mathcal{D}$ if $supp(G) \geq \sigma$ where σ is some minimum support threshold, $0 \leq \sigma \leq 1$.

Let $\mathcal{S}$ be the set of frequent subgraphs of the graph database $\mathcal{D}$ for some minimum support threshold σ. That is, $\mathcal{S} = \{G \in \mathcal{D} | supp(G) \geq \sigma\}$

For any given pair of graphs g_1 and g_2, the *Pearson's Correlation Coefficient* describes the linear correlation between the two graphs in the database is defined as in [7]:

$$\theta(g_1, g_2) = \frac{supp(g_1, g_2) - supp(g_1)supp(g_2)}{\sqrt{supp(g_1)supp(g_2)(1 - supp(g_1))(1 - supp(g_2))}} \tag{2}$$

Finally, the neighbourhood of a vertex v in a graph G will be denoted by $N_G(v)$ or simply $N(v)$ when it is clear which graph is meant. The neighbourhood of v is the induced subgraph of vertices which are adjacent to v in the graph G.

2.2 Formal Graph Prediction Problem Formulation

We define the problem as follows. Given a graph database $\mathcal{D}$ we want to find a certain discrete probability distribution. This distribution is an estimate of how probable a certain edit operation upon the current partial graph is. Let $G_p \subset G$ be called the partial graph which is a subgraph of some unknown supergraph G. The *set of all possible next graphs* given a partial graph is the ball of radius one around the partial graph using the graph edit distance metric. That is, the set of all possible next graphs is $B(G_p, 1)$. Once the discrete probability distribution above has been acquired, it is then possible to attain the most probable next graph $G'_p \in B(G_p, 1)$. This graph is simply the result of performing the most probable edit operation upon G_p.

3 The Method

3.1 Analysis of Dataset

We start by presenting the insights gained by analyzing the dataset. Each floor in the MIT floorplan dataset consists of a set of *spaces* and their connections to other spaces. Floors can be represented as graphs; the spaces can be interpreted as vertices of a graph and the connections as graph edges [4]. A space can be a room surrounded by walls and accessible via doors, but sometimes a space can also have invisible boundaries, e.g. a coffee shop at the end of a corridor.

Connector spaces such as *corridor* and *stair* are crucial parts of any indoor environment since they act as indoor highways. Our intuition tells us that spaces that have the functionality to connect other rooms and floors together should appear with high frequency in natural indoor environments. Table 1 shows the most frequent vertices in the MIT floorplan dataset with their occurrence frequency in all floors. As can be seen, *corridor* and *stair* are in most floors, ranking as the top two frequent spaces. Offices are also a common space in campus buildings.

Furthermore, we would expect to see some common patterns in floorplans. For example, we would expect certain facilities such as lavatories and elevators to

Table 1. Most frequent spaces in the dataset. Here “JAN CL”, “ELEC”, “OFF SV” are abbreviations for janitor closet, electricity cabinet and office service, respectively.

Vertex	Support
STAIR	85%
CORR	78%
OFF	67%
OFF SV	60%
ELEC	60%
JAN CL	57%
LOBBY	48%

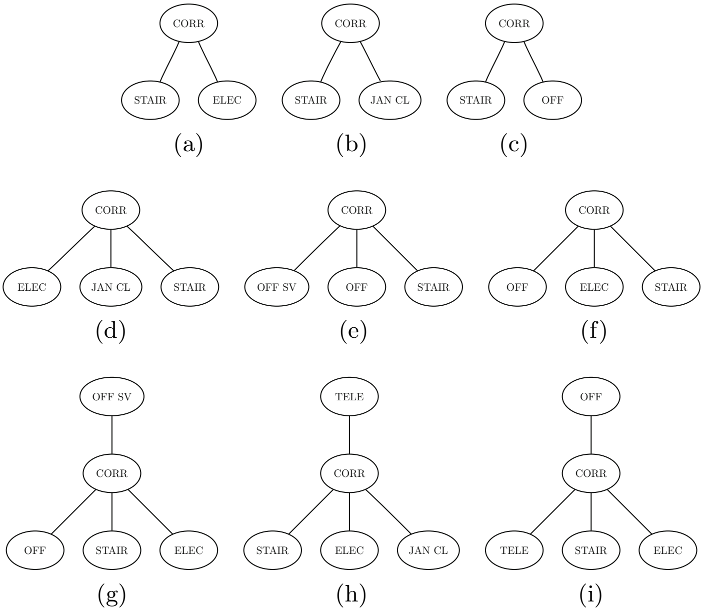


Fig. 2. The three most common frequent subgraphs for graph sizes 3, 4 and 5. The frequencies for subgraphs shown in figures 2a-2c are 37.66, 37.11, 36.56, for figures 2d-2f they are 26.50, 25.04, 25.04 and finally for figures 2g-2i they correspond to 17.18, 17.00, 17.00, respectively.

be at easily reachable locations, or connector spaces such as corridors frequently attached to office rooms. Figure 2 shows the most frequent subgraphs in the dataset for graph sizes 3, 4 and 5. It is remarkable that even for large graph sizes with 4 and 5 vertices, certain patterns are commonplace in the dataset. This supports the hypothesis that indoor topologies consist of commonly occurring smaller parts.

Figure 3a shows the Pearson's correlation coefficient [7] (explained in section 2.1) for the frequent subgraphs in the dataset which occur in more than 16% of all graphs (the frequent subgraph set $\mathcal{S}$ with $\sigma = 0.16$). The graphs are ordered such that the top left pixel is the most frequently occurring subgraph and the top right pixel corresponds to the least frequent. Each pixel represents a frequent subgraph pair and brightness corresponds to high correlations. As an example, figure 3b and 3c correspond to pixel (19,12) or (12,19), which is the highest correlated pair found in this set. Having observed for example the graph in figure 3b, we could say that the edit operation leading to the graph in figure 3c is very probable. The corresponding edit operation would be an *edge addition*, adding an edge between the "OFF" and "P CIRC" vertices.

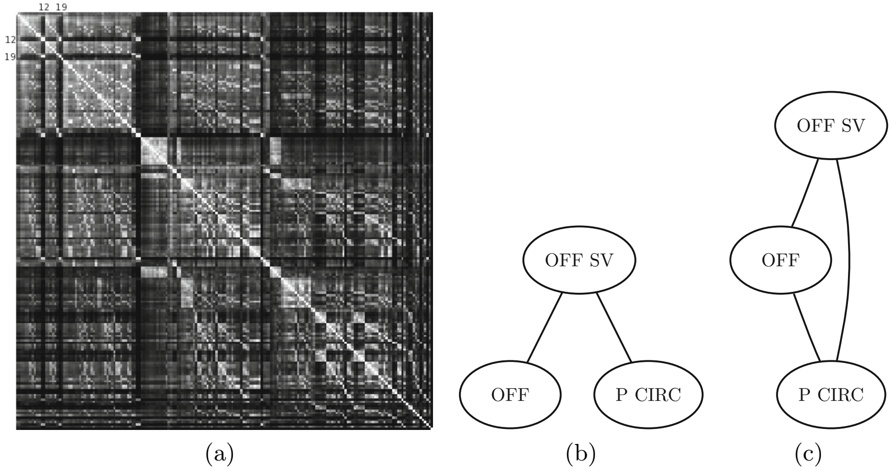


Fig. 3. 3a: Pearson's correlation coefficient for frequent subgraphs occurring in more than 16% of all graphs in dataset. Each pixel represents a frequent subgraph pair and brightness corresponds to high correlations. Subgraphs are ordered by frequency descending from top left pixel. 3b and 3c show the highest correlated pair, corresponding to pixel (19,12) or (12,19) in 3a.

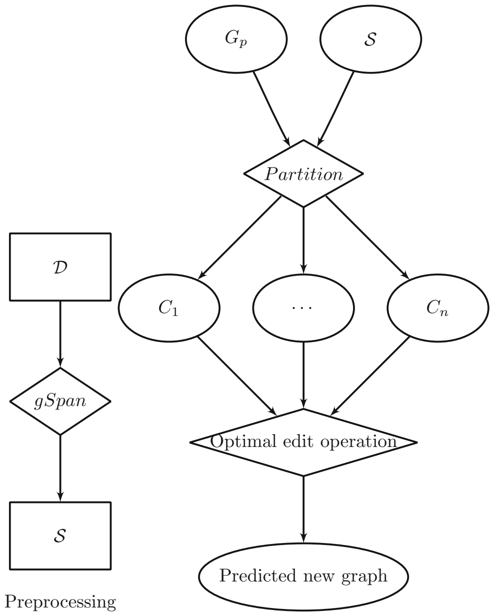
3.2 Method I

Given an initial input graph G_p, we first compute its projected database $\mathcal{D}_{G_p}$. Then, for each graph $E, E \in \mathcal{D}_{G_p}, E \in B(G_p, 1)$, we calculate the edit operation from G_p to E. Finally, the edit operation whose resulting graph has the highest support is determined. This algorithm is naive in the sense that it considers the whole graph at once. This is akin to a hidden Markov Model formulation where the state of the model is the graph itself and actions are edit operations.

The algorithm performs well for small graph sizes. This is encouraging, however we would expect the naive method to fail for larger graphs. By taking into account the overall structure of G_p (defined in section 2.1) as a whole, the algorithm fails to capture the functional patterns with which humans have designed indoor floorplans. As an example, when a rare vertex is connected to a frequently occuring part of the input graph, the algorithm only considers those graphs which include the rare vertex disregarding others, ignoring the functional aspect of subparts of an indoor topology.

3.3 Method II - Prediction with Graph Splitting

In this method, we make use of the frequently occuring subgraphs in the database. We extract frequent subgraphs using the gSpan Algorithm [8]. This provides us with a frequent subgraph database $\mathcal{S}$ which is used in the first step of the prediction. See figure 4a.



(a) Preprocessing

(b) Prediction

Fig. 4. Predition Algorithm Overview. (a) Frequent subgraphs $\mathcal{S}$ are extracted from the graph dataset $\mathcal{D}$. (b) In each iteration, edit operations are hypothesized on selected subsets of the input graph G_p, and the optimal edit operation is executed.

The main steps of this method is given in the following:

1. Split the input partial graph into smaller, overlapping subgraphs which are included in $\mathcal{S}$.
2. For each of these subgraphs of the partial graph, determine the probability of every possible edit operation.
3. Combine the results of the estimates of the edit operations for each subgraph into a final solution for the whole partial graph.

These three steps are summarized in figure 4b.

Step 1: The aim of this step is to divide the partial input graph G_p into a set of overlapping connected subgraphs C where $\forall x \in C, \exists y \in C, x \cap y \neq \emptyset$. The procedure for computing C is given in algorithm 1. The selection of subgraphs plays an important role in prediction quality. We pick the elements of C as much as possible from the frequent graph set $\mathcal{S}$. The rationale behind this is that since indoor topologies consists of multiple functional smaller parts, the algorithm should try to identify those and later expand them as viable predictions. First we determine which of the frequent subgraphs from $\mathcal{S}$ that are present in the current partial graphs, and extract the largest possible such frequent subgraphs set and call it P.

In short, algorithm 1 iteratively checks for the elements of S which are included in G_p (the set P) and which share at least one vertex with the list of subgraphs found so far, C, so as to disregard disconnected subgraphs. Another reason is that computing the list of all possible connected subgraphs of G_p becomes intractable even for small-sized graphs. Therefore we utilize the frequent subgraphs of the graph database to bootstrap this computation and cut down the search space.

Step 2: In this step, we aim to calculate the probability of all possible edit operations for each subgraph of G_p. Let $\mathcal{D}_{C_i}$ be the projected database of any subgraph C_i of G_p, that is, the set of all those graphs which are supersets of C_i. Let x be some graph which is one edit operation away from C_i, that is $x \in B(C_i, 1)$. We then define $\phi(x, C_i) = |\{G' \in \mathcal{D}_{C_i} | x \subseteq G'\}|$. That is $\phi(x, C_i)$ gives the number of times we've observed a specific edit operation upon C_i among all the graphs. The most likely edit operation to perform given that we've observed the subgraph C_i is then given by $\underset{x \in B(C_i,1)}{\arg\max}\, \phi(x, C_i)$. This procedure is given in detail in algorithm 3.

Step 3: Given that we have calculated the most likely edit operation for each of the subgraphs $C_1, ..., C_n$, we have for each of these an optimal edit operation leading to new graphs $C'_1, ..., C'_n$ respectively. We must select one of these, and for any selection C'_j made, the resulting prediction will be $G'_p = \bigcup_{i \in [1,n] \setminus \{j\}} C_i \cup C'_j$. We simply select the edit operation which has the highest support from the graph database. That is, $\underset{C_i, i \in [1,n]}{\arg\max} \phi(C_i, C'_i)$.

Given the function $\phi : \mathcal{G} \times \mathcal{G} \to \mathbb{N}$, it is possible to arrive at an estimate of the discrete probability distribution of the different edit operations upon G_p. The distribution is calculated in a frequentist manner and is given by:

$$P(G'_p = x) = \frac{\phi(x, C_j)}{\sum_{y \in B(C_j,1)} \phi(y, C_j)}, x \in B(C_j, 1) \tag{3}$$

C_j here refers to the selected subgraph and is chosen as detailed above.

Figure 5a shows the initial partial graph which is the input to the prediction algorithm. In this example the input graph is divided into three subgraphs. The output of the first step of the algorithm is shown in black in figure 5b, 5c and 5d. In the second step, the predicted edit operation with the highest support for each subgraph C_i is shown in green. Finally, in the third step, the edit operation with the highest probability is selected.

This splitting of the input graph agrees with the claim that indoor topologies consist of smaller functional parts. Figure 5b shows that while some vertices may be rare (such as "SHAFT"), they can occur in a frequent subgraph pattern, in this case forming a "maintenance" functional group. Figure 5d shows a very common structure, with a corridor as a root node. Finally, in figure 5c, we can see that the algorithm has identified a lobby group. This is also quite common, that the lobby acts as a root node similar to a corridor vertex connected in a tree-like structure.

Algorithm 1. Graph splitting

Input:

- G_p, the current partial graph

Output:

- $C = \{C_1, ..., C_m\}$, the overlapping subgraphs of the partial graph

```
$P \leftarrow \emptyset$
for $s \in \mathcal{S}$ do
    if $s \subseteq G_p \wedge (\neg\exists s' \in \mathcal{S}, s \subseteq s', s' \subseteq G_p)$ then
        $P \leftarrow P \cup \{s\}$
    end if
end for
{$P$ now contains those frequent subgraphs which are contained in the partial
graph $G_p$. They are also the largest possible frequent subgraphs. }
sort(P) by graph size, descending.
$C \leftarrow \{\text{FindCommonFreqSubgraph}(P, G_p, \emptyset)\}$
while $|G_p| \neq |\bigcup_{i=1}^{n} C_i|$ do
    $Found \leftarrow 0$
    for all $c \in C \wedge Found = 0$ do
        $c' \leftarrow \text{FindCommonFreqSubgraph}(P, c, C)$
        if $c' \neq \emptyset$ then
            $C \leftarrow C \cup c'$
            $Found \leftarrow 1$
            break
        end if
    end for
    if $Found = 0$ then
        $D_g \leftarrow G_p \setminus \bigcup_{i=1}^{n} C_i$
        Add the following vertex set to $D_g$: $\bigcup_{v \in V(D_g)} N(v, G_p) \setminus D_g$
        Add the edges (from the edge set of $G_p$) which correspond to the vertex
        additions above.
        $C \leftarrow C \cup \text{GetComponents}(D_g)$
        return $C$
    end if
end while
return $C$
```

Algorithm 2. FindCommonFreqSubgraph

This function will attempt to find another frequent subgraph from the set P that has some vertex in common with some graph C_i (the already established subgraphs of G_p).

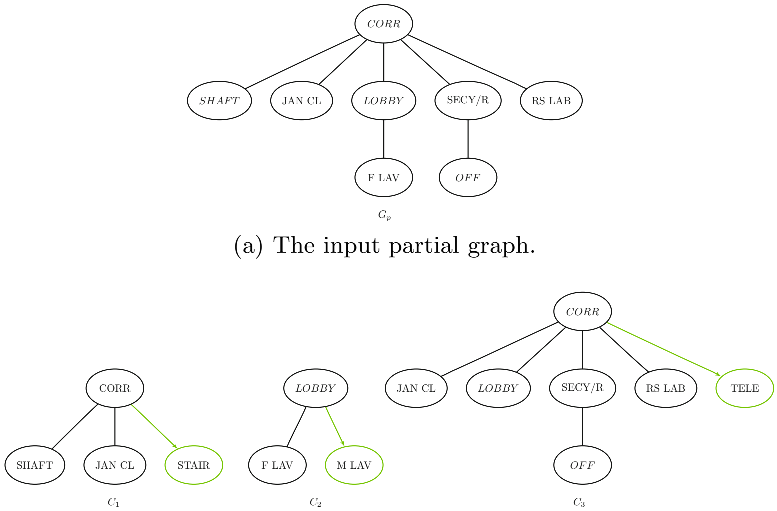
Input:

- P, the sorted sequence of frequent subgraphs that are present in the partial graph
- G, a graph which the result should have some vertex in common with, this is always some C_i except for the initial execution.
- $C = \{C_1, ..., C_n\}$, the thus far added overlapping subgraphs of the partial graph

Output:

- p, the largest frequent subgraph present in the partial graph that has atleast one vertex in common with G (if found). p is also removed from the set P. If no such p could be found, it returns the empty graph $\emptyset$.

```
for all $p \in P$ do
    if HasVertexInCommon$(G, p) \wedge p \nsubseteq \bigcup_{i=1}^{n} C_i$ then
        $P \leftarrow P \setminus \{p\}$
        return p
    end if
end for
return $\emptyset$
```



(a) The input partial graph.

(b) First subgraph (c) Second subgraph (d) Third subgraph

Fig. 5. The overlapping subgraphs of a partial graph and the prediction for each subgraph shown in green

Algorithm 3. Find most likely graph edit operation

Input:

- G, a "small" graph, one subgraph from the output of the graph splitting.
- $\mathcal{D}$, the graph database

Output:

- G', the graph which is the result of performing the optimal edit operation upon G

```
for $x \in \mathcal{D}$ do
  if $G \subseteq x$ then
    for $G' \in B(G, 1) \wedge G' \subseteq x$ do
      {Every $G'$ corresponds to some valid edit operation upon $G$ (that is, both $G$ and $G'$ are contained in this specific graph $x$).}
      $\phi(x, G) \leftarrow \phi(x, G) + 1$
    end for
  end if
end for
return $\arg\max_{x \in B(G,1)} \phi(x, G)$
```

4 Experiments

4.1 Example Runs

To illustrate the method, five different states of a prediction sequence are shown in figure 6. The complete unknown graph G is shown in black dashed lines. The starting initial graph is shown in blue. A predicted edit operation existing in G is shown in green and if it does not exist in G, then it is shown in red.

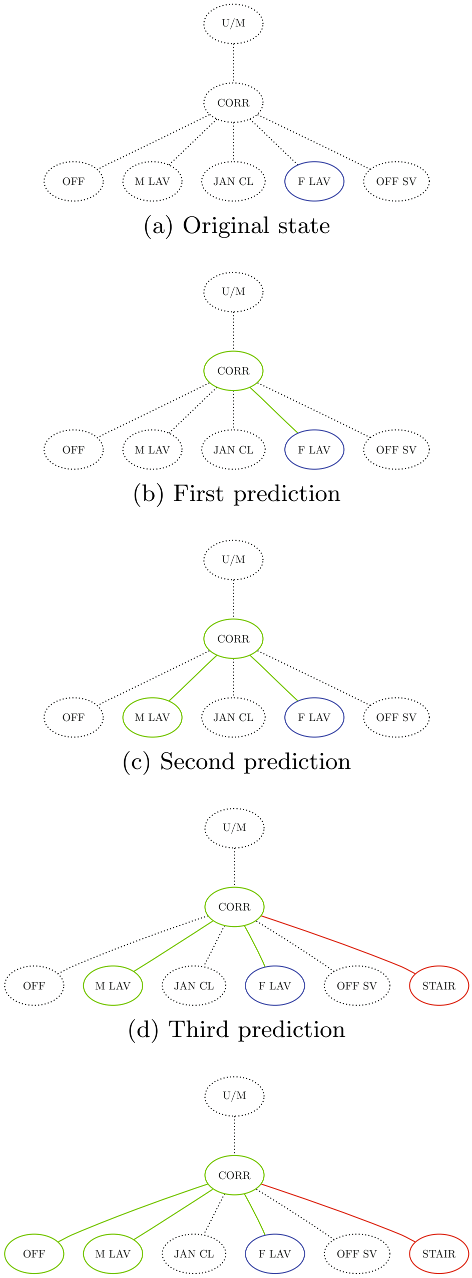
In figure 6a, the partial graph only consists of the female lavatory vertex “F LAV”. The prediction algorithm is then applied to produce figure 6b. The next likely edit operation is to add a corridor “CORR” and connect it to the “F LAV”. Next in figure 6c, we can see the result of executing the prediction algorithm upon the previous graph consisting of “F LAV” and “CORR”. Given that we have observed “F LAV” and “CORR”, the algorithm suggests that it is plausible to have a male lavatory “M LAV” connected to the corridor as well.

As another example, the input graph in figure 7a results in the discrete probability distribution shown in figure 7b. Since this partial graph consists of only two vertices, the only edit operations considered are those of adding a new vertex. On the horizontal axis, the different edit operations are shown as $A \rightarrow B$, where A is some existing vertex of the partial graph and where B is the vertex which should be added and connected to A. Note that edit operations with a probability below 0.02 are not shown. In this case, A can only take the values of janitor closet “JAN CL” or male lavatory “M LAV”. Note that as expected the *corridor* “CORR” vertex has the highest probability of being connected to another vertex by a large margin.

4.2 Quantitative Evaluation

We have compared the results of the two prediction methods. To measure the performance of the algorithms for varying graph sizes, we have selected 2000 partial graphs randomly from the dataset, for each graph size between one and 20. In total 40000 different partial graphs were processed. The selection process works as follows. First we pick a random graph from the dataset $\mathcal{D}$. Then for a given graph size $m = \{1...20\}$, we pick at random m connected vertices which form an input graph. Then, we iterate this process until 2000 partial graphs are selected. Finally, the graphs from which random partial graphs were picked are excluded from the training dataset (multiple partial graphs may come from the same graph).

We counted the number of correct edit operations predicted by each algorithm over the test set, and divided by the total number of partial graph predictions to get a percentage of correct predictions (shown in figure 8). The naive algorithm (*Method I* in section 3.2) is shown in dashed blue and the prediction algorithm with graph splitting (*Method II* in section 3.3) in red. For smaller graph sizes, their performance is almost equivalent. However for larger graphs, the performance of the naive algorithm decreases dramatically compared to the algorithm with splitting. The naive algorithm must compute support for edit operations



(a) Original state

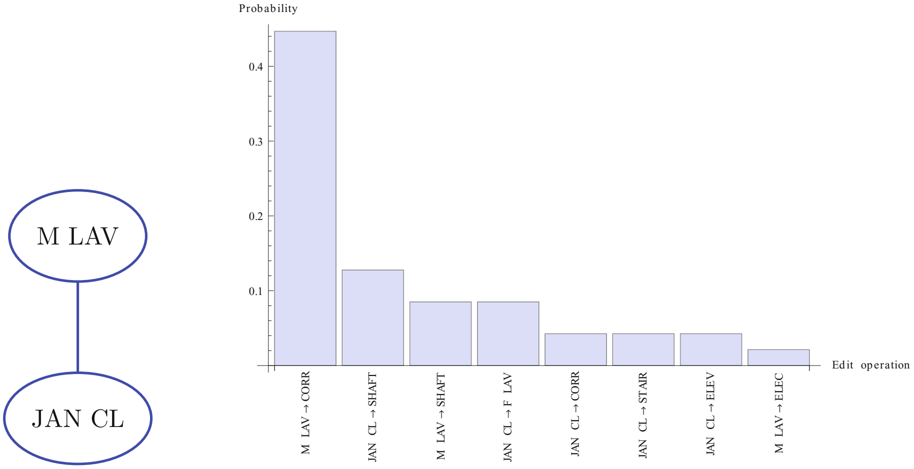
(b) First prediction

(c) Second prediction

(d) Third prediction

(e) Fourth prediction

Fig. 6. The evolution of a predicted graph with four consecutive predictions. The dashed lines are the unknown true graph. The blue node corresponds to the initial input graph, green nodes or edges represent a correct prediction that exists in the true graph whereas red represents a predicted node or edge absent in the actual true graph.



(a) Partial graph

(b) Probability distribution

Fig. 7. The discrete probability distribution for the edit operations of a partial graph. Given the partial graph in (a), vertex addition hypotheses are shown on the x-axis in (b), with corresponding probabilities.

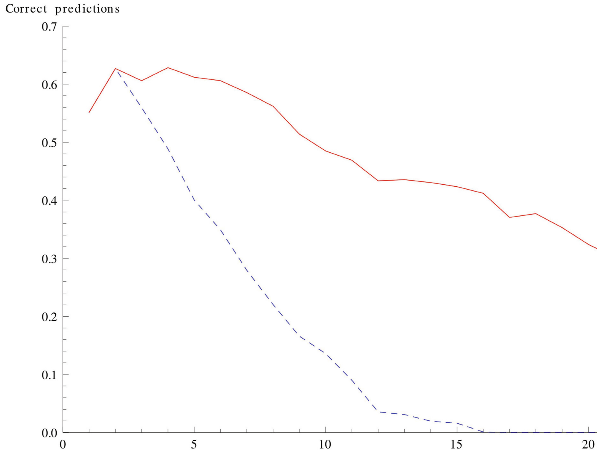


Fig. 8. Comparsion between the two prediction methods over 40,000 partial input graphs. The naive algorithm (*Method I*) is shown as the blue dashed line, and the prediction algorithm with graph splitting (*Method II*) as a solid red line.

on the *whole* graph, and is therefore subject to data sparsity and overfitting as the graph prediction size increases. Method II, however, leverages graph splitting and frequent subgraph extraction to focus on the *functional* components of the graph. This not only prunes the hypothesis space, but also enables greater predictive power through small functional groups, which have more substantial support in the dataset.

5 Conclusion and Future Work

In this paper we have provided an initial analysis of a large real-world indoor topological database. We have shown experimentally that the presented methods predict indoor topologies accurately. To the best of our knowledge, no previous work exists on analyzing and using a large real-world floorplan database for predicting indoor topologies. Furthermore, we have shown that indoor topologies consists of functional smaller parts which in turn can be used to develop methods with better prediction results. The reason for this is such methods capture the rationale behind man-made indoor spaces.
Following this work, we expect a large interest in developping the data-driven methods on indoor environments. We have yet to exploit the rich set of information offered by such datasets.

Future work consists of removing the assumption of only having one of each space category per floorplan, modeling the number of room types, extending the database with data from other environments such as KTH campus, making use of the metric coordinates in the data to have richer predictions and investigate how the predictions generalize for different locations.

Acknowledgements. The authors thank Emily Whiting, Seth Teller and the RVSN group at MIT CSAIL[1] for their helpful pointer to acquire the dataset. We are also grateful to John Folkesson and Kristoffer Sjöö for their feedback on an early draft of the paper. This work was supported by the EU FP7 project CogX.

References

[1] Pronobis, A., Luo, J., Caputo, B.: The more you learn, the less you store: Memory-controlled incremental SVM for visual place recognition. Image and Vision Computing, IMAVIS (March 2010), http://www.pronobis.pro/publications/pronobis2010imavis, doi:10.1016/j.imavis.2010.01.015

[2] Ranganathan, A.: Pliss: Detecting and labeling places using online change-point detection. In: Proceedings of Robotics: Science and Systems, Zaragoza, Spain (June 2010)

[3] Mozos, O., Stachniss, C., Rottmann, A., Burgard, W.: Using AdaBoost for place labeling and topological map building. In: Thrun, S., Brooks, R., Durrant-Whyte, H. (eds.) Robotics Research, vol. 28, pp. 453–472. Springer, Heidelberg (2007) ISBN 978-3-540-48110-2

[1] http://rvsn.csail.mit.edu

[4] Whiting, E., Battat, J., Teller, S.: Topology of urban environments. In: Proc. of the Computer-Aided Architectural Design Futures (CAADFutures), Sydney, Australia, pp. 115–128 (July 2007)

[5] Valiente, G.: Efficient algorithms on trees and graphs with unique node labels. In: Applied Graph Theory in Computer Vision and Pattern Recognition, pp. 137–149 (2007)

[6] Bunke, H., Allermann, C.: Inexact graph matching for structural pattern recognition. Pattern Recognition Letters 1, 245–253 (1983)

[7] Ke, Y., Cheng, J., Yu, J.X.: Efficient discovery of frequent correlated subgraph pairs. In: Proceedings of the 2009 Ninth IEEE International Conference on Data Mining, ICDM 2009, pp. 239–248. IEEE Computer Society, Washington, DC (2009)

[8] Yan, X., Han, J.: gspan: Graph-based substructure pattern mining. In: Proceedings of the 2002 IEEE International Conference on Data Mining, ICDM 2002, p. 721. IEEE Computer Society, Washington, DC (2002)

Object Categorization in Clutter Using Additive Features and Hashing of Part-Graph Descriptors

Zoltan-Csaba Marton, Ferenc Balint-Benczedi, Florian Seidel,
Lucian Cosmin Goron, and Michael Beetz

Intelligent Autonomous Systems, Technische Universität München, Germany
{marton,balintbe,seidelf,goron,beetz}@cs.tum.edu
http://ias.cs.tum.edu

Abstract. Detecting objects in clutter is an important capability for a household robot executing pick and place tasks in realistic settings. While approaches from 2D vision work reasonably well under certain lighting conditions and given unique textures, the development of inexpensive RGBD cameras opens the way for real-time geometric approaches that do not require templates of known objects.

This paper presents a part-graph-based hashing method for classifying objects in clutter, using an additive feature descriptor. The method is incremental, allowing easy addition of new training data without recreating the complete model, and takes advantage of the additive nature of the feature to increase efficiency. It is based on a graph representation of the scene created from considering possible groupings of over-segmented scene parts, which can in turn be used in classification. Additionally, the results over multiple segmentations can be accumulated to increase detection accuracy.

We evaluated our approach on a large RGBD dataset containing over 15000 Kinect scans of 102 objects grouped in 16 categories, which we arranged into six geometric classes. Furthermore, tests on complete cluttered scenes were performed as well, and used to showcase the importance of domain adaptation.

Keywords: segmentation, hashing, classification, scene-graphs, clutter.

1 Introduction

In many if not most service robotic applications, the ability to recognize a large amount of objects (in the range of several thousands, as identified by Biederman [1]) and to localize them is an essential task. In order to match the perception capabilities of humans, Dickinson in [2] advocates that searching for predefined templates is not enough, and that recognition of new exemplars of known categories has to be facilitated. On this premise, Marton et al. [3] use geometric cues for categorization and visual cues for instance classification.

In everyday tasks realistic scenes contain clutter, where objects are not always entirely visible, e.g. due to partial occlusion. Therefore a classification algorithm is presented that deals with detection of general geometric categories of objects

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 17–33, 2012.

in clutter, based on object segments. Learning the different segments and their combinations that form objects is a scalable way to capture the different object categories a robot would encounter. For example, a mug is typically a cylindrical segment, next to a handle, or a teapot is a combination of different rounded shapes with a top and a large handle.

Our approach is similar to the one by Lai and Fox [4] and by Mozos et al. [5] in that it performs part-based categorization in cluttered scenes. We combine the creation of multiple segmentations from [5] with an extended version of creating multiple groupings of these segments [4], and present our approach that was designed to handle multiple instances of objects from several categories, that were labeled according to their general 3D shape.

As in the earlier work, object "parts" are not predefined, but are the results of a segmentation technique, that possibly over-segments objects. The employed segmentation does not produce the same parts for the same object under all circumstances (something that is hard to achieve), but it proved that it produces similar segments for similar objects in [5].

In [5] we assigned each part to a part cluster that was created by unsupervised clustering, and used the class distribution in the "activated" cluster to link the detection to the user-given labels. While in that work information coming from the different parts of the object was combined by a Hough voting scheme for identifying the object's 2D centroid, here we take an approach that is more close to [6]. We identify to what object does each segment belong by considering its descriptor and that of neighboring parts, together with the local topology of the scene. In this sense, it is am improvement over the vocabulary of parts and simple vote accumulating approach from [5].

Although our classification is less complex than the one presented by Lai and Fox [4] it still manages to capture relations between segments in the "soup of segments" approach of Malisiewicz and Efros [7] and can be used to achieve the same effect as their "domain adaptation", as we will show in our experiments. Additionally, we trained on the same number of classes, but with multiple objects per class by grouping objects based on their geometry, and described the topology of the segment groups by graph-theoretic properties in order to improve and speed up classification through hashing.

With the appearance of low-cost 3D sensors like the Kinect, it is expected that large amounts of data will be available for training, and robots can be equipped with it in order to obtain good quality RGB and 3D information at once. Nonetheless, no matter how extensive the training data, a robot might always be confronted with a novel object instance or type when operating in a real dynamic environment. Since teaching all the possible objects a robot might encounter is not always possible, categorization becomes an important step towards learning these novel objects [2]. Thus a mechanism is needed to enable a robot to autonomously acquire new object models as it operates in an environment, and to efficiently add it to its classification framework [3].

Therefore, in this work, we propose a classification architecture that robots can use to make sense of all this training data efficiently using various methods, and describe our approach for identifying object categories in realistic setups, that is extensible with new objects without requiring a full re-training.

As a summary, the main contributions of this work are as follows:

- an efficient part-based object classification method for cluttered scenes, taking into account relations between parts;
- a graph-theoretic hashing method that allows model refinement while having competitive classification performance;
- exemplifying the advantages of multiple segmentations, multiple segment groupings, and "domain adaptation";
- and a lightweight framework that enables easy comparison of different combinations of features and methods.

In the next section we will give an overview of the related work, followed by the description of the framework and the proposed classification method. Then we will present the used segmentation and additive feature, and give details on the creation of the segment-graphs' hash table. We present the evaluation of our method in Section 8 and conclude in Section 9, discussing the possibilities for future improvements.

2 Related Work

Object detection typically involves the computation of descriptive image or 3D features, either for some key-points or for object clusters, for example the ones by Lowe (SIFT) [8] or Rusu et al. (VFH) [9].

In Fergus et al. [10] the authors are proposing an unsupervised scale-invariant learning scheme, in order to detect objects on a wide range of images. Objects are modeled as flexible constellations of segments using a probabilistic representation for all significant aspects of the object. The work exploits the expectation-maximization algorithm in a maximum-likelihood setting. While our method is somewhat different, it can be viewed as the 3D application of the presented principles. The main difference between our approach and the work done by Huber et. al [11] and Fergus et. al [10] is that we are using an additive feature, that enables considering very efficient segment combinations, aided by a hashing procedure of graph features for fast lookups.

Classification using additive features was also performed by Kanezaki et al. [12] using the Liner Subspace Method on the feature space, as presented by Watanabe and Pakvasa [13]. Since the employed Color-CHLAC feature was rotation variant, they had introduced artificial variations in their training data. The learning method itself does exploit the additive property of the feature for classification based on partial views, but not the relations between the different segments of the objects. A related idea to ours was explored by Mian et. al [14], but with a hash table built on local 3D features for CAD fitting, while we have feature independent hashing.

Part-based detection, however, grouped with multiple segmentations, offers several advantages, including efficiency scalability as argued by Mozos et al.

[5], Lai and Fox [4] and by Huber et al. [11]. Segmentation of objects is a well researched domain, with various existing approaches [15,16,17,18], but as argued in [7], rather then relying on a single segmentation that is possibly erroneous, multiple segmentations should be considered.

Typical graph-based object detection approaches employ complex algorithms to find valid configuration. In contrast, we are using the graph representation of the parts in the scene only to capture their spatial configuration, and use the obtained descriptor as hashing keys. This allows for cheap partial re-training when novel objects need to be added, as the training set is partitioned into multiple parts, and separate classifiers. This is in contrast with our previous work [3], but obviously assumes known labels of the novel exemplars.

Regarding the classification framework, a related, but more complex framework, the STAIR Vision Library is presented by Gould et al. [19], designed to support the Stanford AI Robot project. Its machine learning capabilities include SVM, Boosted classifier, classification based on decision trees etc. While the proposed architecture offers roughly the same machine learning capabilities, it is easier to integrate in existing projects as it is implemented as a service in the Robot Operating System framework (`www.ros.org`). Thus its combination with image and 3D processing tools like OpenCV and Point Cloud Library (`www.pointclouds.org`) and robotics software is simplified.

3 Classification Architecture

The framework has a layered architecture consisting of *feature extraction tools*, *dataset tools* and the *classifier manager*.

The feature extraction tools are command line programs for walking directory structures and extracting various 3D and 2D features from the files encountered. The feature extraction process is capable of labeling the instances according to the hierarchical relationships reflected by the directory structure of the dataset.

3.1 General Feature Storage and Command Line Processing Tools

All the feature extraction tools use a common file format to store the extracted features. The dataset files produced in the *feature extraction* layer can be further processed by tools from the *dataset tools* layer. Here various operations are implemented, some of which are:

- serialization tools for loading/saving in ASCII or binary format,
- concatenation of features, combination of datasets and dropping of columns,
- re-labeling of hierarchical dataset to flatten them,
- splitting of datasets for creating training and test partitions (for example for cross-validation),
- bag-of-words (BoW) creation from a feature dataset (used for creating global descriptors out of local ones, like SURF),
- converting classifier confidences to datasets (used for stacking),
- pre-processing like dimensionality reduction, whitening.

3.2 Standardized Interface for Classifiers

The *classifier manager* is integrated in ROS to allow using the trained classifiers in ROS based applications. It manages the life-cycle of classifier instances and acts as a facade for communicating with the classifier instances which must all be derived from a simple interface. Currently Support Vector Machines [20] and mostly classifiers from OpenCV [21] are included.

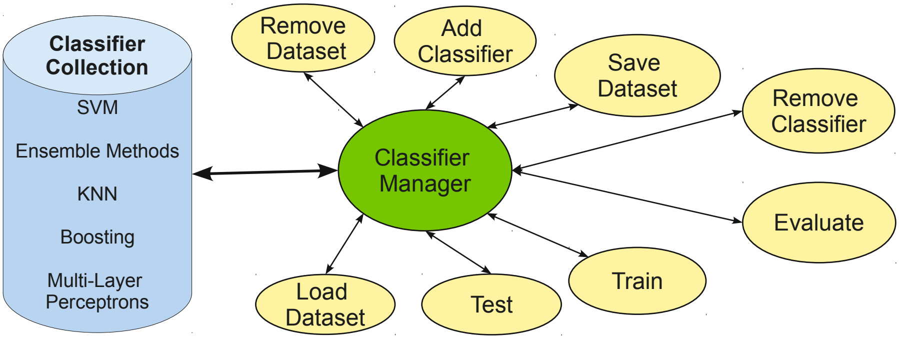


Fig. 1. Simplified scheme of the classification framework. The thin arrows represent possible service requests.

The most common operations for classifiers implemented are shown in Figure 1. These are ROS services that are handled by a *classifier manager* node and each service can operate on different classifiers that run in parallel having an unique identifier. Other services that ease and make classification tasks more efficient include adding a common dataset to be shared by all classifiers, computing and storing a confusion matrix, returning various evaluation statistics (e.g. number of true positives) and timing of feature estimation and classification. All of these primitive operations form a simple domain specific language from which experiments can be constructed using shell scripts or ROS nodes, i.e. C++, Python, Java and Lisp code.

The code along with tutorials and the used data labeled using ground-truth information can be downloaded from our svn repository *http://code.in.tum.de/svn/ias-cf*.

4 Part-Based Recognition

The basic idea of our classification method is that segmenting objects accurately does not always work robustly and will lead to classification failures, but oversegmentation is easily realizable [7,4,5]. These segments that then represent only parts of objects can be used to compute features, and then combined to build up object candidates. We use this approach and use additive features[1] and graph-theoretic description of segment arrangements.

There are of course multiple ways of combining segments and not all of them create a valid object. However, we can test if a combination is valid by checking if the combined feature vector is known. We also exploit the fact that segments and their connections (neighborhood relations) can be treated as a graph, and only certain types of sub-graphs are present in the graph formed by the parts of an object. Checking for subgraph isomorphism is not practical, but there are several descriptors one can employ to rule out isomorphism. Thus, during training we decompose our objects into parts, compute the features for each part, build the part-graph, and generate all sub-graphs along with their combined features. Figure 2 shows an overview of the main steps of our method.

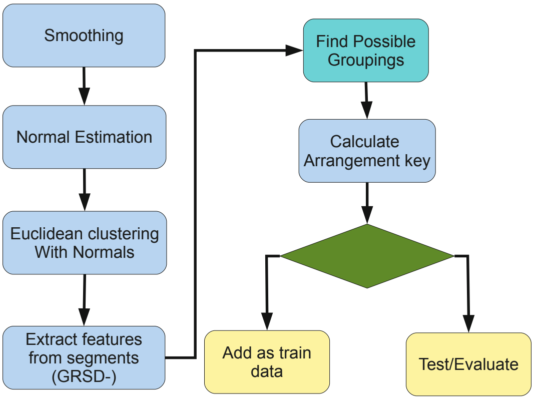

Smoothing
Normal Estimation
Euclidean clustering With Normals
Extract features from segments (GRSD-)
Find Possible Groupings
Calculate Arrangement key
Add as train data
Test/Evaluate

Fig. 2. Main steps of the processing and feature extraction

These features along with the object's class are then saved in a multi-level hash table, where the keys are the number of parts, and the identifiers of the subgraph topology. When testing, the procedure of decomposition and part-graph building is repeated, and starting at each part, all the subgraphs are grown that are not larger than anything seen during training. These are then checked for which objects can they be parts of. Similarities are accumulated in the source part for final classification.

5 Segmentation and Part Graphs

We use the segmentation criteria presented in [5] to over-segment the scans, such that patches with a relatively small curvature are considered, as shown in Figure 3. We use a region-growing approach, that starts at a random point and

[1] If the feature is additive, the descriptor that would be computed for the object is the same as the sum of the features of its segments.

grows the segment until the deviation from the seed normal does not exceed 45 degrees. This way, selecting different seed points result in multiple segmentations of the point cloud into parts. As discussed in the introduction, this process is not completely reproducible, therefore we rely on the large amount of training data to cover all possible cases. As we can produce multiple segmentations by choosing different (random) seed points, different part decompositions can be used for training, which is useful if not enough training examples are available. This was the case in our earlier experiments based on a smaller laser-based dataset [22], where this strategy improved classification rates by 5%.

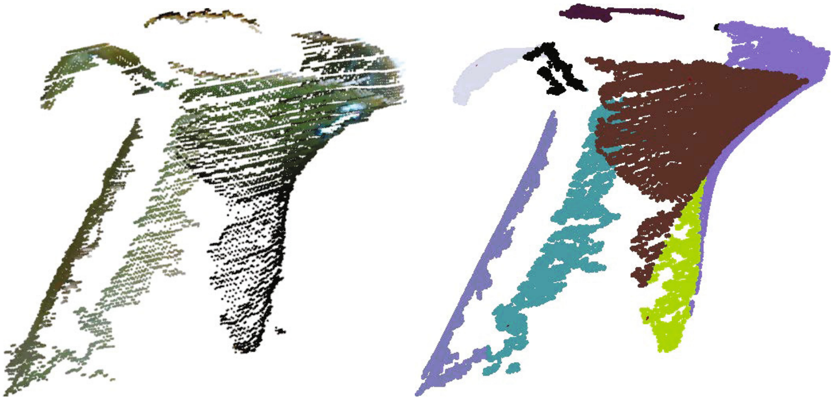


Fig. 3. Cut out object from Kinect frame with visualized color (left) and parts (right)

Since we are dealing with tabletop scenes, the supporting plane can be removed prior to processing, and only points above it considered as in [3]. Small segments are discarded, and for each segment we subsequently compute the GRSD- feature (more detailed description in section 6) and store it for later use. We then extract the part neighborhoods by checking if the physical distance between two parts falls below a threshold, and build a connectivity matrix.

Starting at each vertex, we create all the possible groupings up to a certain size (number of regions in the grouping) in order to obtain our "soup of segments", and create the groups' hash codes. Note that since the graph vertices can be sorted, it is possible to efficiently enumerate all sub-graphs containing a given vertex without repeating already generated ones.

For the hash codes, apart from the number of vertices/parts, we chose to concatenate the sorted list of vertex degrees as well to form a second level of keys. As an alternative for this second level of keys we experimented with using the eigen values of the graph's Laplacian matrix, but found that the results upon evaluation were the same. For this reason, in the upcoming testings and evaluations we used the sorted list of vertex degrees, as they are simpler to compute. This degree order is unique for isomorph graphs, however different graphs can have the same degree order.

6 Simplified and Additive GRSD

Pinz [23] defines categorization as being a *generic object recognition* (generic OR), this being the opposite of a *specific OR*, which deals with the recognition of a specific object (an instance of those categories). Example categories are cups, plates, boxes, etc. while in a specific OR one aims to recognize a specific box, cup etc. Since most household objects from the same category share similar shape, but distinctive appearance, 3D features are typically used for generic ORs, while visual features for specific ones.

An example where the two were combined is presented in [24], using a geometric classifier based on Global Radius-based Surface Descriptor (GRSD) and a visual classifier based on SIFT. As described in [25], we adapted the GRSD feature to be additive, by simplifying it to simple histogram of neighborhoods of surfaces of different type, neglecting the ray-tracing step. This simplified feature (which we called "GRSD-") is very efficient to compute, and we compared its descriptiveness to the original implementation and VFH, a very strong 3D feature.

Table 1. Selected object categories from the RGBD Dataset with good 3D data, grouped into general geometric categories. The number of objects in each category is given in parentheses, with the total number of scans being 80k, of which every fifth is used.

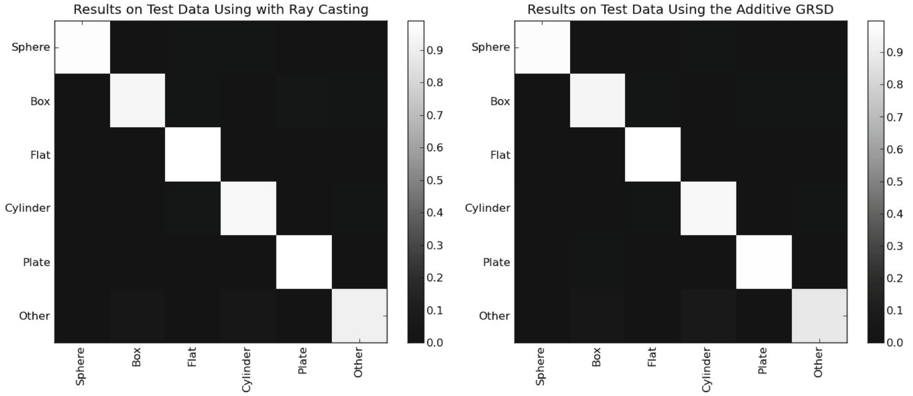
Sphere	Box	Flat (rectangular)	Cylindrical	Plate (disk)	Other
bowl (6)	food box (4)	notebook (5)	coffee mug (8)	plate (7)	cap (4)
ball (6)	sponge (8)	cereal box (5)	food cup (3)		kleenex (5)
		food box (8)	soda can (4)		pitcher (3)
			food can (14)		
			food jar (6)		
			water bottle (6)		

For our tests we used the large RGBD dataset from [26], which contains a total of over 200,000 scans of 300 objects from 51 object categories. As in [26], we use every fifth point cloud from the dataset in our experiments, because the similarity between consecutive point clouds is extremely high. Since in this work we focus on 3D classification, we selected those object categories that have good 3D data (and excluded very small, shiny, transparent objects or noisy scans) and grouped them into geometric categories as shown in Table 1. Please note that the "food box" category contained both regular boxes and large flat boxes, so we had to split them up.

The geometric classes we used were sphere, box, flat (rectangular), cylindrical, plate (disk), as they cover well most of the categories present in the RGBD dataset (for the remaining objects we introduced an "other" category), and as argued in [3] these simple geometric shapes can be used to represent most objects from other publicly available databases as well (e.g. KIT object Models

Table 2. Feature Comparison using SVM

Feature	Number of Dimensions	Accuracy[%]
GRSD	21	95.21
GRSD-	20	95.29
VFH	308	97.35



(a) Original GRSD with ray casting

(b) Additive GRSD with cell neighborhoods

Fig. 4. Confusion matrices for SVM trained on reduced RGBD set

Web Database[2], Household Objects Database from Willow Garage[3], and TUM Semantic Database of 3D Objects[4]).

A comparison between GRSD ,GRSD- and VFH was done training an SVM classifier on the six geometric categories with an RBF kernel on a reduced version of the dataset described in Table 1. Results are shown in Figure 4 and Table 2.

As we can see, the new GRSD version has less dimensions (as transitions from empty to empty voxels are no longer possible) but similar descriptiveness. It approaches the high dimensional VFH's performance, while not being view dependent, and offers the required additive property to easily compute the descriptors of grouped parts.

7 Object Part Hashing

As mentioned earlier, the summed-up training features are saved in a hash table, a classifier is built for the training exemplars in each entry, and the classifier to

[2] http://i61p109.ira.uka.de/ObjectModelsWebUI/

[3] http://www.ros.org/wiki/household_objects_database

[4] http://ias.in.tum.de/software/semantic-3d

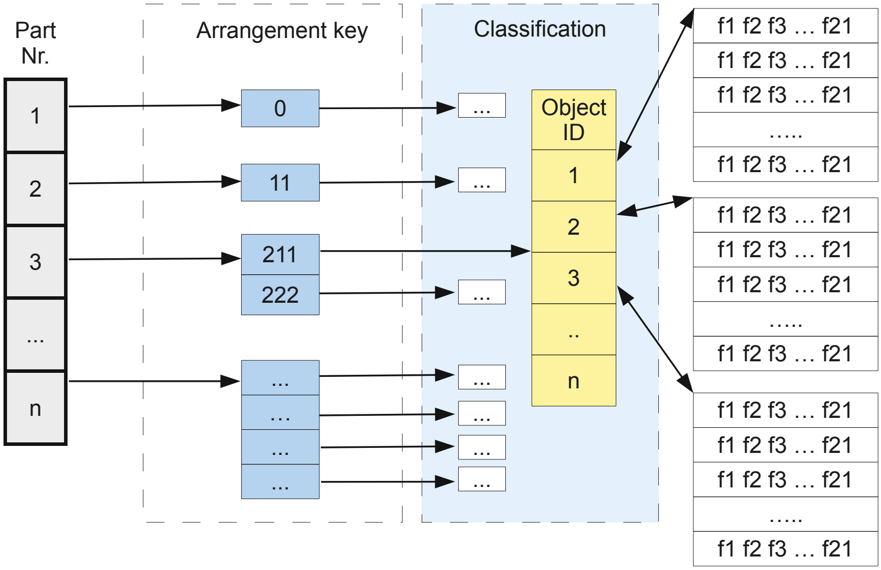


Fig. 5. Structure of the Hash Table

be used for a testing features are looked up in it based on the two keys. The structure of the hash table is shown in Figure 5.

For each class, the maximum classification score is kept, and these scores are summed up for each sub-graph source part. In contrast to the product of the class probabilities for each grouping that was used in [4], we found that this voting approach works better. Similar findings supporting voting were made by Lam and Suen [27] when evaluating combinations of classification results. Similarly, in the case of the experiments with separate objects per point cloud, the parts' scores can be added up (weighted by size) to produce a final classification.

8 Evaluation

Evaluation of the proposed method was done in several steps. In order to be able to test and compare our method with SVM in a fairly short time we reduced the selected dataset presented in Table 1 to roughly 7000 scans of 57 objects in 9 classes presented in Table 3. Testing and training data were split up by using every third scan of an objects for testing purposes and training the classifier on the rest of the scans. After finding the optimal setup for our classifier a final evaluation of the proposed method was conducted on the full dataset. Cluttered scenes were tested on several combinations of the training dataset, presented in more detail in subsection 8.3.

Table 3. Reduced object categories from the selected categories. The number of objects in each category is given in parentheses, with the total number of scans being 30k, of which every fifth is used.

Sphere	Box	Flat (rectangular)	Cylindrical	Plate (disk)	Other
ball (6)	food box (4)	food box (8)	coffee mug (8)	plate (7)	cap(4)
	sponge(8)		food cup (3)		pitcher(3)
			soda can(4)		

8.1 Objects as Separate Clusters

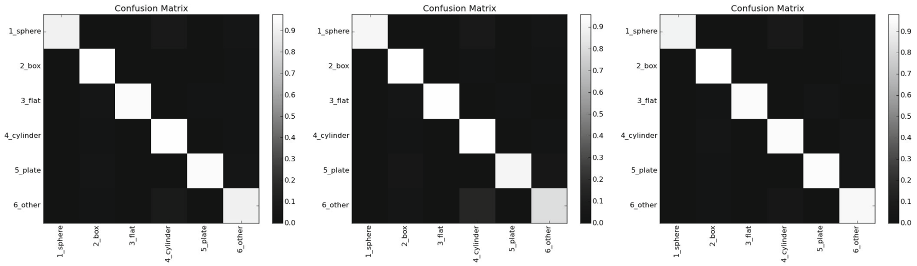
In order to quantitatively evaluate our method, we performed the same test as we did with GRSD- using SVM, and presented our results in Table 4. In addition we tested different distance metrics and found that the Jeffries-Matsuhita distance performs best. Confusion matrices for the different distance metrics using our method are shown in Figure 6.

So far, only whole objects are classified, by summing up the probability distributions of their parts (weighted by size) and selecting the most likely class. As a final evaluation for objects as separate clusters we tested our method on the entire dataset from Table 1, obtaining a 92% successful classification rate.

In order to quantify the advantage of using the part arrangement keys, we evaluated the results for using only the part numbers as hash keys. On the smaller

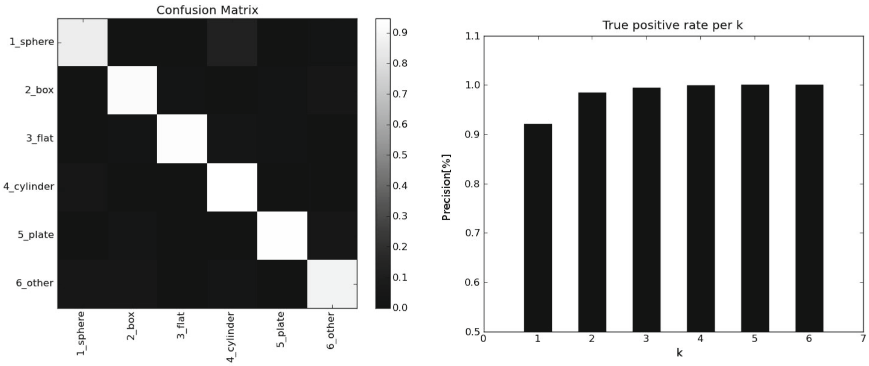
Table 4. Classification results using different distance metrics as compared to the baseline obtained for GRSD-

GRSD- distance metric	Accuracy[%]
Euclidean	94.5
Manhattan	92.5
Jeffries-Matsuhita	95.5
Baseline (SVM – not part-based)	95.3



(a) Euclidean (b) Manhattan (c) Jeffries

Fig. 6. Confusion matrices for different distance metrics



(a) (b)

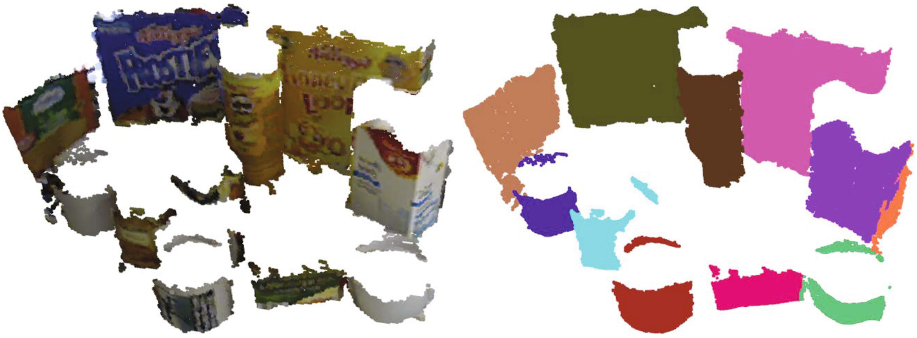
Fig. 7. Confusion matrix and cumulative score for the entire RGBD dataset

dataset this increased classification time by 21.4% and dropped the success rate by 1.2%. On larger databases this difference will only be accentuated even more. The advantage of incorporating the spatial relations between parts is highlighted in the next subsection as well.

Another interesting observation is shown in Figure 7 (b). When taking into account the cumulative results, considering the second most likely result already improves our classification rate by 5%. Knowing this, we could be able to take advantage of the possible re-segmentations of the testing object in case we obtain similar top votes.

8.2 Comparison to Segmentation-Based Classification

The main advantage of our approach is that it does not rely on a correct segmentation and that it takes relations between parts into account (both through the



(a) original (b) segmented

Fig. 8. Cluttered scene and its geometry-based segmentation according to [28]

hashing and the cumulative voting of part groupings). To exemplify this, consider the results for the test scene shown in Figure 8. Using the GRSD- feature, our approach classifies around 90% of the points correctly, while a geometric segmentation followed by a nearest neighbor classification obtained a correct classification for around 65% of the points. Even though the segmentation has few errors for this scene, misclassification rate is quite hight, but using our part-grouping approach results are much better. For a full evaluation of the method on larger scenes, please see the following subsection.

8.3 Complete Cluttered Scenes

Figures 9 and 10 show the two additional tabletop scenes on which we tested our approach. The color red represents the *sphere* class, blue *cylinders*, yellow *boxes*, and cyan the *flat* class. We are currently in the process of capturing and labeling a larger set of scenes, that will present more varied object types. Manual labeling is a time-consuming process, but we feel that the previous subsection provided enough support for the validity of the method. Based on the results on these datasets, here we will present additional aspects of our work as well.

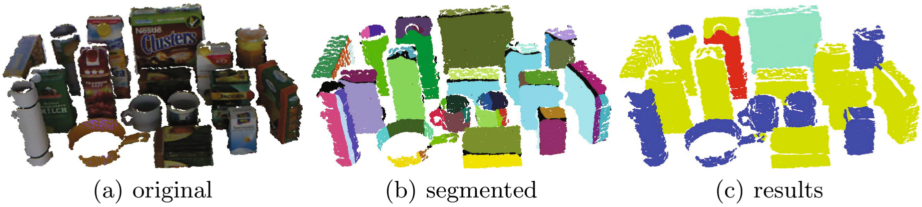

(a) original (b) segmented (c) results

Fig. 9. Segmentation and classification on a cluttered table scene 1

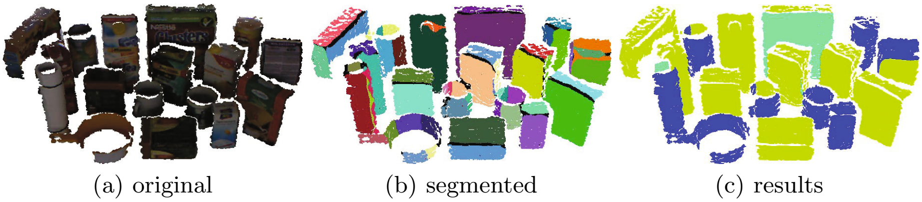

(a) original (b) segmented (c) results

Fig. 10. Segmentation and classification on a cluttered table scene 2

The cluttered scene testing was run on the two tabletop scenes, using different datasets or combinations of datasets as training data shown in Table 5. As it is expected result vary depending on the type of dataset we used for training. Although results when testing objects as separate clusters were good for the RGBD datasets, when testing the scenes the results were far from satisfying. This is because the scenes contain very different kinds of objects.

Table 5. Classification results for tableto p scenes using different training datasets

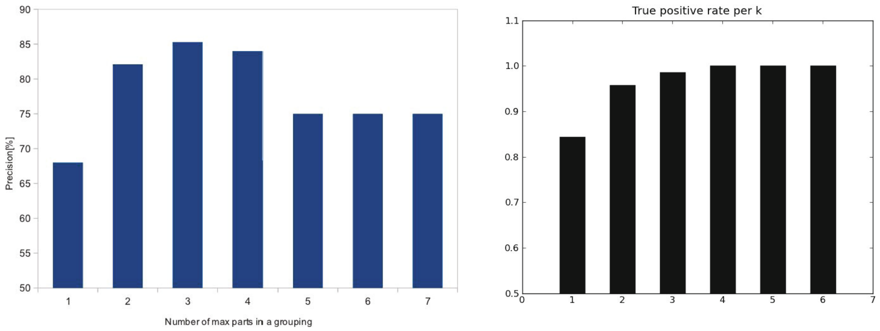
Datasets:	RGBD-Small	RGBD	VOSCH	Small+VOSCH	RGBD+VOSCH
Scene 1					
per point	73%	47%	76%	**84%**	54%
per segment	83%	49%	67%	**83%**	54%
Scene 2					
per point	76%	54%	70%	**80%**	58%
per segment	76%	41%	72%	**74%**	47%

In order to diversify our training data we combined the RGBD dataset with the "VOSCH" Kinect scan dataset used in [25], consisting of 63 similar objects than in our scenes, captured from different viewpoints with an angular step of 15 degrees. Similarly to [4], we found that this "domain adaptation" improves results, as seen in Table 5. However, as the results on the larger RGBD dataset suggest, identifying the correct weighting of the two data sources is necessary, possibly based on an evaluation set. Apparently, as the number of objects increases, confusions get more frequent, therefore the weight of the domain specific objects need to be increased.

In the case of the smaller dataset, the combination with the scans from VOSCH improved over the results on both separate sets of training data, highlighting the importance of mixing various sources of information while keeping specific specialties. Related ideas are discussed in [29] as well, where the task and environment adaptation improves perception capabilities. Thanks to the hashing approach, handling large databases and dynamically adding new objects is alleviated, as training times are reduced, and only affected groups have to be re-trained.

One of the important contributions of our work is the grouping of neighboring parts and taking into account the results of these groups. When grouping neighboring parts there is an upper threshold on the maximum number of parts that can form a grouping. When training the classifier this threshold was set to 8. In the case of tabletop scenes we experimented with this threshold and found that the optimal maximum number of parts that can form a grouping is actually less then the one used for single objects. Results of this evaluation are shown in Figure 11 (a). It can be observed that grouping the segments greatly improves the classification process up to a given number of parts. However, if we choose the threshold to be too high in a cluttered scene, we risk grouping parts together that do not belong to the same object.

As a final experiment, we repeated the experiment from Figure 7 (b) for the combined (per-point) results of the two scenes as well, see Figure 11 (b). Again, we found that the top votes are correct in the majority of cases, with the success rate increasing by nearly 12% to 95.8% with the first two votes being considered.



(a) effect of part sub-graph size

(b) percentage of correct top k votes

Fig. 11. Results for different maximum parts in a grouping, and different number of top votes considered

9 Conclusion and Future Work

In this paper we have shown the advantages of exploiting multiple segmentations and part-graph descriptors to deal with object categorization in clutter. The proposed methods were evaluated on a large RGBD dataset, and on Kinect scans of cluttered tabletop scenes, and showed promising results.

Given the class distributions for the different parts in the complete scenes, a subsequent verification or fitting step would be needed to find the exact object locations and poses. Nonetheless, these probabilities alone provide the assignment of parts to objects, and offer valuable input about likely object hypotheses. Additionally, as argued in [10], partial occlusion and shape variability is handled well by object part relations, a property we could enhance further by exploiting the additive property of the used feature.

Probably the most critical issue is the relatively non-descriptive simplified GRSD-. Future work will concentrate on using an additive version of the more descriptive VFH or the improved GRSD-/Color-CHLAC combination [25]. Additive features also allow for partial correspondence, by checking if all bins of a feature vector are smaller than those of a trained vector. This way occlusions don't "break" a detection, but special classification methods are needed in order to take advantage of this (for example, Linear Subspace Method as in [25]).

Currently only nearest neighbors classification is used in combination with hashing, but we plan to exploit the advantages of the classification framework to use more discriminative methods as well.

Acknowledgments. This work was supported by the DFG cluster of excellence CoTeSys (Cognition for Technical Systems) at the Technische Universität München; by the PR2 Beta Program of Willow Garage, Menlo Park, CA; and by the project "Doctoral studies in engineering sciences for developing the knowledge based society – SIDOC" contract no. POSDRU/88/1.5/S/60078.

References

1. Biederman, I.: Recognition-by-components: A theory of human image understanding. Psychological Review (1987)
2. Dickinson, S.: The evolution of object categorization and the challenge of image abstraction. In: Dickinson, S., Leonardis, A., Schiele, B., Tarr, M. (eds.) Object Categorization: Computer and Human Vision Perspectives (2009)
3. Marton, Z.C., Pangercic, D., Blodow, N., Beetz, M.: Combined 2D-3D Categorization and Classification for Multimodal Perception Systems. The International Journal of Robotics Research (2011)
4. Lai, K., Fox, D.: Object recognition in 3d point clouds using web data and domain adaptation. The International Journal of Robotics Research 29(8), 1019–1037 (2010)
5. Mozos, O.M., Marton, Z.C., Beetz, M.: Furniture Models Learned from the WWW – Using Web Catalogs to Locate and Categorize Unknown Furniture Pieces in 3D Laser Scans. Robotics & Automation Magazine 18(2), 22–32 (2011)
6. Marton, Z.C., Rusu, R.B., Jain, D., Klank, U., Beetz, M.: Probabilistic Categorization of Kitchen Objects in Table Settings with a Composite Sensor. In: Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, USA, October 11-15 (2009)
7. Malisiewicz, T., Efros, A.A.: Improving Spatial Support for Objects via Multiple Segmentations. In: Proceedings of the British Machine Vision Conference (2007)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
9. Rusu, R.B., Bradski, G., Thibaux, R., Hsu, J.: Fast 3d recognition and pose using the viewpoint feature histogram. In: Proceedings of the 23rd IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Taipei, Taiwan (October 2010)
10. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: CVPR, pp. 264–271 (2003)
11. Huber, D., Kapuria, A., Donamukkala, R.R., Hebert, M.: Parts-based 3d object classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2004) (July 2004)
12. Kanezaki, A., Nakayama, H., Harada, T., Kuniyoshi, Y.: High-speed 3d object recognition using additive features in a linear subspace. In: ICRA, pp. 3128–3134 (2010)
13. Watanabe, S., Pakvasa, N.: Subspace method in pattern recognition. In: Proceedings of 1st International Joint Conference on Pattern Recognition (1973)
14. Mian, A.S., Bennamoun, M., Owens, R.: Three-dimensional model-based object recognition and segmentation in cluttered scenes. IEEE Trans. Pattern Anal. Mach. Intell. 28, 1584–1601 (2006)
15. Bergström, N., Björkman, M., Kragic, D.: Generating Object Hypotheses in Natural Scenes through Human-Robot Interaction. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 827–833 (September 2011)
16. Mishra, A.K., Aloimonos, Y.: Visual Segmentation of "Simple" Objects for Robots. In: Robotics: Science and Systems (RSS) (2011)
17. Fowlkes, C.C., Martin, D.R., Malik, J.: Local figureground cues are valid for natural images. Journal of Vision 7(8) (2007)
18. Comaniciu, D., Meer, P., Member, S.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 603–619 (2002)

19. Gould, S., Russakovsky, O., Goodfellow, I., Baumstarck, P., Ng, A.Y., Koller, D.: The stair vision library (v2.4) (2010), http://ai.stanford.edu/~sgould/svl
20. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
21. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
22. Balint-Benczedi, F., Marton, Z.-C., Beetz, M.: Efficient part-graph hashes for object categorization. In: 5th International Conference on Cognitive Systems, CogSys 2012 (2012)
23. Pinz, A.: Object categorization. Found. Trends. Comput. Graph. Vis. 1, 255–353 (2005)
24. Marton, Z.C., Pangercic, D., Rusu, R.B., Holzbach, A., Beetz, M.: Hierarchical object geometric categorization and appearance classification for mobile manipulation. In: Proceedings of 2010 IEEE-RAS International Conference on Humanoid Robots, Nashville, TN, USA, December 6-8 (2010)
25. Kanezaki, A., Marton, Z.C., Pangercic, D., Harada, T., Kuniyoshi, Y., Beetz, M.: Voxelized Shape and Color Histograms for RGB-D. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Workshop on Active Semantic Perception and Object Search in the Real World, San Francisco, CA, USA, September, 25–30 (2011)
26. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view rgb-d object dataset. In: Proc. of International Conference on Robotics and Automation, ICRA (2011)
27. Lam, L., Suen, C.Y.: Optimal combinations of pattern classifiers. Pattern Recognition Letters 16(9), 945–954 (1995)
28. Goron, L.C., Marton, Z.C., Lazea, G., Beetz, M.: Segmenting cylindrical and box-like objects in cluttered 3D scenes. In: 7th German Conference on Robotics (ROBOTIK 2012), Munich, Germany (May 2012)
29. Horswill, I.: Integrating vision and natural language without central models. In: Proceedings of the AAAI Fall Symposium on Embodied Language and Action (1995)

Efficient Object Categorization with the Surface-Approximation Polynomials Descriptor

Richard Bormann, Jan Fischer, Georg Arbeiter, and Alexander Verl

Fraunhofer IPA, Nobelstr. 12, 70569 Stuttgart, Germany
{richard.bormann,jan.fischer,georg.arbeiter,
alexander.verl}@ipa.fraunhofer.de
http://www.ipa.fraunhofer.de/

Abstract. Perception of object categories is a key functionality towards more versatile autonomous robots. Object categorization enables robots to understand their environments even if certain instances of objects have never been seen before. In this paper we present the novel descriptor Surface-Approximation Polynomials (SAP) that directly computes a global description on point cloud surfaces of objects based on polynomial approximations of surface cuts. This descriptor is directly applicable to point clouds captured with time-of-flight or other depth sensors without any data preprocessing or normal computation. Hence, it is generated very fast. Together with a preceding pose normalization, SAP is invariant to scale and partially invariant to rotations. We demonstrate experiments in which SAP categorizes 78 % of test objects correctly while needing only 57 ms for the computation. This way SAP is superior to GFPFH, GRSD and VFH according to both criteria.

Keywords: Object Categorization, Robot Vision, 3D Descriptor.

1 Introduction

Mobile service robots which are intended to serve people in natural household environments need to retrieve rich information about their surroundings to accomplish tasks given to them. A major part of this perception problem is the recognition of objects for interaction. Although powerful object detection algorithms exist today they do not suffice for a versatile operation. Neither should every single object occurring in the environment be learned by the robot in advance nor would this even be realizable with respect to current hardware limitations. Object categorization solves this problem by identifying the class of formerly unseen object instances. Hence, the perception problem decreases significantly in size. This work focuses on the categorization of small and medium-scale rigid household objects with a simple shape.

The use of point cloud data is a good starting point towards successful object categorization in this case since many common object classes in households

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 34–53, 2012.

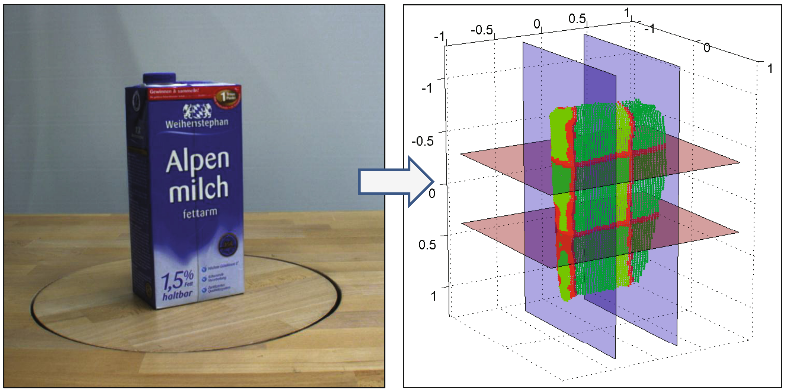


Fig. 1. Computation of the SAP descriptor: the point cloud of the milk box is scale normalized and cut with planes. The red surface points at the cuts contribute to the polynomial approximation.

expose strong similarities in shape whereas texture may differ significantly. Moreover, algorithms for categorization should evaluate fast as there is limited computing power and energy supply available on a mobile robot while users expect fast responses. To our knowledge, the best algorithms with respect to runtime use Global Fast Point Feature Histograms (GFPFH) [1], Global Radius-based Surface Descriptors (GRSD) [2] or Viewpoint Feature Histograms (VFH) [3], that describe the properties of an object's point cloud based on point normals. They are computed within less than a second by several complex steps which also include the denoising of the data that is necessary for reliable normal estimation.

In this work, we present the novel Surface-Approximation Polynomials (SAP) descriptor, which is tailored for the goal of fast and normal-free object categorization on point cloud data. It is based on a pose normalization of the object's point cloud and the approximation of polynomials along cuts of the surface (see Figure 1). We successfully apply this descriptor directly to noisy and unprocessed point cloud data generated by SwissRanger SR4000 and PMD CamCube depth sensors. We will show that this descriptor allows for a categorization performance of 78 % correctly categorized objects on a dataset of 14 classes. This is superior to GFPFH, GRSD and VFH by 9.5 % and more. At the same time the SAP descriptor can be computed within only 57 ms which is faster than each of the three aforementioned descriptors. We also demonstrate the scale invariance of SAP and the partial rotation invariance to tilt and pan. Finally, we provide an outlook for the addition of rotational invariance around the camera axis.

The remainder of this paper is structured as follows: Section 2 provides a literature review on object categorization techniques in different contexts and Section 3 explains the algorithm of the SAP descriptor as well as the categorization framework. In Section 4 we present various experiments which demonstrate the performance and properties of the SAP descriptor. We conclude with a summary and an outlook on future work in Section 5.

2 Related Work

Object categorization is a topic of high interest in the computer vision and the robotics community. However, both areas are quite different concerning their data, constraints and objectives. Computer vision approaches usually rely on plain color images and aim at tasks like image retrieval. A good overview over current methods is provided by Galleguillos and Belongie [4]. Modern techniques in this area mostly use derivatives of Bag-of-Words models (BoW) on local descriptors with Support Vector Machines (SVM) as classifier and attain image categorization results of around 70% on the Caltech-101 dataset [5]. While computer vision research rather focused on categorizing large amounts of classes recently [6], high precision in the predictions is actually more important to robotics. We believe that the use of depth sensors supports this goal since 3D shape is often very characteristic for object classes.

Hence, we have to deal with 3-dimensional data from our objects of interest. Similar to local 2D images features, many local 3D descriptors have been devised. Some popular examples are Spin Images [7], Shape Index [8], 3D SURF [9] and SHOT [10]. For a broader overview on local 3D descriptors and a comparison regarding the object classification task we refer to Knopp et al. [11], who present orientation invariant 3D object categorization based on Bag-of-Words from 3D SURF features and a Hough Transform voting method. They evaluate their algorithm on high resolution full 3D models and obtain 95.5% accuracy for 8 categories. However, the processing for one model takes 20.66s on the 3D SURF features and even more on other local features. Likewise, Toldo et al. [12] present up to 100% accuracy on synthetic data for 6 classes while having computation times around 50s. They apply Bag-of-Words on segmented parts of the object using the Shape Index descriptor and classify with an SVM. Examples of global features for the shape matching task are Shape Distributions [13] and Spherical Harmonics [14]. Although these approaches provide very high accuracies, they do not meet several aspects in robotics: first, highly resoluted 3D meshes are normally not available with current 3D sensors attached to robots. Moreover, the object's surface is regularly captured and categorized from a single view instead of a full 360° model. Finally, computation times should be as low as possible but definitely not in the order of seconds or higher. Therefore, we will propose a 3D descriptor for categorizing single shot object surfaces which is very fast to compute and robust to low quality sensory input yet powerful enough for high categorization rates. Pu et al. [15] present an approach for 3D model retrieval that computes the global descriptor from various slices through the 3D mesh. Their key idea is similar to our approach, however, their work focuses on instance retrieval of highly resoluted full 3D models.

Related work which addresses the issues of robotics is available with unsupervised and supervised category learning. The first problem was investigated by Endres et al. [16] who collect histograms of discretized spin images of objects segmented from 3D laser scanner data and cluster them with Latent Dirichlet Allocation (LDA) in an unsupervised fashion. The resulting classes correspond to balloons, boxes, chairs, swivel chairs and humans with 90.38% accuracy. Labels are assigned within 0.5 s. However, as we like to access the semantic information included in the class labels, we need to use a supervised learning approach.

Based on a linear SVM classifier Bo et al. [17] propose color and depth kernel descriptors which can categorize 86.2 % of the test objects into 51 classes. Further popular descriptors are Surflet-pair-relation histograms [18], Global Fast Point Feature Histograms (GFPFH) [1], which are very similar to the first, and Global Radius-based Surface Descriptors (GRSD) [2]. GFPFH builds on FPFH which computes the sum of angle histograms between angles of normals of each surface point and its neighboring points. These histograms are classified as geometric primitives. Histograms over the occurence of geometric primitives along lines between any two voxels of the object's point cloud yield the GFPFH feature. Its accuracy on a 4 class problem is 96.69% with a computation time of below a second. The GRSD descriptor is composed similarly to GFPFH descriptor from local RSD features, which basically represent the local minimum and maximum curvature around a point. It can categorize 85% of unseen objects correctly into six classes needing around 0.2s for each computation. Another descriptor that is very similar to GFPFH but that also encodes the viewpoint at the visible object surface is the Viewpoint Feature Histogram (VFH) [3]. VFH includes the camera axis in the computation of FPFH histograms to establish viewpoint dependent signatures for the trained objects.

Our descriptor is different from these methods insofar as it does not rely on normal computations and local feature representations. Instead, we construct the descriptor directly in a global fashion on the point cloud data and hence avoid the data preparation and normal computation which can consume quite some time if no GPU is available. This way SAP can be computed very fast within 57 ms. The next section details our approach.

3 Methods

This section describes the concepts of the SAP descriptor and the employed framework for categorizing unknown objects. We present a simple and easy-to-compute descriptor on point cloud data of objects. The SAP descriptor is specially designed for the needs of efficient object categorization on a robot being compact, scale invariant and having little computational demands.

Within the categorization framework, we approach the following two kinds of problems provided that the robot can obtain some descriptor for every object in its surroundings:

1. The robot must find objects of a certain category k in its environment. It will label descriptors as either being members of category k or not. This is a binary classification problem.
2. The robot has to assign a category label to every object found in its surroundings. This problem is essentially a classification problem with multiple classes.

The next section starts with an explanation how the needed data is acquired and preprocessed.

3.1 Data Acquisition and Segmentation

The SAP descriptor solely needs a single shot point cloud $\mathcal{P}$ of the object of interest segmented from the scene. We test the descriptor on two databases captured with a SwissRanger SR4000 sensor and a PMD CamCube, respectively. We did not use the Kinect sensor because it was not yet on the market when the databases were collected. The Swissranger 4000 has a resolution of 176x144 pixels and a depth accuracy of around 1% of the measured coordinates. The PMD CamCube can capture depth images with a slightly higher resolution of 204x204 pixels. Both databases will be introduced in Section 4.1.

For recording, the objects were placed on a rotary disc embedded into a table surface. We placed the depth camera approximately 1 m away from the object center and captured depth images of the objects from a slightly elevated viewing angle in front of a mostly homogeneous background (see Fig. 1). By rotating each object incrementally on the rotary table we recorded point clouds from all sides that can typically be observed by the robot when searching for objects on surfaces of the height of a table. This way, we captured 36 to 72 views per object.

For computing the SAP descriptor the obtained point clouds need to be segmented. We require that objects can only be placed on top of a plane, for example a table as in the case of the database recordings. We use a parametric model fitting algorithm from the PCL library [19] for the iterative estimation and removal of larger planes. This method searches for the plane parameters a, b, c, d and points (x, y, z) satisfying the corresponding plane equation

$$ax + by + cz + d = 0 \quad . \tag{1}$$

Sample points are drawn from the point cloud and associated with a plane according to the RANSAC [20] algorithm. RANSAC iteratively draws triples of points, solves the plane equation and searches for further points supporting this plane. The algorithm terminates when the plane with most supporting points is found with high probability. The volume above this plane is considered as the space of potential objects. Multiple objects inside this volume are separated by Euclidean clustering so that we can only examine simple scenes with this approach. Nevertheless, providing a fancy segmentation algorithm that works in arbitrary situations is beyond the scope of this paper. We refer to the literature for approaches that work properly in many cases [1,21,22]. In the following we describe the SAP descriptor and the categorization methods.

3.2 Surface-Approximation Polynomials Descriptor

The underlying idea of the SAP descriptor is to generate a scale normalized view of an object's surface, cut it with 2-dimensional planar subspaces perpendicular to the camera plane and approximate the projections of the cut with polynomials of even order. For a better understanding of the idea and the single steps we refer to Figure 2. To receive a scale-invariant description, the pose of point cloud $\mathcal{P}$ is normalized by computing the centroid $\mathbf{m}$ of $\mathcal{P}$ and its rescaling factor $s = \max_{\mathbf{p} \in \mathcal{P}} \{|p_x - m_x|, |p_y - m_y|\}$, where p_x ans p_y are the x- and y-coordinates

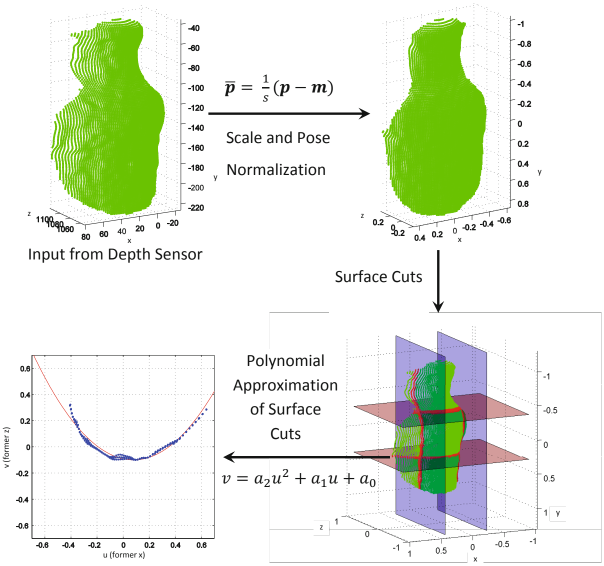


Fig. 2. Computation scheme of the SAP descriptor. The upper left image shows the input point cloud. Then, pose and scale normalization is applied, surface cuts are extracted and finally approximated with a polynomial.

of point $\mathbf{p}$ and m_x and m_y are components of the centroid $\mathbf{m}$. The coordinate system is defined with the z-axis pointing from the camera center towards the scene. The x-axis runs horizontal in the image plane, the y-axis is vertical. Every point $\mathbf{p}$ is then translated to shift the point cloud's center into the origin and scaled by $\frac{1}{s}$. Due to the computation of the rescaling factor s, the $\bar{p}_x$- and $\bar{p}_y$-coordinates of each point $\bar{\mathbf{p}}$ of the transformed point cloud $\bar{\mathcal{P}}$ fall into the range $[-1, 1]$. In summary, the transformation is:

$$\bar{\mathbf{p}} = \frac{1}{s} \cdot (\mathbf{p} - \mathbf{m}) \quad , \tag{2}$$

$$\mathbf{m} = \frac{1}{|\mathcal{P}|} \sum_{\mathbf{p} \in \mathcal{P}} p \quad , \tag{3}$$

$$\mathbf{s} = \max_{\mathbf{p} \in \mathcal{P}} \{|p_x - m_x|, |p_y - m_y|\} \quad . \tag{4}$$

Translating the center of the point cloud to the origin ensures translation invariance with respect to the coordinate system of the depth sensor while the scaling operation effects that the point cloud is resized to a common scale.

After normalization, we sample points from the surface which are approximately located on straight lines parallel to the x- and y-axes. This can be thought of as picking points which approximately lie within cutting $x-z$-planes ($y = \text{const.}$) or $y-z$-planes ($x = \text{const.}$). Specifically, we define to sample points for n_x lines parallel to the x-axis and n_y lines parallel to the y-axis. These lines are equally spaced within the [-1, 1] interval. The cutting planes are illustrated in the lower right image of Figure 2 in rose (y = const.) and purple (x = const.) color. The points associated with cuts are displayed in red.

Following, we approximate the points assigned to each cut with a polynomial of order n_p which essentially comprises the information coded in the point locations into $n_p + 1$ parameters of the polynomial. The polynomials are computed with a standard regression approach: suppose that the coordinates of the points projected into the 2-dimensional subspace are renamed from x or y to u and from z to v. We aim to find the parameters $\mathbf{a} = (a_0, a_1, \ldots, a_{n_p})^{\mathrm{T}}$ of the polynomial $v = a_0 + a_1 u + \ldots + a_{n_p} u^{n_p}$. If we have L sample points on the cutting line, with $L \geq n_p + 1$, we obtain L constraints that can be rephrased in vector notation:

$$v_i = a_0 + a_1 u_i + \ldots + a_{n_p} u_i^{n_p} \ , \ \forall i = 1, \ldots, L \ , \tag{5}$$

$$\mathbf{v} = \mathbf{U} \cdot \mathbf{a} \ , \tag{6}$$

$$\mathbf{v} = [v_1, \ldots, v_L]^{\mathrm{T}} \ , \tag{7}$$

$$\mathbf{U} = \begin{bmatrix} 1 & u_1 & \ldots & u_1^{n_p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & u_L & \ldots & u_L^{n_p} \end{bmatrix} \ . \tag{8}$$

We can easily generate vector $\mathbf{v}$ and matrix $\mathbf{U}$ from the L point samples and solve the linear regression problem with a standard approach, e.g. Singular Value Decomposition. If the available point data is insufficient, we assign zeros to $\mathbf{a}$.

Finally, we concatenate the parameter vectors ${}^{\mathbf{i}}\mathbf{a}^{\mathrm{T}}, i = 1, \ldots, n_x + n_y$, of the n_x and n_y approximated polynomials into one vector $\hat{\mathbf{c}} = [{}^{\mathbf{1}}\mathbf{a}^{\mathrm{T}}, \ldots, {}^{\mathbf{n_x+n_y}}\mathbf{a}^{\mathrm{T}}]$. The final SAP descriptor

$$\mathbf{c} = \left[\frac{\lambda_1}{\gamma}, \frac{\lambda_2}{\lambda_1}, \frac{\lambda_3}{\lambda_1}, \hat{\mathbf{c}} \right] \tag{9}$$

consists of the parameter vector $\hat{\mathbf{c}}$ and some general size information of the object. The eigenvalues λ_1, λ_2 and λ_3 (in descending order) are obtained from a Principal Component Analysis on point cloud $\mathcal{P}$. They encode the size of the object in the three principal directions. λ_2 and λ_3 are expressed relative to λ_1 to encode the relations between the side lengths. λ_1 instead is saved with its full magnitude, except for the constant scaling factor γ, so that the broad information about the object's absolute size is preserved. The constant γ is solely necessary to rescale the entry of λ_1 approximately into the range $[0, 1]$.

The computational complexity for computing the SAP descriptor is $\mathcal{O}(|\mathcal{P}|)$ where $|\mathcal{P}|$ is the number of points in the point cloud. In detail, the single computations have the following complexities:

$$\text{scale and pose normalization:} \quad \mathcal{O}(|\mathcal{P}|), \tag{10}$$

$$\text{point assignment to cuts:} \quad \mathcal{O}((n_x + n_y) \cdot |\mathcal{P}|), \tag{11}$$

$$\text{SVD for polynomial approximations:} \quad \mathcal{O}(L_{max}(n_p + 1)^2 (n_x + n_y)), \tag{12}$$

$$\text{PCA for eigenvalues:} \quad \mathcal{O}(3^2 \cdot |\mathcal{P}|), \tag{13}$$

where $L_{max} \ll |\mathcal{P}|$ is the largest number L of points on a cutting line.

In summary, the SAP descriptor is basically a collection of parameters from polynomials fitted into the normalized surface of an object and three size variables. The three parameters that can be tuned are the numbers of cuts n_x and n_y parallel to the x- and y-directions as well as the order n_p of the polynomial.

3.3 Category Learning

After the computation of descriptors we have to apply a method that separates the different object categories by finding the characteristics in their descriptors. Recall that we aim at enabling the algorithm for two tasks: the search for an object instance of a certain class, which is a binary one-against-all classification problem, and the detection of the category of an unknown object, which is a multi-class decision problem. Therefore, the classification algorithm is based on N binary one-against-all classifiers that distinguish each of the N object classes against the others. A straightforward extension of these binary classifiers for multi-class distinction without learning a new classifier will be discussed in Section 3.4.

We use Random Forest classifiers [23] for the N binary one-against-all classification problems because they compared favorably to Support Vector Machines [24] and K-Nearest Neighbors as we will see in Section 4.2. Each Random Forest is trained in regression mode assigning a 1 as desired output if the provided descriptor is from the category the classifier is trained on and assigning a 0 if the sample originates from any other class. Queried with a sample $\mathbf{x}$, the Random Forest for class k will output a number $w_k = w_k(\mathbf{x})$ between 0 and 1. We define a decision boundary $\theta_i \in [0, 1], i = 1, \ldots, N$, for each of the N Random Forest classifiers which allows to interpret the outputs w_k: we say that classifier k asserts that sample $\mathbf{x}$ belongs to its class if $w_k \geq \theta_k$. Dependent on the value of the decision boundary θ_k we obtain different results regarding

the true positive rate ρ_{tpr} (recall): $$\rho_{\mathrm{tpr}} = \frac{h_{\mathrm{tp}}}{h_{\mathrm{tp}} + h_{\mathrm{fn}}} \quad , \tag{14}$$

the false positive rate ρ_{fpr}: $$\rho_{\mathrm{fpr}} = \frac{h_{\mathrm{fp}}}{h_{\mathrm{fp}} + h_{\mathrm{tn}}} \quad \text{and} \tag{15}$$

the precision ρ_{pr}: $$\rho_{\mathrm{pr}} = \frac{h_{\mathrm{tp}}}{h_{\mathrm{tp}} + h_{\mathrm{fp}}} \tag{16}$$

of the classifier. Here we denote the number of true positive classifications h_{tp}. These are positive samples which were actually classified positive. The number of false positive classifications is called h_{fp}, the number of false negative classifications h_{fn} and the number of true negative classifications h_{tn}.

We set each θ_i to the optimal value suggested by the performance on the training set according to the following measure:

$$\theta_i = \arg \min_{t \in [0,1]} d_{\mathrm{ROC}} \left(\rho_{\mathrm{tpr},i}(t), \rho_{\mathrm{fpr},i}(t) \right) \tag{17}$$

$$d_{\mathrm{ROC}}(\rho_{\mathrm{tpr}}, \rho_{\mathrm{fpr}}) = \sqrt{(1 - \rho_{\mathrm{tpr}})^2 + \rho_{\mathrm{fpr}}^2}. \tag{18}$$

Hence, we are searching for the θ_i that minimize the distance of the ROC plot to the optimal point with $\rho_{\mathrm{tpr},i} = 1$ and $\rho_{\mathrm{fpr},i} = 0$ for each class i. If we search for an object of class i and the corresponding classifier outputs a value greater or equal than θ_i when provided with a descriptor $\mathbf{x}$ we consider the sample as belonging to the queried class i.

However, this method does not work properly if the category of a sample has to be determined since in some cases more than one binary classifier might assert that sample $\mathbf{x}$ belongs to its class. Therefore, we convert the output w_k of each classifier $k = 1, \dots, N$ into a likelihood $L(a_k|\mathbf{x})$ of descriptor $\mathbf{x}$ belonging to class k. We define

$$L(a_k|\mathbf{x}) = \frac{\mathrm{e}^{\alpha \cdot m_k(\mathbf{x})}}{\mathrm{e}^{\alpha \cdot m_k(\mathbf{x})} + \mathrm{e}^{-\alpha \cdot m_k(\mathbf{x})}} \quad , \tag{19}$$

$$m_k(\mathbf{x}) = \begin{cases} \frac{w_k(\mathbf{x}) - \theta_k}{1 - \theta_k} \ , & w_k(\mathbf{x}) \geq \theta_k \\ \frac{w_k(\mathbf{x}) - \theta_k}{\theta_k} \ , & w_k(\mathbf{x}) < \theta_k \end{cases} \qquad \forall k = 1, \dots, N \quad . \tag{20}$$

Mapping $m_k(\mathbf{x})$ maps the output value $w_k(\mathbf{x})$ piecewise linearly from the range $[0, \dots, \theta_k, \dots, 1]$ to $[-1, \dots, 0, \dots, 1]$ where the decision threshold θ_k is mapped to 0. Then for all positive decisions of classifier k, that is $w_k(\mathbf{x}) \geq \theta_k$, it holds that $m_k(\mathbf{x}) \geq 0$ and for all negative decisions we have $m_k(\mathbf{x}) < 0$. Equation (19) is inspired by the conversion of the tree ensemble output to a probability in AdaBoost [25]. This function maps negative values of $m_k(\mathbf{x})$ to low probabilities, positive values of $m_k(\mathbf{x})$ to high probabilities and assigns uncertainty, that is 0.5, if $m_k(\mathbf{x})$ equals 0.Whereas the potentially different values of the decision thresholds $\theta_k, k = 1, \dots, N$, prevent a direct comparison of the outputs $w_k(\mathbf{x})$, the conversion as shown renders the certainties $L(a_k|\mathbf{x})$ of different classifiers comparable. Parameter $\alpha > 0$ in equation (19) is a scale factor which defines the slope of the mapping and the minimally $\frac{\mathrm{e}^{-\alpha}}{\mathrm{e}^{-\alpha} + \mathrm{e}^{\alpha}}$ and maximally $\frac{\mathrm{e}^{\alpha}}{\mathrm{e}^{\alpha} + \mathrm{e}^{-\alpha}}$ possible probability. α must be carefully adjusted to distribute the occurring $w_k(\mathbf{x})$ over the whole range of probabilities.

The next section details how we can obtain a category decision from the likelihoods of the binary classifiers with respect to their reliabilities.

3.4 Extension for Multi-class Categorization

Most of the popular classification methods of machine learning are essentially distinguishing between two classes. Several approaches exist for a multi-class extension of those binary classifiers, like one-against-all or one-against-one schemes. In one-against-all solutions, there exists one basic classifier for each class which discriminates a single class against the remainder of classes. The decision for a certain class is found either by choosing the result of the classifier with the highest certainty or by constructing a decision cascade beginning with the strongest classifier [26]. Another approach is the one-against-one scheme which contains a basic classifier for each pair of classes. The decision is determined by collecting the votes of all these classifiers. For the reasons discussed in Section 3.3 we decided for the one-against-all scheme whose size only grows linearly with the number of classes.

Furthermore, each of the N one-against-all classifiers outputs a certainty $L(a_k|\mathbf{x})$ that descriptor $\mathbf{x}$ belongs to class k as explained in Section 3.3. One could be tempted to classify a descriptor with the respective class of the classifier yielding the highest certainty. However, this can easily lead to wrongly biased multiclass decisions since the individual reliabilities of the classifiers are not considered. A simple example illustrates the problem: Suppose the classifier for class j is outputting a probability of 1.0 for every sample so that we would always choose it. Nevertheless, this classifier is only correct in those few cases when the sample is indeed from class j. That is visible on the low precision of classifier j, which indicates that only few of the positive outputs are indeed correct. The multiclass performance of this classifier would not be better than guessing. This finding suggests that we have to incorporate the precision of each classifier, which is a measure of reliability.

Consequently, we apply a probabilistic decision scheme which outputs a probability distribution for the class to choose. The proposed scheme incorporates the different reliabilities of the classifiers in a principled way. Assume there are N different binary classifiers, Random Forests in our example, each for one of the N classes of objects. Presented with a data sample $\mathbf{x}$, their outputs are the likelihoods $L(a_1|\mathbf{x}), \ldots, L(a_N|\mathbf{x})$, where $L(a_k|\mathbf{x})$ stands for the certainty of classifier k that $\mathbf{x}$ is a sample of its class (see Section 3.3). Applying the formula for total probability, the probability that sample $\mathbf{x}$ is a descriptor of class o_i can be expressed as

$$p(o_i|\mathbf{x}) = \sum_{k=1}^{N} p(o_i|a_k, \mathbf{x}) p(a_k|\mathbf{x}) \quad . \tag{21}$$

We approximate

$$p(o_i|a_k, \mathbf{x}) \approx p(o_i|a_k) = \frac{p(a_k|o_i)p(o_i)}{p(a_k)} \tag{22}$$

The decision accuracy term $p(a_k|o_i)$, which describes the probability that the binary classifier of class k considers a sample positive when it is actually a sample of class i, is determined from cross-validating the binary classifiers while the class frequency prior $p(o_i)$ is set to a uniform distribution but could also be obtained from the training dataset. Having these distributions we can calculate $p(a_k) = \sum_{i=1}^{N} p(a_k|o_i)p(o_i)$. The prior $p(a_k|\mathbf{x}) = \beta L(a_k|\mathbf{x})$ is proportional to the certainties $L(a_k|\mathbf{x})$. However, as we are only interested in the object class $\hat{o} = \arg\max_{o_i, i=1,\ldots,N} p(o_i|\mathbf{x})$ with highest probability, we do not have to compute β. In summary, we determine the multiclass decision $\hat{o}$ as

$$\hat{o} = \arg \max_{o_i, i=1,\ldots,N} \sum_{k=1}^{N} \frac{p(a_k|o_i)p(o_i)}{p(a_k)} L(a_k|\mathbf{x}) \quad . \tag{23}$$

This approach allows to weight the probabilities with which the classifiers are voting for their class with the reliabilities of these classifiers.

Table 1. Number of objects captured from each class in the IPA-1 and IPA-2 database

IPA-1		IPA-2			
class	# objects	class	# objects	class	# objects
ball	3	binder	10	dishliquid	9
book	10	book	11	drink carton	9
bottle	10	bottle	10	computer mouse	8
coffeepot	7	can	10	pen	10
cup	10	coffeepot	10	scissors	5
drink carton	4	cup	10	screen	10
flowerpot	7	dishes	10	silverware	29
plush toy	3				
toy car	3				

4 Evaluation

This section evaluates the categorization performance of the SAP descriptor and compares the results with other approaches. We additionally demonstrate the properties of the SAP descriptor concerning computation time and invariance to camera distance and rotations. All experiments measuring categorization rates were carried out with a 10-fold randomized leave-out-one cross-validation in which we left one randomly chosen object for the test set that did not occur within the training data. Categorization rates are the ratios of correctly classified views of the test objects to the total number of views from test objects. Categorization rates are computed individually for each class and reported as the average over all classes. All results were determined using the following databases.

4.1 Datasets

We recorded two datasets with 9 and 14 classes of household objects, respectively, according to the procedure described in Sec. 3.1. The dataset with 57 objects from 9 classes is called set IPA-1. For this set, each object was placed on a rotary table and captured 72 times yielding consecutive views in 5° steps. The Swissranger depth camera was mounted at the height of a robot viewing slightly downwards onto the table. The average number of points per segmented object is 6144. The left column of Table 1 summarizes the classes contained in the database as well as the number of object instances captured.

The second dataset, named IPA-2, was captured with a PMD CamCube and contains 36 views per object. The average number of points per object is 26491. The middle and right columns of Table 1 provide an overview over the distribution of the 151 objects into the 14 classes. A detailed description of the IPA-2 object database can be found in [27]. This set is publicly available at `http://www.kyb.mpg.de/nc/employee/details/browatbn.html`.

4.2 Results

We compare the performance of our categorization framework and the SAP descriptor on the larger IPA-2 database with the Global Radius-based Surface Descriptor (GRSD), the Global Fast Point Feature Histogram (GFPFH) descriptor and the Viewpoint Feature Histograms (VFH) descriptor, which have been applied to similar tasks. For computing the latter three descriptors, we used the implementations of the PCL library [19] and implemented the supplementary preprocessing according to the descriptions in the original papers [2,1,3]. Additionally, the point clusters of the objects were centered in front of the camera before any data processing so that no random deviations of the camera viewpoint could diminish the descriptive power of these descriptors. We found that this measure increased the recall rate of VFH by almost 2 %, for example. Then, the point clouds were downsampled with a voxel filter of leaf size 1.5 cm (GRSD, GFPFH) or 0.5 cm (VFH), respectively, to speed up the following normal estimation. Finally, we called the functions which compute the respective descriptors from the point cloud and its normals. For accumulating the local RSD features, we utilized the GFPFH-function as this is almost exactly the algorithm which also computes GRSD and because no other implementation was available. The labels for each voxel were estimated with the `getSimpleType()` function, which appears to realize the method described in the original paper. As we do not have point-wise labels in our datasets, we could not train the voxel labels w.r.t. to the FPFH descriptors as explained in [1]. Instead, we clustered the FPFH descriptors with k-means into 5 classes and used these classes as labels for the GFPFH computation. Using more clusters resulted in a decrease in performance.

In the following, the naming scheme for the variants of SAP descriptors is SAP-n_x-n_y-n_p, where n_x, n_y describe the number of cuts along the x- and y-coordinate axes and n_p denotes the degree of the approximating polynomial. The multi-class classification rates and their standard deviations on dataset IPA-2 are compiled in Table 2 for all tested descriptors. For the sake of completeness we also cite the results from Browatzki et al. [27] who conducted similar experiments on this database. We can observe that the SAP descriptor appears to be more powerful for categorization problems than VFH, GRSD or GFPFH with an increase in multi-class categorization performance of 9.5 % to 23.5 %. SAP also performs 5 % better than the best of the four descriptors tested in [27]. For SAP, GRSD and GFPFH we report the results when using the categorization framework of this paper whereas the VFH descriptors are categorized with a Support Vector Machine.

Table 3 summarizes the classification rates of these descriptors with respect to the used classifier. We compare a K-Nearest Neighbors (KNN) classifier with $k = 1$, a multi-class Support Vector Machine (SVM)[1] and our proposed framework with Random Forests. It shows that our method can almost always attain the top performance. Only with the VFH descriptor the multi-class SVM is 1.1 % better. The results for KNN do not improve when the number of considered neighbors

[1] Both classifiers as implemented in OpenCV [28], however the SVM originates from libsvm [29].

Table 2. Comparison of several descriptors regarding multi-class categorization performance, average computation time per view and average throughput in points per second. The evaluation was carried out on the IPA-2 database.

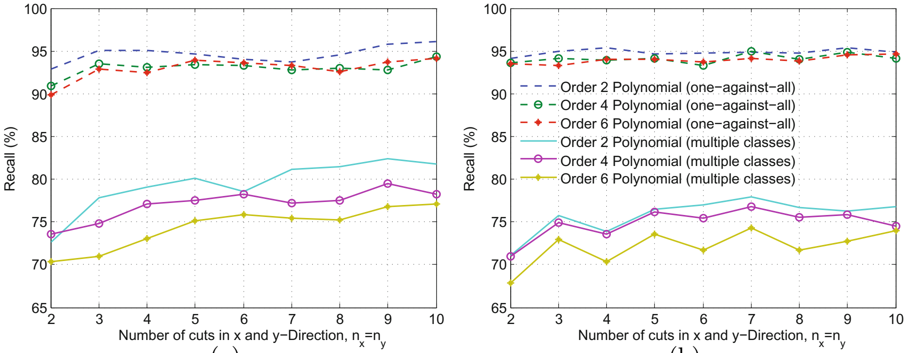
Descriptor	Performance	Computation Time	Throughput
Shape Distributions [27]	25.4 %	31 ms	∼ 855 000 pts/s
Shape Index [27]	34.6 %	78 ms	∼ 339 000 pts/s
Shape Context 3D [27]	55.2 %	234 ms	∼ 113 000 pts/s
Depth Buffer [27]	72.9 %	**16 ms**	**∼ 1 656 000 pts/s**
GFPFH	54.4±6.2 %	921 ms	28 928 pts/s
GRSD	56.1±5.8 %	957 ms	27 841 pts/s
VFH	68.4±6.7 %	93 ms	205 883 pts/s
SAP-7-7-2	**77.9±5.5 %**	57 ms	338 534 pts/s

Table 3. Comparison of different classifiers for categorizing the objects from the IPA-2 database with several descriptors

Classifier	KNN ($k = 1$)	SVM	Random Forests
Performance with SAP-7-7-2	55.4 %	45.5 %	**77.9 %**
Performance with VFH	60.4 %	**68.4 %**	67.3 %
Performance with GRSD	56.0 %	51.5 %	**56.1 %**
Performance with GFPFH	51.6 %	46.8 %	**54.4 %**

k is increased. The SVM uses a one-against-one multi-class extension. It was trained with automatic parameter tuning through a 10-fold cross-validation. The margin between the three classifiers is by far the largest with the SAP-7-7-2 descriptor. The reason for this effect is the inhomogeneous descriptor which consists of an absolute size measure, two relative size measures and parameters of polynomials. Random Forests do not require the input data to be normalized to a common magnitude. Consequently, they work well with the data we provide. However, we would need to construct an adequate metric for KNN or SVM since these classifiers rely on normalized data.

Besides the categorization performance we also report average computation times and throughput for the descriptor computation from a single view, including necessary preprocessing. To avoid biased representations the computation time for the classifier is not included. Nevertheless, our Random Forests-based approach evaluates very fast with only 8 ms on average for the SAP-7-7-2 descriptor (45-dimensional). The results for the descriptors Shape Distributions, Shape Index, Shape Context and Depth Buffer were determined by Browatzki et al. [27] and are obtained on a 3 GHz DualCore machine with 2 GB RAM. We estimated the throughput for these values. All code was written in C++. The computation times for GFPFH, GRSD, VFH and SAP were determined on a mobile Intel I7 2.8 GHz Processor using only a single core. It shows that the computation time of the SAP descriptor is almost four times slower than the top performer Depth Buffer, however at the gain of 5% better categorization



(a) (b)

Fig. 3. Comparison of different configurations of the SAP descriptor with varying numbers of surface cuts and degrees of the approximating polynomials (a) on IPA-1 database and (b) on IPA-2 database

rates. The runtime of SAP allows to compute the descriptor with over 17 Hz. That is SAP could classify up to 17 objects in a scene within one second. VFH suffers from a slightly longer computation time because of the preceding normal computation. GFPFH and GRSD can only categorize one object per second.

4.3 Parameters and Properties of the SAP Descriptor

As explained in Section 3.2 the SAP descriptor has three parameters which are supposed to have a significant influence on its descriptive power and cannot be trivially chosen. Therefore, we examine the categorization performance with respect to the numbers of cuts n_x and n_y parallel to the x- and y-axes and with respect to the degree n_p of the approximating polynomial. The number of cuts is always kept equal for both dimensions, that is $n_x = n_y$, since objects can have the same extent in both directions. It is not possible to reduce n_y for slim objects because the employed classifiers expect descriptors of fixed length. Furthermore, it is not suitable to divide a constant total number of cuts to variable numbers n_x and n_y either, as this approach means comparing parts of the descriptors which contain spatially unrelated data for different objects or at least for objects from different classes.

The dependency of classification performance on these parameters is illustrated in Figure 3(a) for the IPA-1 dataset and in Figure 3(b) for the IPA-2 dataset. Both diagrams report results for the binary classification problem of separating one object class against the others as well as for the multi-class labeling task where each object view has to be assigned one of the class labels. The general trend that polynomials with higher degree n_p cause a lower categorization performance becomes evident in the binary and multi-class case. Manual inspection of the descriptors suggests that higher order polynomials are less stable and tend to model the noise from the sensor. As to expect, increasing the number of cuts allows the descriptors to capture more details and improves the categorization results. Furthermore,

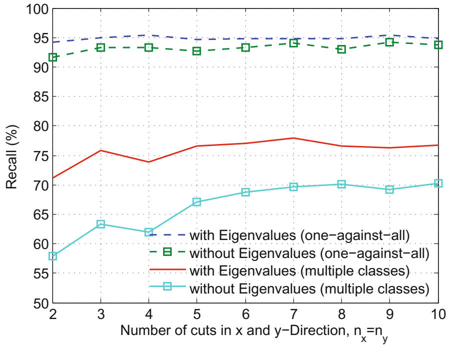
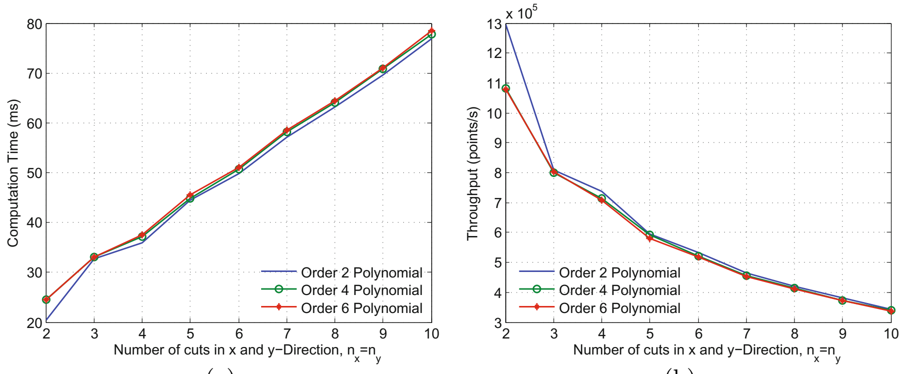


Fig. 4. Recall rates for the SAP descriptor with varying numbers of surface cuts and a polynomial degree of 2 when the descriptor either contains the size information from the PCA eigenvalues. Evaluated on the IPA-2 database.

dataset IPA-1 contains less object classes than IPA-2 and consequently, the recall rates are higher with fewer classes. Finally, we can conclude that SAP-7-7-2 is a reasonable choice considering the categorization performance on both datasets and the computation time. Therefore, we selected this configuration as standard for the comparison to other descriptors and within the following experiments revealing further properties of SAP.

Besides a suitable configuration we also need to know whether it makes sense to concatenate the absolute and relative size information from PCA with the polynomial parameters to form the SAP descriptor. Thus, we analyzed both components of the SAP descriptor alone. We found that using only the three values of the size component the categorization performance decreases to 62.2 %. On the other hand, there is a similarly significant drop in the recall rates if only the polynomial parameters appear in the descriptor as Figure 4 indicates. Consequently, combining both cues in the SAP descriptor proves to be resonable.

Next, the influence of the parameters of the SAP descriptor on the computation time shall be dissected. Figure 5(a) displays the average computation times for the computation of a SAP descriptor from one object view for increasing numbers of n_x, n_y and n_p. As the theoretical analysis in Section 3.2 predicts there is a linear increase in computation time with rising numbers of surface cuts. However, the influence of the degree of the polynomials is less visible because of the very small differences in computation time. We therefore suppose, that the SVD for polynomial fitting has a significantly lower impact on the computation time than the effort for assigning points to the cuts, which is also linear with $n_x + n_y$ but independent of n_p. We also display statistics about processed model points per second in Figure 5(b) to provide a measure which is independent of the number of points per object view. The general trend coincides with the computation time result as throughput behaves essentially reciprocal to computation time: the more surface cuts and the higher the degree of the polynomials, the lower the throughput.



(a)

(b)

Fig. 5. Dependency of (a) computation time and (b) throughput of the SAP descriptor for increasing numbers of surface cuts and polynomial degrees measured on set IPA-2

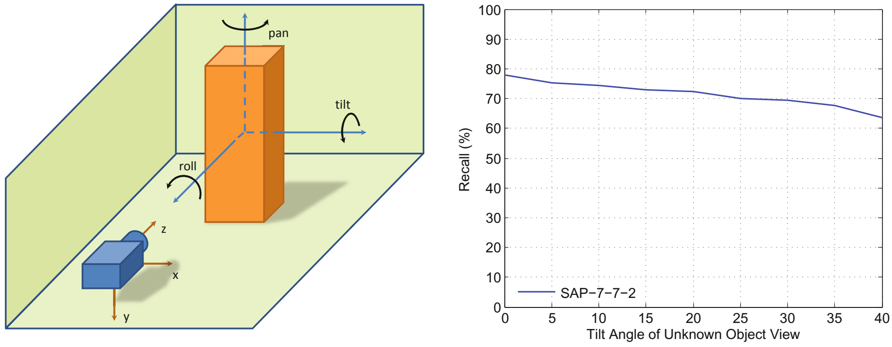
Table 4. Effect of lower resolution point cloud data on the SAP-7-7-2 descriptor. The resulting recall rates are a measure of robustness to scale changes.

Percentage of Points	100 %	50 %	25 %	10 %	6.25 %	4 %
Camera Distance Factor	1	1.4	2	3.2	4	5
Performance	77.9 %	78.2 %	77.1 %	74.9 %	74.8 %	73.6 %

4.4 Scanning Distance and Rotation Invariance

The last part of the evaluation deals with the robustness or invariance of the SAP descriptor against common transformations of objects. First, we analyze the invariance against scale changes. This happens when the camera moves closer to the object or farther away from it. Although the range sensors still capture the real size of the objects because they provide metric measurements, the sampling density of the point clouds decreases quadratically with the distance to the camera. A consequence is that noisy pixels can have more impact since their percentage of the measured points increases. SAP is computing regressions over many points and should therefore naturally expose a high robustness to scale changes. We simulated scale changes by randomly sampling decreasing amounts of points from the original depth data. The recall rates reported in Table 4 indicate that SAP has indeed a high scale invariance. Please notice that sampling 25 % of the original points corresponds to doubling the distance to the camera and sampling 10 % is approximately the triple distance. Up to this distance the categorization performance does not decrease more than 3.0 %.

Rotations are another important kind of transformation that regularly occur between objects in the real world and the camera. In Figure 6(a) we define three kinds of basic rotations: pan, tilt and roll. Arbitrary rotations consist of these three basic rotations. In the following, we examine to which extent SAP can handle each of of them.



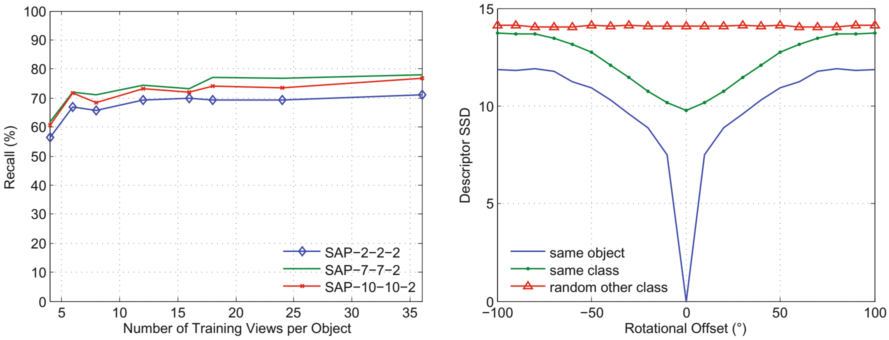
(a) (b)

Fig. 6. Analysis of the rotational robustness of the SAP descriptor: (a) definition of the rotational axes and (b) robustness with respect to tilting rotations

First, we evaluate the robustness to tilting rotations of the object. This kind of rotation occurs for example when the camera watches the object from a different height and angle. To gauge the robustness of the SAP descriptor against tilting rotations, we trained the categorization system with the original data from the IPA-2 database and the same data from every object view tilted by angle α against the camera. Then we measured the categorization performance on object views of objects outside the training set which were tilted by angle $\alpha/2$. This way we can predict how many different tilt angles have to be present in the data of training objects to allow for successful categorization at the intermediate tilt angles. In Figure 6(b) the recall rates are plotted against the tilt angles of the test data. We can see that SAP can still categorize 73.0 % of the test object views, which were tilted by 15°, while the training set only contained object views at tilt angles of 0° and 30°. Consequently, it would suffice to capture object views at different tilt angles every 30° of training instances to successfully model a class.

A similar analysis can be done for pan rotations. As we already have 36 views of each object around the pan direction in the IPA-2 database, an analysis about the rotational stability around the pan axis can be conducted by excluding more and more views from the training set. Testing is done with all views. Figure 7(a) shows the relation between utilized training views and recall rates. We learn that 18 views are enough to maintain a high recall rate of 77.0 %. This corresponds to capturing depth images of the training objects in a pan distance of 20°.

Fig. 7(b) furthermore shows the results of descriptor repeatability tests on the IPA-2 database which underline that the descriptor's similarities of neighboring viewing angles are very high. Similarity between two descriptors c_1 and c_2 is measured as the sum of squared differences (SSD) $SSD = \|c_1 - c_2\|^2_{L_2}$. Moreover, with increasing angular offset the SAP descriptors are still much more similar to the original object than to objects of any another class and also very similar to descriptors from other objects of the same class.



(a) (b)

Fig. 7. Analysis of the rotational robustness of the SAP descriptor along the pan axis: (a) recall analyzed against the number of equally distributed views of the training objects for three configurations of the SAP descriptor. (b) Averaged sum of squared differences between SAP-7-7-2 descriptors obtained from different viewing angle offsets for descriptors originating from the same object, objects of the same category and random non-class objects.

SAP, as introduced in Section 3.2, has no means to compensate roll rotations. This implies that SAP can only recognize instances of the learned classes as long as they are standing in a similar upright position as the training objects. Of course, this is not generally the case in reality. Therefore, we devised a rotational transformation which precedes the SAP computation. This transformation compensates roll rotations of the captured object by projecting the 3D points into the image plane and computing a repeatable orientation. Then the point cloud is rolled to a canonical orientation so that the following SAP computation always runs on a well-adjusted point cloud. Several experiments showed the success of this idea with recall rates around 77.0 %. A paper about the analysis of this extension of the SAP descriptor is in preparation.

5 Conclusions

In this paper we introduced the novel descriptor SAP for categorizing simple household objects. SAP directly computes global features on possibly noisy point cloud data. We showed that if SAP descriptors are sampled appropriately from different views of training objects, SAP can obtain very good categorization results of 78% within a short computation time of 57 ms per computation. These results compare favorably to GFPFH, GRSD and VFH. Further experiments proved the invariance of the SAP descriptor to scale, to tilt rotations up to ±15° and to pan rotations up to ±10°. These results provide useful suggestions how to sample views from the training objects to ensure a complete coverage.

We also discussed an extension for full rotation invariance around the camera axis briefly. The results of our experiments are encouraging. Thus, future work will be devoted to a careful evaluation of this approach. Furthermore, we plan to analyze whether it is possible to create some artificial views automatically to decrease the number of views which have to be captured from training objects. Finally, the SAP descriptor shall be tested with the Kinect depth sensor on numerous real world scenes.

Acknowledgments. This research was partly funded from the EU FP7-ICT-287624 Acceptable robotiCs COMPanions for AgeiNg Years (ACCOMPANY).

References

1. Rusu, R.B., Holzbach, A., Beetz, M., Bradski, G.: Detecting and segmenting objects for mobile manipulation. In: ICCV, S3DV Workshop (2009)
2. Marton, Z.C., Pangercic, D., Blodow, N., Beetz, M.: Combined 2D-3D Categorization and Classification for Multimodal Perception Systems. The International Journal of Robotics Research 30(11), 1378–1402 (2011)
3. Rusu, R.B., Bradski, G., Thibaux, R., Hsu, J.: Fast 3D recognition and pose using the viewpoint feature histogram. In: Proceedings of the International Conference on Intelligent Robots and Systems (IROS), Taipei, Taiwan (2010)
4. Galleguillos, C., Belongie, S.: Context based object categorization: A critical survey. Computer Vision and Image Understanding (CVIU) 114, 712–722 (2010)
5. Zhang, H., Berg, A.C., Maire, M., Malik, J.: SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2126–2136 (2006)
6. Deng, J., Berg, A.C., Li, K., Fei-Fei, L.: What Does Classifying More Than 10,000 Image Categories Tell Us? In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 71–84. Springer, Heidelberg (2010)
7. Johnson, A., Hebert, M.: Using spin images for efficient object recognition in cluttered 3d scenes. IEEE Trans. PAMI 21(1), 433–449 (1999)
8. Koenderink, J.J., van Doorn, A.J.: Surface shape and curvature scales. Image Vision Computing 10, 557–565 (1992)
9. Knopp, J., Prasad, M., Willems, G., Timofte, R., Van Gool, L.: Hough Transform and 3D SURF for Robust Three Dimensional Classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6316, pp. 589–602. Springer, Heidelberg (2010)
10. Tombari, F., Salti, S., Di Stefano, L.: Unique Signatures of Histograms for Local Surface Description. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 356–369. Springer, Heidelberg (2010)
11. Knopp, J., Prasad, M., Van Gool, L.: Orientation invariant 3D object classification using hough transform based methods. In: Proc. of the ACM Workshop on 3D Object Retrieval, pp. 15–20 (2010)
12. Toldo, R., Castellani, U., Fusiello, A.: A bag of words approach for 3D object categorization. In: Proc. of Int. Conference on Computer Vision, pp. 116–127 (2009)
13. Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D.: Shape distributions. ACM Tr. on Graphics 21(4), 807–832 (2002)
14. Kazhdan, M., Funkhouser, T., Rusinkiewicz, S.: Rotation invariant spherical harmonic representation of 3D shape descriptors. In: Symposium on Geometry Processing (June 2003)

15. Pu, J., Yi, L., Guyu, X., Hongbin, Z., Weibin, L., Uehara, Y.: 3D model retrieval based on 2D slice similarity measurements. In: Proceedings of the 3D Data Processing, Visualization, and Transmission, pp. 95–101 (2004)
16. Endres, F., Plagemann, C., Stachniss, C., Burgard, W.: Unsupervised discovery of object classes from range data using latent dirichlet allocation. In: Proc. of Robotics: Science and Systems (2009)
17. Bo, L., Ren, X., Fox, D.: Depth Kernel Descriptors for Object Recognition. In: IROS (September 2011)
18. Wahl, E., Hillenbrand, U., Hirzinger, G.: Surflet-pair-relation histograms: A statistical 3D-shape representation for rapid classification. In: 3-D Digital Imaging and Modeling, pp. 474–481 (2003)
19. Rusu, R.B., Cousins, S.: 3D is here: Point Cloud Library (PCL). In: Proc. of Int. Conference on Robotics and Automation (ICRA), Shanghai, China (2011)
20. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24, 381–395 (1981)
21. Marton, Z.C., Rusu, R.B., Jain, D., Klank, U., Beetz, M.: Probabilistic Categorization of Kitchen Objects in Table Settings with a Composite Sensor. In: Proc. of the Int. Conf. on Intelligent Robots and Systems, St. Louis, MO, USA (2009)
22. Collet Romea, A., Srinivasa, S., Hebert, M.: Structure discovery in multi-modal data: a region-based approach. In: Proceedings of ICRA (2011)
23. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
24. Burges, C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121–167 (1998)
25. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Annals of Statistics 28, 2000 (1998)
26. Mozos, O.M., Burgard, W.: Supervised learning of topological maps using semantic information extracted from range data. In: IROS, pp. 2772–2777 (2006)
27. Browatzki, B., Fischer, J., Graf, B., Bülthoff, H., Wallraven, C.: Going into depth: Evaluating 2D and 3D cues for object classification on a new, large-scale object dataset. In: Proc. of Int. Conf. Computer Vision Workshop on CD4CV, pp. 1–7 (2011)
28. Bradski, G., Kaehler, A.: Learning opencv: Computer vision with the opencv library (2008)
29. Chang, C., Lin, C.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011)

Online Semantic Mapping of Urban Environments

Nikos Mitsou[1,2], Roderick de Nijs[1], David Lenz[1], Johannes Frimberger[1], Dirk Wollherr[1], Kolja Kühnlenz[1], and Costas Tzafestas[2]

[1] Institute of Automatic Control Engineering (LSR), Technische Universität München, D-80290 Munich, Germany
{nmitsou,rsdenijs,david.lenz,jfrimberger,dw,koku}@tum.de
[2] School of Electrical and Computer Engineering, Division of Signals, Control and Robotics National Technical University of Athens, Greece
ktzaf@cs.ntua.gr

Abstract. In this paper we present an integrated approach for efficient online 3D semantic map building of urban environments and the subsequent extraction of qualitative spatial relationships between the different objects in the scene. We split this process into three stages, where we combine a state of the art image segmentation and classification algorithm with an online clustering algorithm to obtain a coherent representation of the environment. Finally, a graph representation is extracted which can then be used for spatial reasoning and human robot interaction. We present first results from data collected by a mobile robot which operates in city areas.

1 Introduction

The ability to recognize and understand the environment is a major goal for robotic intelligence. The development of Simultaneous Localization and Mapping (*SLAM*) techniques has enabled robots to construct metric and topological representations of indoor and outdoor environments while keeping track of their own position [1,2]. However, more sophisticated environment models are required for the creation of intelligent and fully autonomous systems. Apart from storing information about occupied and free areas [3] or topological connections between areas in the environment [4], nowadays semantic maps have been introduced that contain additional layers of information such as the type of rooms, objects found, spatial relations between those objects [5], etc.[1]

The advantages of the presence of semantics in the robot internal representation of the environment are manifold. The possibility for high-level reasoning and inference is opened, allowing for the creation of more versatile and intelligent components, such as efficient task planning [6]. Moreover, understanding

[1] There are some inconsistencies in the literature on the use of the concepts of "*semantic*" and "*semantic mapping*". In this paper, the term *semantic* is used to refer to an entity (e.g. map or label) that contains meaningful information to humans.

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 54–73, 2012.

the environment on a similar semantic level as humans enables robots to communicate and exchange information with humans in an easier way. This can be used as an additional means of verifying and gathering information [7]. Robots can acquire this higher level knowledge about their environment either through a human expert or by processing the available sensor data, for which reliable inference algorithms are required. As an example, depth information in the form of point clouds and appearance information from cameras can be used by robots to construct semantically annotated maps.

In this work, we consider a robot exploring an urban environment and collecting images through its cameras. The purpose of the robot is to generate a 3D semantic map of the environment. In contrast to many existing works where the full data is available from the start, here we build the map incrementally in an online setting. The proposed system starts with classifying the regions of the incomming images and projecting the labeled pixels into the space using the stereo reconstruction. A probabilistic grid is used to efficiently store the labeled point clouds, which are then processed by an online clustering algorithm that extracts surfaces from the constructed 3D grid and generates objects in the environment. While this map is being generated, the discovered entities are organized into a graph structure with labeled relationships between the entities. In brief, our contributions are as follows: we propose a framework for 3D semantic mapping based on image classification, object extraction and spatial reasoning and we utilize state of the art solutions for each corresponding part. Except for the accuracy in the results, focus has been also given into the efficiency of the system with performance close to real time in our experimental settings.

The paper is organized as follows: In Section 2, an overview of the proposed architecture is provided. Section 3 presents a brief survey of related work organized by the different components of the architecture. Sections 4, 5 and 6, give detailed descriptions of the different modules. Experimental results are presented in Section 7 and a discussion and an outlook are provided in Section 8.

2 System Overview

The architecture of the proposed system is presented in Fig. 1. As inputs, it receives an image from the robot's camera, a disparity map of this image and the robot odometry measurements. After processing this data, the system outputs a 3D semantic map of the environment with spatial relations between the discovered objects.

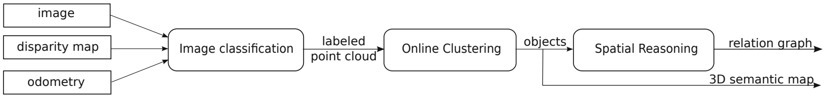


Fig. 1. Overview of main system components and their connections

Our system consist of three main components:

- *Image segmentation and classification component*: Incoming images are split into visually similar regions and a Conditional Random Field (CRF) is used to classify these regions with the aid appearance and 3D features from stereo cameras. The resulting data is projected back into the localization frame of the robot as a labeled point cloud.
- *Object extraction component*: The goal of this component is to extract objects from the sequence of the labeled point clouds in an online fashion. A naive accumulation of these results can lead to erroneous maps due to noise in both the spatial properties and the labels of the point clouds. A probabilistic grid is constructed to provide spatio-temporal smoothing over the classes and an online clustering algorithm is applied for surface and object extraction.
- *Spatial reasoning component*: Qualitative spatial relations between objects are extracted and a graph structure of the environment is generated.

Since the focus of this paper is on mapping of urban environments, eight possible classes are considered: *sidewalk*, *building*, *tree*, *pedestrian*, *car*, *sky*, *street* and *grass*. These semantic classes are chosen because of their relevance to robot navigation and human robot communcation.

3 Related Work

Many different approaches have been proposed for automatically annotating the environment with semantic labels. For indoor environments and room labeling for instance, solutions that consider object detection [8,5], multimodal sensor classification [9] and classification of 3D point clouds [10,11] can be found. Examples for outdoor environments are terrain classification [12] or urban scene classification [13]. In order to store this knowledge, [14] proposes a hierarchical abstraction structure for semantic spatial knowledge.

Since our approach integrates multiple components involving classification and efficient data storage and reasoning, we review the related work of each component in the following subsections.

3.1 Image Segmentation

During the last years, there has been a growing interest in urban scene understanding. Most works focus on segmenting images into regions with a specific semantic class associated. In urban environments, many classes have little texture or share a similar appearance, making it hard for classifiers to discriminate between classes. From a computer vision perspective, this motivates the use of probabilistic models of increasing complexity to capture relationships between image regions. These models create a graph-like structure based on the image and construct planar, hierarchical [15], or higher order [16] random fields on this graph, of which the Maximum-a-posteriori (MAP) solution results in the optimal joint image labeling. From a robotics perspective, there has been interest in combining data from multiple sensors and modeling temporal relationships in the data. In [13], urban environments are classified using a temporal analysis of laser scans and camera images through a conditional random fields (CRF). Similarly, [17] use the same input and present a generative model combined with a Markov random field (MRF) for temporal smoothing. Also interesting, in [18], semantic classes and depth estimates are estimated jointly. Other works have as a main goal to infer the geometric properties of urban scenes, either using stereo [19] or monocular images [20]. In this work, we perform semantic image segmentation using appearance and 3D information, and use it as an intermediate step for efficiently generating 3D semantic maps.

3.2 Point Cloud Clustering

The clustering problem has been addressed in many contexts and by researchers of different fields [21] [22]. In the field of robotics, it has been extensively used for different purposes; one of them is for understanding and interpreting 3D point clouds. In [23] for example, the authors use outlier removal, clustering and segmentation in order to reconstruct an indoor environment. In their experiments, different point clouds are merged and their method is applied over the merged point cloud that describes the whole environment. In [24] and [25], two approaches for 3D semantic mapping of urban environments has been presented. The received point clouds are segmented and planes are extracted. The above mentioned methods use 3D laser range finders to generate accurate point clouds of the environment and apply clustering over the whole dataset of point clouds.

Lately, methods that perform clustering in an online fashion has been proposed. In [26] the authors use a triangular mesh for their spatial representation model and update it incrementally by processing only those points of a new scan that do not overlap with their existing models. In [27], an incremental segmentation algorithm for the merge and expansion of existing clusters has been proposed. Initially, all neighboring points of a new scan are found. If the neighboring points are already assigned to an existing cluster, the new point is also assigned to the same cluster. If more than one clusters are candidates for assignment, then a merging of these clusters is performed. In both the above techniques, 3D data is generated by rotating a 2D laser range finder.

3.3 Spatial Relations

Many different approaches for spatial relations generation have been proposed. In terms of qualitative spatial relations, the focus in literature lies mostly on relations like *disjoint*, *intersecting*, *contains*, etc in 2D or the generalization of these relations to 3D space [28]. In order to calculate these relations in the 2D space, in [29] projections of the objects to the x and y axis of both objects are considered to differentiate between the relations. [30] uses 3D spatial relations in order to reason about locations of objects and answer queries such as: *if object A is southwest of B and B is north of C, where can C be in relation to A?*

Spatial relations have been extensively used in the robotics field. One of the main reasons for building spatial relations among objects in the environment is for Human Robot Interaction purposes. In [31], for example, spatial relations were combined with natural language processing in order to make the robot capable of describing the positions of objects in the environment in terms of *left, right, somehow left*, etc. The inverse procedure, i.e. identifying objects based on these relations given in directions has been studied by [32]. In [33], the same purpose was fulfilled by using global directions such as east, west etc. In [34], the authors present heuristics to understand relations used in route descriptions by a human to the robot.

Recent literature mainly focuses on indoor applications. Especially relations like *in*, *on* are used to describe locations of objects. For example in [35], the authors calculate the mechanical support *on* based on distance and contact criteria. This idea is further expanded in [36], where the authors are using *in* and *on* to reason about chains of relations to reason about locations of objects. In the knowledge-base of [37], the same type of spatial relations are not calculated implicitly but only when there are queried to confirm or reject a hypotheses.

4 Image Segmentation and Classification

The goal of the classification model is to find the most likely label assignment for the different regions in an input image. It employs data acquired from cameras and an associated disparity map to compute 3D features and reproject the resulting classification into the world frame, which can be provided by any localization system, e.g.,SLAM.

4.1 Classification Model

The classification problem is formulated as finding the optimal configuration for a pairwise conditional random field (CRF), which was introduced by [38] under the context of natural language processing and later extended to the realm of image segmentation and classification by [39]. Given a graph $G = (V, E)$, and random variables Y_i associated to every $i \in V$ can take a values from y_i from a label set $\mathcal{L} := \{1, ..., l\}$ A CRF with pairwise interactions is a graphical model that assigns a probability to each label configuration $\boldsymbol{y} = (y_1, ..., y_n)$ given input $\boldsymbol{x}$ by

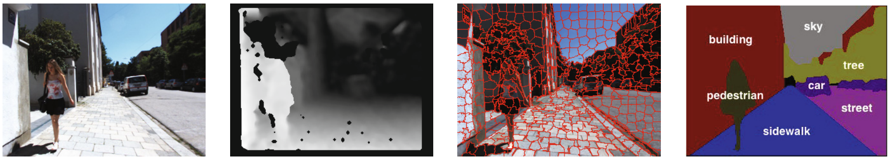
$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{w}) &= \frac{1}{Z(\boldsymbol{x};\boldsymbol{w})}\exp(-E(\boldsymbol{y},\boldsymbol{x};\boldsymbol{w})) \\ E(\boldsymbol{y},\boldsymbol{x};\boldsymbol{w}) &= \sum_{i\in V}\phi_i(y_i,\boldsymbol{x_i};\boldsymbol{w}) + \sum_{(i,j)\in E}\phi_{ij}(y_i,y_j,\boldsymbol{x_{ij}};\boldsymbol{w}), \end{aligned} \quad (1)$$

where $\phi_i(y_i, \boldsymbol{x_i}; \boldsymbol{w})$ is a unary cost and measures the compatibility of label y_i with the node feature vector $\boldsymbol{x_i}$, and $\phi_{ij}(y_i, y_j, \boldsymbol{x_{ij}})$ expresses the compatibility between the labels taken by the neighboring nodes Y_i and Y_j, given the edge features $\boldsymbol{x_{ij}}$. Obtaining the normalization term

$$Z(\boldsymbol{x};\boldsymbol{w}) = \sum_{\tilde{\boldsymbol{y}}}\exp E(\tilde{\boldsymbol{y}},\boldsymbol{x};\boldsymbol{w}),$$

requires summing over all possible assignments to $\boldsymbol{y}$, and ensures that the probability distribution integrates to one. The described formulation allows to include neighborhood relationships between nodes through the pairwise terms, assigning higher probabilities to labels that are also compatible with the neighboring labels. The goal is then to find the labeling $\boldsymbol{y}^*$ such that $\boldsymbol{y}^* = \text{argmax}_{\boldsymbol{y}}\, p(\boldsymbol{y}|\boldsymbol{x})$. Although solving this problem is hard, in practice good solutions can be found using loopy belief propagation or α-expansions, commonly used in computer vision.

Learning involves maximizing the log-likelihood of training data through iterative techniques such as stochastic gradient descent[40] and L-BFGS. Although $\phi_i(y_i, \boldsymbol{x_i})$ and $\phi_{ij}(y_i, y_j, \boldsymbol{x_{ij}})$ are linear functions, making the problem convex in the parameters $\boldsymbol{w}$, evaluating $Z(\boldsymbol{x};\boldsymbol{w})$ is computationally infeasible for large, loopy graphs. This forces to either manually set the parameters $\boldsymbol{w}$ or rely on approximations to $Z(\boldsymbol{x};\boldsymbol{w})$ that can be obtained from inference techniques such as max-sum loopy belief propagation or tree-reweighted belief propagation[41].



(a) Input image (b) Disparity map (c) Superpixel segmentation (d) Manually labeled image

Fig. 2. Image segmentation: Inputs, manually labeled image and segmented regions

Building the Graph. Since neighboring regions in an image normally share the same semantic class and doing inference on the 4-connected structure of the image pixels is very slow, it is efficient to group pixels together and reduce the computational load. For this purpose, we use a preprocessing algorithm to partition the image into patches - commonly named superpixels - while respecting the natural object boundaries, using some criteria. Various techniques, e.g. Normalized Cuts [42], Turbopixels [43] and SLIC superpixels [44], are openly available and can accomplish this task at various frames per second. It is desirable that the resulting regions group similar pixels but are not too greedy, as the superpixels might *leak* between classes. In this work we use SLIC because they are fast to compute and segment the image into small, uniform regions, while respecting boundaries.

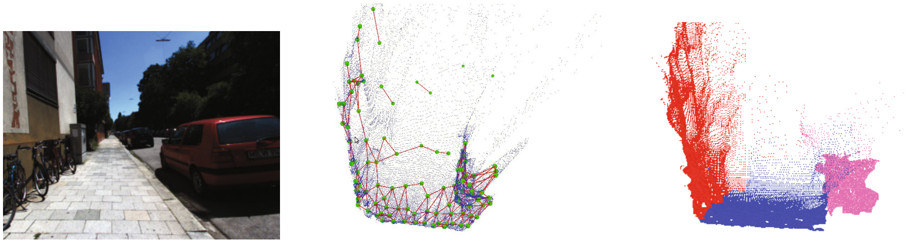
The graph G is built using the superpixel segmentation of the image together with the depth map. Specifically, a superpixel k will have a vertex or node $v \in V$ assigned if its depth value is known for every pixel. This eliminates noisy, far and infinite depth regions - like sky-. Furthermore, a pair of superpixels will share an edge $e \in E$ if they are part of V and their spatial distance (3D) from their centers is below some threshold value d. After this step, we have $|V|$ superpixels, each of which is linked to n_k pixels.

Unary and Pairwise Factors. The unary and pairwise terms in (1) assign a cost to every state by a linear combination of weights and features. Although this has the advantage that the log-likelihood of the training data will be a convex function of the weights, it suffers from only being able to model linear relationships between features and class probabilities. In order to incorporate non-linear dependencies between image features and class probabilities, we first use a local classifier for each node $v_i \in V$, $i = 1, 2, ..., |V|$ from image features and use the output class responses as the input $\boldsymbol{x_i}$ for the unary factors[1] and the appended class probabilities $\boldsymbol{x_{ij} = [x_i, x_j]}$ of each node for the pairwise factors. In practice, for the node potentials we will have, including bias, $f_u = |\mathcal{L}| + 1$ features and thus $p_u = f_u|\mathcal{L}|$ parameters. For the edges we have $f_e = 2|\mathcal{L}| + 1$ features and a parameter linking them to each class, giving $p_p = f_e|\mathcal{L}|$ parameters. Consequently, the CRF imposes a class-dependent smoothness, but does not learn which combination of distinct classes is more likely. We leave out modeling class co-occurences because, in contrast to other datasets, in urban scenarios the same classes appear in every image, and may neighbor or not depending on the perspepective. For the same reason, determining a class prior based on global image features, often named *context*, is not useful in the urban setting.

Local Classifier Features. The features are the most important step in image classification, as good features facilitate the job of the classifier. State of the art features generally combine strong low level local image descriptors, such as SIFT[45], with a specialized Bag-of-Words model to create a good representation for superpixels. Unfortunately, this approach is to slow for online settings, for which we propose to use a more efficient simple superpixel descriptor. On one hand we use a typical bank of filters consisting of mean, variance, Gaussian, Laplacian of Gaussian and derivative of Gaussian filters as well as a Histogram of Gradients. To speed up the calculation, following the technique presented in [46] these features are calculated very fast by approximating them through integral images. The responses of the features at every pixel are averaged over each superpixel. Additionally, a second set of features comes from the depth map delivered by stereo. These features are calculated from corresponding 3D points associated to each superpixel. These include the height above ground, the curvature and the projections of the normal vector onto the vertical and a horizontal plane. The features are chosen to be as viewpoint invariant as possible. For the same reason, a feature that captures the distance to camera path is left out, [47], since it is not invariant even under lateral displacements.

Output. After performing inference on the CRF, the resulting labels $\boldsymbol{y}^*$ are directly propagated from the superpixels to the linked pixels. Finally, the N

[1] The response of the local classifier for each class is in $\mathbb{R}$. Experiments with mapping the class-responses to $[0, 1]$ with a logistic function and normalizing for interpreting them as unary potentials decreased performance in our case For this reason, we simply use the responses directly as features for the CRF . However, [13] did have success with normalization.



(a) Input image (b) CRF adjacency structure (c) Labeled point cloud

Fig. 3. a) Shows input image. b) The adjacency structure used for the CRF through the connected green dots. c) A labeled point cloud, where each color represents a class.

pixels, with $N = \sum_{|V|} n_k$, are projected into the world coordinate frame of the robot as a labeled 3D point cloud $S = \{p_1, p_2, ...p_N\}$, with $p_i = \{p_x, p_y, p_x, label\}$, and where *label* is the mentioned label of the superpixel, which can then be used for further the clustering process. This result can be seen on Fig. 3(c).

5 Online Object Extraction

The point clouds generated in the previous step can be considered as raw spatial information. The robot has to abstract and further interpret this data in order to perceive spatial structures and recognize objects. However, the point clouds can be quite noisy mainly due to the following reasons: First, regions in the image might be classified wrongly, thus the corresponding points can be associated with wrong labels. Secondly, the accuracy of the projection of the pixels of the image depends on many parameters such as the number of successfully identified matching points. The error in the position of a stereo point can vary a lot from a few millimeters up to many centimeters.

Thus, in this work a 3D probabilistic grid is adopted in order to ensure spatio-temporal smoothing over the classes. An online clustering method that operates over the grid and reveals actual structures in the environment follows. Since most objects in the environment consist of sets of planar surfaces. the goal of the clustering method is to reveal the underlying surfaces that have generated our observed point clouds.

5.1 Basic Concepts

The robot stream consists of a sequence of labeled point clouds S_1, S_2, . . . , S_k arriving at time points t_1, t_2, . . . , t_k. Each point cloud S_i consists of a set of 3D labeled points, $S_i = \{p_1, p_2, \ldots, p_n\}$.

Due to the huge amount of the data generated by the robot, it is not feasible to retain all labeled 3D points. Therefore, a dynamic *3D grid* is used to partition the data space into cells of length ξ. The original points are assigned to the corresponding grid cells and we work hereafter on the grid instead of the original raw data.

Since the stereo point clouds are noisy, points with different labels might reside inside the same cell. To estimate the class label of a cell, the class distribution is preserved within each cell. In particular, for each class the probability of the class in the cell and the number of points that belong to this class (density) are stored. This information is updated as more data are accumulated over time and is expected to correctly approximate the true class label of the cell.

5.2 Online Clustering

Upon the arrival of a new labeled point cloud from the stream, our algorithm proceeds as follows: It first maps the points into the grid structure (grid update step). It then extracts the local clusters from this scan (local cluster extraction step). Finally, it merges the local clusters extracted from this scan to the so far generated global clusters (global cluster extraction step). To clarify the two concepts, local clusters are extracted from the data of a single scan and global clusters are generated from the accumulated scans of many time points. Below, these steps are described in more detail.

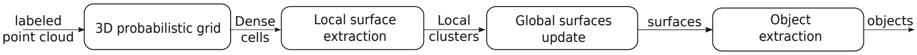


Fig. 4. Overview of object extraction

Step 1: Grid Update. Let G_{t-1} be the grid till time point $t-1$ and S_t be the new scan arriving at t. The goal of the first step is to register S_t to G_{t-1} and thus generate the updated grid G_t.

First, the new scan S_t is mapped onto the grid. For each point $p_k \in S_t$, the corresponding cell c is found. If c has not yet been created, the grid is expanded to include it. The label information of p_k is used to update the class distribution of the cell c. In particular, if $p(label_1^{t-1}), p(label_2^{t-1})...$ is the class distribution of cell c before the addition of p_k, this information is updated as follows:

$p(label_j^t|c, p_k) = p(label_j^{t-1}|c) * p(label_j|p_k.label)$

The probability $p(label_j|p_k.label)$ denotes the certainty about the classification process and it can be directly acquired by the confusion matrix of the previous section.

After all points of S_t have been mapped to the grid, the set of all new dense grid cells NC_t (cells whose density is above a given threshold) that will be used for the extraction of local clusters is generated. Existing grid cells that switched their majority class, i.e. are still dense but under a different majority class are also added to this set. However, cells that do not have a certain majority class, i.e. its probability is less than a threshold are considered noisy and not added to the set.

Step 2: Local Cluster Extraction. The goal of this step is to generate the local clusters LC_t from the updated grid G_t at t. Due to the limited range abilities of the robot, the scan S_t covers the environment only partially. Thus, the clusters generated in this step correspond only to parts of objects (e.g. a small part of a building), therefore they are called *partial* or *local* clusters.

A new cluster clu is created starting from a random cell c in NC_t (the set from Step 1). The normal vector and the class distribution of clu are initialized to the normal vector and label of c.

The algorithm expands the cluster based on the grid cells $c' \in NC_t$ that are directly connected to c. The cell c' is added to the cluster clu, if corresponding surfaces have similar orientations and their majority class agree. The normal vector of clu is then updated in order to consider the influence of c': a weighted average is performed over the normal vector of the cell and the vector of the cluster to update the later. The procedure continues until clu cannot be further expanded. The algorithm restarts from another cell c'' that has not been visited yet and continues until no more unvisited cells exist in NC_t. In this way, all local clusters LC_t from the current scan S_t are discovered.

Step 3: Global Cluster Update. The goal of the last step is to merge the local clusters LC_t produced at t with the old global clusters $clus_{t-1}$ produced till $t-1$ and thus produce the new global clusters $clus_t$. The clusters of this step are called global because they merge partial observations/clusters into full surfaces (e.g. a whole side of a building).

Intuitively, a partial cluster is considered as a *continuation* of a global cluster if they are close to each other in the grid, if their surface orientations are similar and if they correspond to the same class. The vicinity between the two clusters is defined in terms of the adjacency of their cells in the grid. It is equal to the number of directly connected cells between the two clusters.

The transitions *absorption*, *merge* and *birth* introduced in [48] for monitoring cluster evolution are adopted. A local cluster is considered to:

i be absorbed if there is only one similar global cluster with respect to orientation and label and their vicinity in the grid exceeds a *cluster vicinity threshold* $minUnits$.

ii be merged if two or more similar global clusters exist, as defined above. This is the case where parts of the same surface exist in different global clusters due to partial occlusions

iii start a *new* global cluster (birth) if nothing of the above occures.

In the case of absorption or merging, a final new global cluster is generated that contains all the cells from the local cluster and the global cluster(s), the weighted average of their normal vector and their label.

In order to generate complete objects such as cars and buildings which usually consist of different surfaces, neighboring surfaces with similar labels are merged together.

In Fig. 5, a 3D map generated from the raw labeled point clouds, the corresponding probabilistic grid and the extracted objects are portrayed. It is clear

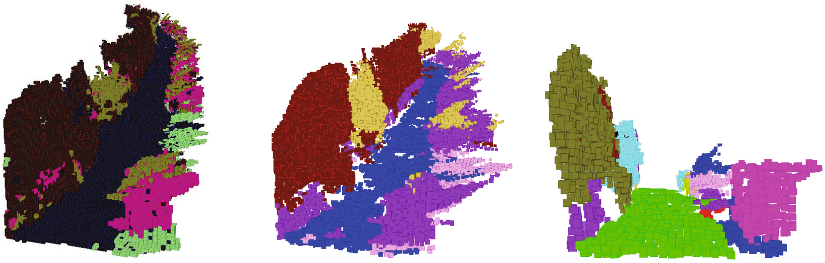

(a) Generated raw 3D map. (b) Filtered grid (c) Detected objects

Fig. 5. a) A map generated for the whole street using octomaps [49]. b) The 3D grid constructed by our algorithm of the same street. c) The extracted objects: cars on the right side and buildings on the left side. The colors in the first two images follow the ones in Fig. 2(d).

that our filtered grid is less noisy than the original raw map. For each cell, the majority class is displayed. Empty spots indicate that the corresponding cells have mixed class distributions.

6 Qualitative Spatial Relations

Based on the objects of the last section, we want to generate a graph-like structure of pairwise spatial relationships between objects and between each object and the robot. This allows the robot to reason about qualitative spatial aspects or to express the environment in human-understandable form which allows for a more natural communication.

Two different classes of spatial relations are considered based on the spatial deictics used in [34] but with a focus on close-by relations:

i Absolute Relation: It considers the relation of two objects without any reference direction. Here, an object can be *next, above, below* of another object.

ii Relative Relation: It considers the relation of an object to the robot which has a looking direction as a reference. An object can be *left, right, in front, behind, above, below* of the robot.

Absolute Spatial Relation. In order to extract the absolute spatial relations, the closest points between objects have to be found. A naive approach of considering the centroid for calculating the relations would fail since our objects can be of arbitrary shape. For example, the centroid of a street can be a few hundred meter away from a car, but the car might still be on the street. Therefore, we find the closest pair of points for two clusters by iterating over all point pairs.

From a pair of closest points, the distance vector is calculated as

$$\triangle \boldsymbol{r} = \boldsymbol{r}_2 - \boldsymbol{r}_1 \tag{2}$$

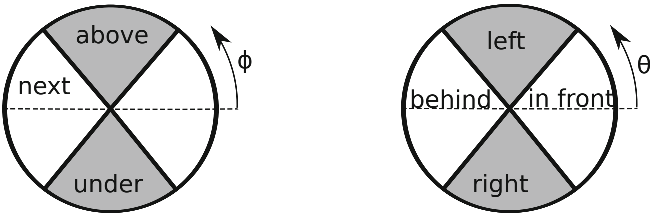
where $\boldsymbol{r}_1, \boldsymbol{r}_2$ are the closest points on cluster 1 and cluster 2 respectively. We transform the difference vector $\triangle \boldsymbol{r} = (\triangle x, \triangle y, \triangle z)$ into spherical coordinates as in [50]:

$$r = \sqrt{\triangle x^2 + \triangle y^2 + \triangle z^2} \qquad \in [0, \infty] \tag{3}$$

$$\phi = \arctan\left(\frac{\triangle y}{\triangle x}\right) \qquad \in [-\pi, \pi] \tag{4}$$

$$\theta = \arccos\left(\frac{\triangle z}{r}\right) \qquad \in [0, \pi] \tag{5}$$

For categorizing different relations, regions on the elevation angle θ are defined as shown in Fig. 6(a). Relations that do not match with the above regions are categorized as *none*.



(a) Absolute relation

(b) Refinement of 'next'

Fig. 6. Zones of the spherical coordinates for the different spatial relations. (a) shows the categorization for the absolute relations based on the elevation angle into *next*, *above* and *below*. (b) Additional categories in the relative relations based on the azimuth-angle.

In order to be more robust to noise, instead of using only the closest pair of points, we use the k nearest pairs to evaluate the relation between the two objects. The relationship is evaluated for each pair of points and a relation-membership distribution is generated. The probability of observing a relation R is given by the equation:

$$P(X = R) = \frac{N_R}{k} \tag{6}$$

where N_R is the number of times t relation R has been observed. In this way, multiple object relations can be extracted. For example, a sidewalk can be interpreted as next to a building, but also as partly under a building.

Relative Spatial Relation. In order to extract relative qualitative spatial relations, the robot position and orientation should be taken into account. Therefore, the same approach as with the generation of absolute relations is considered. The k-nearest points of the object to the robot position are found. We transform each of the individual points from the global coordinates (x_i, y_i, z_i) into a local robot coordinate system $(x_{l,i}, y_{l,i}, z_{l,i})$. We further transform these local Cartesian coordinates into spherical coordinates $(r_{l,i}, \phi_{l,i}, \theta_{l,i})$ to calculate the distance and the viewing angles with respect to the robot. The relations *above* and *below* can be inferred as in the case of absolute relations whereas the relation *next* can be further analyzed based on the orientation angle $\phi_{l,i}$ into the categories *behind* and *in front* as shown in Fig. 6(b).

Relation Graph. The pairwise relations can be modeled into a graph where the nodes are the objects of the environment and the edges are their relations. An example of such a representation is presented in Fig. 9 (and is explained in Section 7). This graph could be used for querying e.g. "find all objects to the left of the robot".

Although the complexity of graph construction is quadratic to the number of objects, there is no need to rebuild the graph from scratch each time a new point cloud is registered. The graph can be updated online by considering only the objects that have been affected due to the addition of the new point cloud.

7 Experimental Results

7.1 Classification on Image Level

CamVid Dataset. To test the correctness of the image classification stage we evaluated on the popular urban dataset CamVid[51], which offers high resolution images from the perspective of a car driving through a city, and various hundreds of labeled images. Since no dense 3D points are available on this dataset, we ignore the features derived from the point clouds and train only on appearance features. Also, the graph structure is taken directly from the superpixel adjacency structure, without pruning edges that link superpixels spatially far apart. Classification accuracy by using this modified version of the CRF model are presented in Table 1. These results are obtained with a reduced set of eight common classes, chosen to be mostly similar to those in our own dataset.

Table 1. Image segmentation results on the CamVid and IURO datasets. *Street sign for CamVid, Grass for IURO dataset.

Dataset	Sidewalk	Building	Tree	Pedestrian	Car	Sky	Street	Ss/Gr*	**Average**	**Global**
Camvid	26.6	86.4	69.3	20.0	85.2	92.8	95.8	14.1	62.3	80.1
IURO	92.9	94.8	82.6	0	83.0	96.3	64.8	13.6	64.8	90.9

IURO Dataset. In order to test our algorithm from the point of view of a mobile robot and with dense 3D data, a new dataset is being created with areas recorded in the neighbourhood of the city center TUM campus in Munich from a pedestrian robot perspective. This dataset includes sensor data from stereo cameras, the associated disparity maps [52], a laser range finder, and currently 140 labeled images with 8 semantic classes. The results on Table 1 were obtained by performing leave-one-out crossvalidation on the different streets. They show that most classes are correctly identified with good accuracy, although classes with few training examples, like pedestrians, are harder to identify reliably and generally have a lower accuracy. Discrepancies in results between the two datasets are probably related to the different appearances that classes take in each dataset, making them more distinguishable in one than in the other.

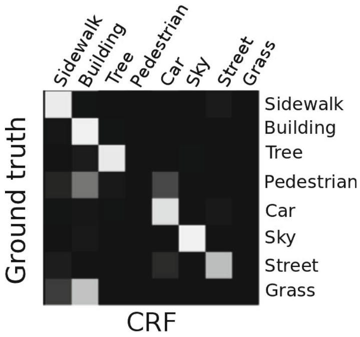


Fig. 7. Confusion matrix of the classification algorithm the IURO dataset

7.2 Object Extraction

In order to evaluate the object extraction step, different experiments were conducted. In the first one, we evaluate the filtering abilities of the 3D probabilistic grid. The purpose of the second experiment is to measure how well the resulting objects match with the objects in the environment. In the last experiment, we present the semantic map of a sidewalk in the IURO dataset. In these experiments, the ACE robot [53] was used. As a localization method, a laser based scan matching algorithm was applied to correct local odometry errors.

Noise Filtering through the 3D Grid. In this experiment, our goal is to evaluate the effect of the 3D grid on our data. We repeated the experiment on the environment of Fig. 2(a) three times with different cell sizes each time and counted the cells whose majority class has probability less than different thresholds.

In Table 2, it is shown that the smaller the cell size parameter is, the fewer noisy cells are observed. For example, for cells with 5cm size, only 4.99% have majority class with probability less than 50% and 18.32% with probability less than 90%. With the increase of the cells size, more points fall inside each cell thus the number of the noisy cells increases. Another finding of this experiment is that even when the cells size parameter is selected to be small, there still exist cells with mixed label distributions. This is due to the noise in the labeled point clouds and it indicates the 3D grid can be applied for filtering of these clouds.

Table 2. Percentance of noisy cells for different noise threshold

Cell size	50%	60%	70%	80%	90%
05 cms	04.99	13.25	13.69	16.81	18.32
10 cms	12.34	20.14	20.73	23.06	24.80
15 cms	18.77	26.61	27.24	29.20	30.61

Object Evaluation. In order to evaluate the quality of the extracted objects, a comparison with ground truth data is performed for two sidewalks in the city of Munich. In Table 3, the precision and recall of the object detection is presented. From the results, we can see that our algorithm has correctly detected all sidewalks and areas with trees. However, the detection mainly of cars is not so accurate. The main problems is that our algorithm fails to differentiate between two cars if they are parked too close to each other and groups them into one car object.

Table 3. Precision of correctly identified objects

	Sidewalk	Building	Cars	Trees
Precision	100%	91.4%	69.6%	100%
Recall	100%	100%	76.9%	100%

3D Semantic Map. In Fig. 8, a semantic map of a sidewalk in the city of Munich (shown in Fig. 2(a)) is presented. The length of the sidewalk is approximately 60m. 280 images were used for the construction of the map. By inspecting the generated map, it is easy to observe that the sidewalk (blue), the buildings (red) on the one side and the cars (purple) on the other side of the sidewalk are successfully captured. It is also worth noticing that the garage openings between buildings and a few areas with trees (yellow) are also portrayed well in the map. A few spurious cars can be found between buildings and the sidewalk. Although no real cars exist in these areas, the main reason for these detections is the parked bicycles and motorbikes on the side of these buildings.

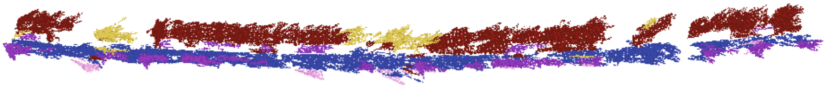


Fig. 8. Semantic map of Fig. 8 (the class colors are as in Fig. 5)

7.3 Spatial Relation Graph

The result when applying the algorithm to the objects shown in Fig. 8 can be seen in Fig. 9. For example, *Sidewalk #1* is under *Building #1* and the robot while *Car #2* is above *Street #1*. Comparing to ground truth data in the IURO dataset, the relation graph correctly classifies the relation in 86.5% of the cases and only deviate in certain cases from intuition. Most of these cases are, that due to a lack of seen points under a car, it is unlikely to get the correct relation of a car being on a street.

7.4 Execution Times

As an indication of the efficiency of the presented system, we present the average execution times of our three main sub-modules for the construction of the previous map. The image classification execution time was 0.48 seconds per image and the object extraction module required 0.37 seconds to process each point cloud. Since these two modules can run in parallel, the execution time of our system is 0.48 seconds on average. The output of the spatial reasoning module is not necessary to be available on time. Thus, the relations between the objects of the previous map were calculated offline at the end of the experiment and the execution time was 0.7 seconds.

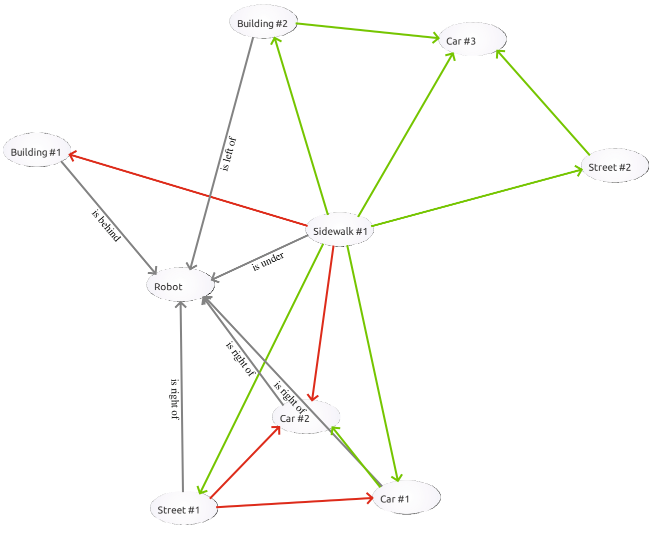


Fig. 9. Resulting spatial relations graph for Fig. 5(c), where only the strongest connection between two clusters is shown. Green encodes *next*-relationships, whereas red encodes *under*-relationships. The arrow indicates the direction of the relation, i.e. Sidewalk # 1 is under Car # 2.

8 Conclusions

In this paper, we proposed a system for semantic mapping in urban environments consisting of an image segmentation/classification component, an object extraction component and a spatial relation generation component. Starting with a stereo camera, our image labeling approach is able to classify important classes in urban environments with high precision and generate labeled point clouds. The clustering further improves the classification results through outlier removal and temporal smoothing. In addition to the semantic 3D-grid that is created, the spatial relation graph gives an abstract representation of the environment, which can help in reasoning or human-robot interaction. This multilayer representation of semantic information allows using different abstractions for different planning or sensing tasks.

For future work, several directions are considered. First of all, an experiment on a larger, more complex scene will be conducted. This would allow for a more thorough analysis of the pipeline presented here. Furthermore, ways to make the individual parts more integrated could be evaluated. For example, the current clustering and the relations between these clusters could be fed back to the image processing part for improving the classification process. It has to be verified whether the spatial relations extracted here are suitable for interaction with humans. Depending on this analysis, the types of relations could be further expanded in order to perform more sophisticated reasoning. Objects being left or right of each other under consideration of the orientation of objects would be one of these possible extensions.

Acknowledgements. This work is part of the IURO project supported by the 7th Framework Programme of the European Union, ICT Challenge 2 Cognitive Systems and Robotics, contract number 248317, by the DFG excellence initiative research cluster Cognition for Technical Systems CoTeSys, and by the Institute for Advanced Study (IAS), Technische Universität München.

References

1. Bazeille, S., Filliat, D.: Incremental topo-metric slam using vision and robot odometry. In: 2011 IEEE International Conference on Robotics and Automation (ICRA), pp. 4067–4073. IEEE (2011)
2. Doh, N.L., Lee, K., Chung, W.K., Cho, H.: Simultaneous localisation and mapping algorithm for topological maps with dynamics. Control Theory & Applications, IET 3(9), 1249–1260 (2009)
3. Elfes, A.: Using occupancy grids for mobile robot perception and navigation. Computer 22(6), 46–57 (1989)
4. Thrun, S., Bücken, A.: Integrating grid-based and topological maps for mobile robot navigation. In: Proceedings of the National Conference on Artificial Intelligence, pp. 944–951 (1996)
5. Zender, H., Martínez Mozos, O., Jensfelt, P., Kruijff, G.J.M., Burgard, W.: Conceptual spatial representations for indoor mobile robots. Robotics and Autonomous Systems 56(6), 493–502 (2008)
6. Galindo, C., Fernández-Madrigal, J.A., González, J., Saffiotti, A.: Robot task planning using semantic maps. Robotics and Autonomous Systems 56, 955–966 (2008)
7. Skubic, M., Perzanowski, D., Blisard, S., Schultz, A., Adams, W., Bugajska, M., Brock, D.: Spatial language for human-robot dialogs. IEEE Systems, Man, and Cybernetics, Part C: Applications and Reviews 34, 154–167 (2004)

8. Vasudevan, S., Siegwart, R.: Bayesian space conceptualization and place classification for semantic maps in mobile robotics. Robotics and Autonomous Systems 56(6), 522–537 (2008)
9. Pronobis, A., Mozos, O.M., Caputo, B., Jensfelt, P.: Multi-modal semantic place classification. The International Journal of Robotics Research 29, 298–320 (2010)
10. Nüchter, A., Hertzberg, J.: Towards semantic maps for mobile robots. Robotics and Autonomous Systems 56(11), 915–926 (2008)
11. Rusu, R.B., Marton, Z.C., Blodow, N., Holzbach, A., Beetz, M.: Model-based and learned semantic object labeling in 3d point cloud maps of kitchen environments. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009, pp. 3601–3608. IEEE (2009)
12. Wolf, D.F., Sukhatme, G.S.: Semantic mapping using mobile robots. IEEE Transactions on Robotics 24(2), 245–258 (2008)
13. Douillard, B., Fox, D., Ramos, F., Durrant-Whyte, H.: Classification and semantic mapping of urban environments. International Journal of Robotics Research 30(1), 5–32 (2011)
14. Kuipers, B.: The spatial semantic hierarchy. Artificial Intelligence 119(1-2), 191–233 (2000)
15. He, X., Zemel, R.S., Carreira-Perpinán, M.A.: Multiscale conditional random fields for image labeling. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, vol. 2, pp. 695–702. IEEE (2004)
16. Sturgess, P., Alahari, K., Ladicky, L., Torr, P.H.S.: Combining appearance and structure from motion features for road scene understanding. In: BMVC (2009)
17. Posner, I., Cummins, M., Newman, P.: A generative framework for fast urban labeling using spatial and temporal context. Auton. Robots 26(2-3), 153–170 (2009)
18. Ladicky, L., Sturgess, P., Russell, C., Sengupta, S., Bastanlar, Y., Clocksin, W., Torr, P.: Joint optimisation for object class segmentation and dense stereo reconstruction (2010)
19. Geiger, A., Lauer, M., Urtasun, R.: A generative model for 3d urban scene understanding from movable platforms. In: CVPR (2011)
20. Geiger, A., Wojek, C., Urtasun, R., Geiger, A., Lauer, M., Urtasun, R., Geiger, A., Ziegler, J., Stiller, C., Lenz, P., et al.: Joint 3d estimation of objects and scene layout. In: Neural Information Processing Systems
21. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Computing Surveys (CSUR) 31(3), 264–323 (1999)
22. Xu, R., Wunsch, D., et al.: Survey of clustering algorithms. IEEE Transactions on Neural Networks 16(3), 645–678 (2005)
23. Rusu, R., Marton, Z., Blodow, N., Holzbach, A., Beetz, M.: Model-based and learned semantic object labeling in 3D point cloud maps of kitchen environments. In: IROS (2009)
24. Wolf, D., Howard, A., Sukhatme, G.S.: Towards geometric 3d mapping of outdoor environments using mobile robots. In: IROS. IEEE (2005)
25. Valencia, R., Teniente, E.H., Trulls, E., Andrade-Cetto, J.: 3D mapping for urban service robots. In: IROS, IEEE (2009)
26. Marton, Z.C., Rusu, R.B., Beetz, M.: On fast surface reconstruction methods for large and noisy datasets. In: ICRA (2009)
27. Klasing, K., Wollherr, D., Buss, M.: Realtime segmentation of range data using continuous nearest neighbors. In: ICRA (2009)
28. Zlatanova, S., Rahman, A.A., Shi, W.: Topological models and frameworks for 3D spatial objects. In: Computers & Geosciences, pp. 419–428 (2004)

29. Jungert, E.: Qualitative spatial reasoning for determination of object relations using symbolic interval projections. In: Proceedings 1993 IEEE Symposium on Visual Languages, pp. 83–87 (August 1993)
30. Li, C., Lu, J., Yin, C., Ma, L.: Qualitative spatial representation and reasoning in 3d space. In: Intelligent Computation Technology and Automation, ICICTA 2009, vol. 1, pp. 653–657 (2009)
31. Skubic, M., Perzanowski, D., Schultz, A., Adams, W.: Using spatial language in a human-robot dialog. In: Proceedings of the 2002 International Conference on IEEE Robotics and Automation, ICRA 2002, vol. 4, pp. 4143–4148. IEEE (2002)
32. Moratz, R., Tenbrink, T., Bateman, J.A., Fischer, K.: Spatial Knowledge Representation for Human-Robot Interaction. In: Freksa, C., Brauer, W., Habel, C., Wender, K.F. (eds.) Spatial Cognition III. LNCS (LNAI), vol. 2685, pp. 263–286. Springer, Heidelberg (2003)
33. Bloch, I., Saffiotti, A.: On the representation of fuzzy spatial relations in robot maps. In: Intelligent Systems for Information Processing, pp. 47–57 (2002)
34. Bauer, A., Gonsior, B., Wollherr, D., Buss, M.: Heuristic rules for human-robot interaction based on principles from linguistics - asking for directions. In: AISB Convention - Symposium on New Frontiers in Human-Robot Interaction (2009)
35. Sj, K., Aydemir, A., Moerwald, T., Zhou, K., Jensfelt, P.: Mechanical support as a spatial abstraction for mobile robots. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4894–4900. IEEE (2010)
36. Sj, K., Pronobis, A., Jensfelt, P.: Functional topological relations for qualitative spatial representation. In: The 15th International Conference on Advanced Robotics (2011)
37. Tenorth, M.: Knowledge Processing for Autonomous Robots. PhD thesis, Technische Universitaet Muenchen (2011)
38. Lafferty, J., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
39. Kumar, S., Hebert, M.: Discriminative random fields: A discriminative framework for contextual interaction in classification. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, pp. 1150–1157. IEEE (2003)
40. Vishwanathan, S.V.N., Schraudolph, N.N., Schmidt, M.W., Murphy, K.P.: Accelerated training of conditional random fields with stochastic gradient methods. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 969–976. ACM (2006)
41. Wainwright, M.J., Jaakkola, T.S., Willsky, A.S.: Tree-based reparameterization framework for analysis of sum-product and related algorithms. IEEE Transactions on Information Theory 49(5), 1120–1146 (2003)
42. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888–905 (2000)
43. Levinshtein, A., Stere, A., Kutulakos, K.N., Fleet, D.J., Dickinson, S.J., Siddiqi, K.: Turbopixels: Fast superpixels using geometric flows. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(12), 2290–2297 (2009)
44. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels. Technical Report 149300 EPFL (June 2010)
45. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
46. Bhatia, A., Snyder, W.E., Bilbro, G.: Stacked integral image. In: 2010 IEEE International Conference on Robotics and Automation (ICRA), pp. 1530–1535. IEEE (2010)

47. Zhang, C., Wang, L., Yang, R.: Semantic Segmentation of Urban Scenes Using Dense Depth Maps. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 708–721. Springer, Heidelberg (2010)
48. Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., Schult, R.: Monic: modeling and monitoring cluster transitions. In: Proceedings of the 12th SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 706–711. ACM (2006)
49. Wurm, K.M., Hornung, A., Bennewitz, M., Stachniss, C., Burgard, W.: Octomap: A probabilistic, flexible, and compact 3d map representation for robotic systems. In: Proc. of the ICRA 2010 Workshop on Best Practice in 3D Perception and Modeling for Mobile Manipulation (2010)
50. del Pobil, A.P., Escrig, M.T., Jaen, J.A.: An attempt towards a general representation paradigm for spatial reasoning. In: International Conference on Systems, Man and Cybernetics, 'Systems Engineering in the Service of Humans', Conference Proceedings, vol. 1, pp. 215–220 (October 1993)
51. Brostow, G.J., Fauqueur, J., Cipolla, R.: Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters (2008)
52. Geiger, A., Ziegler, J., Stiller, C.: Stereoscan: Dense 3d reconstruction in real-time. In: IEEE Intelligent Vehicles Symposium, Baden-Baden, Germany (2011)
53. Bauer, A., Klasing, K., Xu, T., Sosnowski, S., Lidoris, G., Muhlbauer, Q., Zhang, T., Rohrmuller, F., Wollherr, D., Kuhnlenz, K., et al.: The autonomous city explorer project. In: IEEE International Conference on Robotics and Automation, ICRA 2009, pp. 1595–1596. IEEE (2009)

SURE: Surface Entropy for Distinctive 3D Features

Torsten Fiolka[1], Jörg Stückler[2], Dominik A. Klein[3],
Dirk Schulz[1], and Sven Behnke[2]

[1] Fraunhofer Institute for Communication, Information Processing
and Ergonomics FKIE, Wachtberg, Germany
{torsten.fiolka,dirk.schulz}@fkie.fraunhofer.de

[2] Autonomous Intelligent Systems Group, University of Bonn, Germany
stueckler@ais.uni-bonn.de, behnke@cs.uni-bonn.de

[3] Intelligent Vision Systems Group, University of Bonn, Germany
kleind@iai.uni-bonn.de

Abstract. In this paper, we present SURE features – a novel combination of interest point detector and descriptor for 3D point clouds and depth images. We propose an entropy-based interest operator that selects distinctive points on surfaces. It measures the variation in surface orientation from surface normals in the local vicinity of a point. We complement our approach by the design of a view-pose-invariant descriptor that captures local surface curvature properties, and we propose optional means to incorporate colorful texture information seamlessly. In experiments, we compare our approach to a state-of-the-art feature detector in depth images (NARF) and demonstrate similar repeatability of our detector. Our novel pair of detector and descriptor achieves superior results for matching interest points between images and also requires lower computation time.

Keywords: Depth image interest points, local shape-texture descriptor.

1 Introduction

Interest points paired with a descriptor of local image context provide a compact representation of image content. They can be used in various applications such as image registration [15,13,21], robot simultaneous localization and mapping (SLAM) [27], panorama stitching [3], photo tourism [7], as well as place [29] and object recognition [5,18,31].

Many applications require that a detector repeatedly finds interest points across images taken from various view poses and under differing lighting conditions. Since the scale of surface regions in the image depends on the distance of the sensor from the observed surface, the detector must also retrieve a repeatable scale if distance is not directly measured. This scale can then be used to normalize the size of the image region in which local image context is described.

Descriptors, on the other hand, are designed to distinguish well between different shapes and textures. They are often judged in terms of precision-recall

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 74–93, 2012.

relations [17]. However, one must admit that descriptor distinctiveness depends clearly on the variety of shapes and textures that appear at the selected interest points. Thus, a detector is preferable that finds interest points in various structures and highly expressive regions.

In this paper, we propose a new approach for extracting shape features at surface points through a measure of surface entropy (SURE). Our features combine a novel pair of interest point detector and local context description. Our approach can be applied to depth images as well as unorganized 3D point clouds. An entropy-based interest measure selects points on surfaces that exhibit strong local variation in surface orientation. We complement our approach by the design of a descriptor that captures local surface curvature properties. We also propose means to incorporate color and texture cues into the descriptor when RGB information is available for the points. We implement both detector and descriptor to process point clouds efficiently on a CPU. Our approach extracts features at a frame rate of about 5 Hz from RGB-D images at VGA resolution.

In experiments, we measure the repeatability of our interest points under view pose changes for several scenes and objects. We compare our approach with state-of-the-art detectors and demonstrate the advantages of our approach. We also assess the distinctiveness of our descriptor and point out differences to state-of-the-art methods.

2 Related Work

2.1 Interest Point Detection

Feature detection and description has been a very active area of research for decades. The computer vision community extensively studies detectors in intensity images. Nowadays, interest point detection algorithms are designed to be invariant against moderate scale and viewpoint changes [35]. There is not a single method that is always best in every application, but some noteworthy stick out from the bulk: The Harris-Affine [16] detector that recognizes corner structures based on the second moment matrix, the MSER [15] detector that identifies groups of pixels that are best separable from their surrounding, and the well known SIFT [14] or optimized SURF [1] detectors that are based on intensity blobs found by a difference of Gaussians filter. One recent example is the SFOP [6] detector for combination of corners, junctions, and blob-like features from a spiral model.

Most related to our method, also the entropy measure based on image intensities has been investigated for interest point detection [10,11]. It has been successfully applied to object recognition [5] due to the high informativeness of maximum entropy regions. Lee and Chen [12] picked up this idea of features based on histogram distributions and extended it to intensity gradients and color. They used the Bhattacharyya coefficient to identify local distributions that distinguish themselves most from the surrounding. Both approaches are not capable of real-time processing. In our approach, we adopted the entropy measure for 3D normal orientations in order to get stable-placed features determined by multiple surfaces.

However, those methods purely based on intensity image data suffer problems emerging from projective reduction to 2D space. Moreels and Perona [18] evaluated affine detectors for recognition of conspicuously shaped 3D objects and found out that none "performs well with viewpoint changes of more than 25-30°".

With the steadily increasing availability of depth measuring sensors, recently various methods have been developed to extract interest points from dense, full-view point clouds. The notion of scale has a different interpretation in 3D data. It now depicts the 3D extent of a structure which has been only intrinsic to the scale in 2D images. In depth images, the 2D projection of a structure at a specific 3D scale still varies with distance to the sensor. Few approaches have been proposed that detect interest points at multiple 3D scales and that automatically select a scale for which an interest point is maximally stable w.r.t. repeatability and localization.

Pauly et al. [22], for example, measure surface variation at a point by considering the eigenvalues of the local sample covariance. Novatnack et al. [20] extract multi-scale geometric interest points from dense point clouds with an associated triangular connectivity mesh. They build a scale-space of surface normals and derive edge and corner detection methods with automatic scale selection. For depth images [20], they approximate geodesic distances by computing shortest distances between points through the image lattice. Surface normals are computed by triangulating the range image. Our approach does not require connectivity information given by a mesh. Unnikrishnan et al. [37] derive an interest operator and a scale selection scheme for unorganized point clouds. They extract geodesic distances between points using disjoint minimum spanning trees in a time-consuming pre-processing stage. They present experimental results on full-view point clouds of objects without holes. In [32], this approach has been applied to depth images and an interest detector for corners with scale selection has been proposed. Steder et al. [29] extract interest points from depth images without scale selection, based on a measure of principal curvature which they extent to depth discontinuities. However, our approach is not restricted to depth images and can be readily employed for full-view point clouds.

2.2 Local Descriptors

The SIFT-descriptor [14] has been successfully used in computer vision applications. It describes the local gradient pattern in spatial histograms of gradient magnitudes and orientations. It is made rotation-invariant by aligning the histograms to the dominant gradient orientation at the interest point.

Several improvements to the SIFT descriptor have been proposed. SURF [1] sums Haar wavelet responses as a representation of the local gradient pattern. Recently, Calonder et al. [4] and Rublee et al. [24] demonstrated that binarized pixel comparisons at randomly distributed sample points yield a robust and highly efficient descriptor that outperforms SIFT or SURF.

Other approaches do not solely focus on gradient descriptions of texture. Shape Contexts [2], for instance, build a histogram of contour points in the local neighborhood of a point. Tuzel et al. [36] propose to use covariance of feature values in local image regions as a descriptor.

Johnson and Hebert [9] introduce spin-images to describe local shape context in 3D point clouds. In this approach, cylindrical coordinates of the local point distribution are described in a 2D image-like histogram. The surface normal at an interest point is chosen as the cylindrical axis, and the polar angle is neglected to project the points into 2D.

Shape Context [8,19] has been extended to 3D in order to describe the distribution of points in log-polar histograms. Tombari et al. [34] extract a local reference frame at a point and extract histograms of normals. In [33] they extend their approach to also capture the distribution of color. However, this method strongly depends on the stability of the reference frame.

Rusu et al. [26] quantify local surface curvature in rotation-invariant Fast Point Feature Histograms (FPFH). They demonstrate that the histograms can well distinguish between shapes such as corners, spheres, and edges.

Steder et al. [29] proposed the NARF descriptor for depth images. They determine a dominant orientation from depth gradients in a local image patch and extract radial depth gradient histograms. In conjunction with the NARF detector, Steder et al. [30] applied this descriptor for place recognition.

3 Entropy-Based Interest Points in 3D Point Clouds

3.1 Interest Points of Local Surface Entropy

Our detector is based on statistics about the distribution of local surface normals. We are interested in regions of maximal diversely oriented normals, since they show promise to be stably located at transitions of multiple surfaces or capture entire (sub-)structures that stick out of the surroundings. To identify such regions, we measure the entropy

$$H(X_{\mathcal{E}}) = -\sum_{x \in X_{\mathcal{E}}} p(x) \log p(x), \tag{1}$$

where $X_{\mathcal{E}}$ is a random variable characterizing the distribution of surface normal orientations occurring within a region of interest $\mathcal{E} \subseteq \mathbb{R}^3$. We extract interest points where this entropy measure achieves local maxima, i.e. where $X_{\mathcal{E}}$ is most balanced.

Entropy Computation from Point Clouds. Depth sensors usually measure surfaces by a set of discrete sample points $Q = \{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n\}, \boldsymbol{q}_k \in \mathbb{R}^3$. We approximate the surface normal at a sample point $n(\boldsymbol{q}_k)$ looking at the subset of neighboring points $\mathcal{N}_k = \{\boldsymbol{q}_l \in Q | \, \|\boldsymbol{q}_k - \boldsymbol{q}_l\|_1 < r\}$ within a given support range r. Then, $\hat{n}_r(\boldsymbol{q}_k)$ equals the eigenvector corresponding to the smallest eigenvalue of the sample covariance matrix $\text{cov}(\mathcal{N}_k)$.

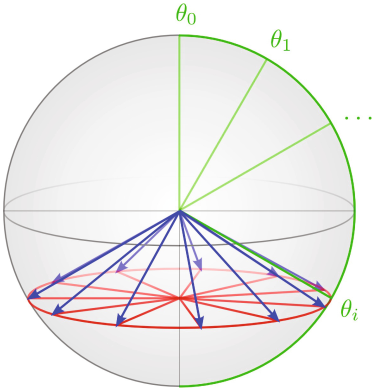


Fig. 1. Construction of an approx. uniform sphere partition. Green: equidistant inclination angles; red: sphere to cone section $C(\theta_i)$ and a_{θ_i} equidistant azimuth angles; blue: resulting orientation vectors on inclination level θ_i.

We discretize the surface normal distribution $X_{\mathcal{E}}$ by use of an orientation histogram in which we count the occurrences of surface normal orientations for a spherical surface partition. We follow the approach by Shah [28], subdividing the spherical surface into approximately equally sized patches. Those are specified by their centrical azimuth and inclination angles. To achieve an uniform decomposition of the sphere, we first choose t equidistant inclination angles $\theta_i = \frac{\pi i}{t}, i \in \{0, \ldots, t-1\}$. Then, for each of these inclination angles, we calculate a number of

$$a_{\theta_i} := \lfloor 2\ t \sin(\theta_i) + 1 \rfloor \propto C(\theta_i) \tag{2}$$

equidistant azimuth angles. This way, the sample density in azimuth is proportional to the circumference $C(\theta_i)$ of the section of the sphere with a cone of inclination θ_i. Transforming from spherical into Cartesian coordinates, we obtain a set of normalized vectors $\boldsymbol{v}_{i,j}$ pointing to the centers of histogram bins. Figure 1 depicts the construction of these vectors.

Each estimated surface normal at a point $\boldsymbol{q}_m \in Q \cap \mathcal{E}$ contributes to the histogram bin $x_{i,j}$ with a weight

$$w_{i,j} = \begin{cases} 0 & \text{, if } \hat{n}_r(\boldsymbol{q}_m) \cdot \boldsymbol{v}_{i,j} < \cos\alpha \\ \frac{\hat{n}_r(\boldsymbol{q}_m) \cdot \boldsymbol{v}_{i,j} - \cos\alpha}{1-\cos\alpha} & \text{, else} \end{cases}, \tag{3}$$

where α denotes the maximal angular range of influence. Finally, we normalize the histogram before calculating the surface normal entropy according to Equation 1.

3.2 Efficient Implementation Using an Octree

For efficient data access and well-ordered computation, we set up an octree structure containing the 3D point cloud inferred from the RGB-D image given by the sensor. In order to measure local surface entropy, our octree enables uniform sampling in 3D space. Furthermore, we exploit the multi-resolution architecture of the octree for fast volume queries of point statistics.

An octree organizes points from a 3D space into cubic subvolumes that are connected in a tree. The root node of the tree spans a volume chosen to fit the extent of the data. Edges between parent and child nodes represent a subset-relation. Each parent node branches into eight children constituting a partition of the parent's volume into eight equally sized octants. This branching is repeated iteratively until a predefined resolution, that equals a maximum depth of the tree, is reached.

The multi-scale structure of the octree allows for efficient bottom-up integration of data, facilitating the calculation of histograms, as well as search queries for local maxima in arbitrary volumes. In each node, we store histogram, integral and maximum statistics for different attributes of all points that are located within the volume of the node. These values can be computed efficiently by passing the attributes of points on a path from leave nodes to the root of the tree. This direction, every parent node accumulates and merges data received from its child nodes.

When querying for statistics inside an arbitrary 3D volume, we recursively descend the tree: if a node is fully inside the queried volume, its statistics are integrated into the response; if it is completely outside, this branch is discontinued; otherwise its child nodes are examined the same way. This is valid since each node already integrates the data of all leaves below in its own statistics. An easily understood example for data statistics is the average position of points within a certain volume $\mathcal{V}$. By integrating over the homogeneous coordinates of points $\boldsymbol{s} = (x, y, z, w)^T = \sum_{\boldsymbol{q}_i \in \mathcal{V}} (x_i, y_i, z_i, 1)^T$, one retains the mean via normalization $\bar{\boldsymbol{q}} = \frac{1}{w} \boldsymbol{s}$.

3.3 Interest Point Detection

The surface normal entropy function depends on two scale parameters: one is the radius r of vicinity $\mathcal{N}$ for the estimation of a surface normal orientation; the other is the extend of a region of interest $\mathcal{E}$, where the distribution of normals and thus the local surface entropy is gathered. These volumes are chosen to be cubic and appropriate to fit the intrinsic octree resolutions. The maximal depth ($\widehat{=}$resolution) of the octree is usually determined by the normal sampling interval at the finest scale that is specified to be a common multiple of the other dimensions. This way, range queries are processed most efficiently. Usually, sampling interval sizes of surface normals as well as normal orientation histograms are set to be at least half of the diameter of their respective local support volume.

All these parameters have to be chosen carefully. The histogram scale $\mathcal{E}$ corresponds directly to the size of the interest points, at which local structures

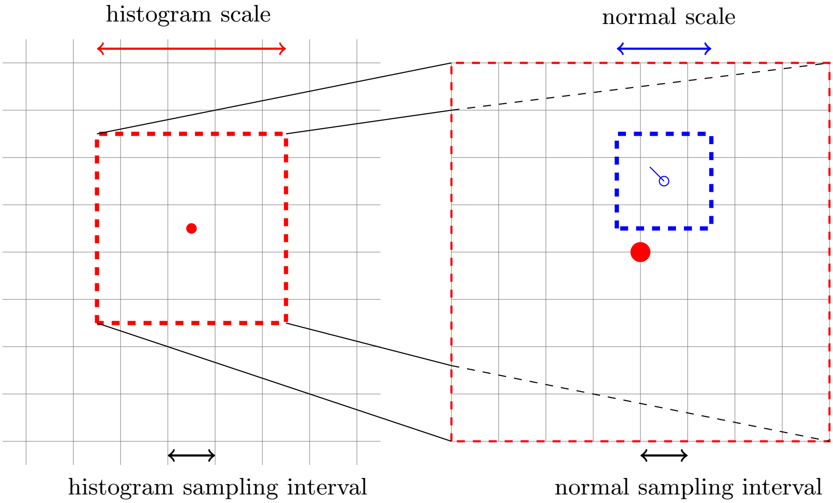


Fig. 2. Scheme of the different parameters for calculating normals and entropy

become salient. Its sampling interval is a trade-off between preciseness and speed. According to the Nyquist-Shannon sampling theorem, a minimal sampling frequency of twice the region size is needed to reconstruct the surface entropy function, i.e. not to miss the occurrence of a local maximum. We choose the normal scale r to a constant fraction of the histogram scale. Accordingly, the sampling interval for normals must also obey the sampling theorem. Reproducing the effect of a lowpass filter for removal of artifacts, we consider an entropy sample to be an interest point candidate, if it exceeds all its spatial neighbors within a dominance region. In addition, the candidate is only kept if it exceeds a global entropy threshold $H_{\min}$. The latter is checked, because noisy sensor data, image borders, and depth jumps occasionally induce interest point candidates on planar surfaces.

While surface entropy along an ideal ridge would be constant in theory, sensor noise and discretization artifacts will induce spurious measurements at these structures and thus cause local maxima of surface entropy. Such interest point candidates should be filtered out by inspection of the local prominence, since their position is loose in one dimension. Inspired by cornerness measures from image based interest point operators, we test for a considerable variance of surface entropy in all directions. First, we compute the local center of surface entropy mass within the region $\mathcal{E}_{\boldsymbol{q}}$ around a sample point $\boldsymbol{q}$

$$\mu_H(\mathcal{E}_{\boldsymbol{q}}) := \frac{1}{\sum_{\boldsymbol{q}_i \in \mathcal{E}_{\boldsymbol{q}}} H(X_{\mathcal{E}_{\boldsymbol{q}_i}})} \sum_{\boldsymbol{q}_i \in \mathcal{E}_{\boldsymbol{q}}} H(X_{\mathcal{E}_{\boldsymbol{q}_i}})\, \boldsymbol{q}_i. \tag{4}$$

Then, the sample covariance matrix of local surface entropy mass equals to

$$\mathrm{cov}_H(\mathcal{E}_{\boldsymbol{q}}) := \frac{1}{\sum_{\boldsymbol{q}_i \in \mathcal{E}_{\boldsymbol{q}}} H(X_{\mathcal{E}_{\boldsymbol{q}_i}})} \sum_{\boldsymbol{q}_i \in \mathcal{E}_{\boldsymbol{q}}} H(X_{\mathcal{E}_{\boldsymbol{q}_i}}) \left((\boldsymbol{q}_i - \mu_H(\mathcal{E}_{\boldsymbol{q}}))(\boldsymbol{q}_i - \mu_H(\mathcal{E}_{\boldsymbol{q}}))^T\right). \tag{5}$$

By decomposition of $\mathrm{cov}_H(\mathcal{E}_{\boldsymbol{q}})$ we derive the eigenvalues λ_1, λ_2, and λ_3 sorted by value in ascending order. Finally, our local prominence check is defined

$$P(\mathcal{E}_{\boldsymbol{q}}) = \frac{\lambda_1}{\lambda_3} \geq P_{\min}, \tag{6}$$

where we used $P_{\min} = 0.15$ in our experiments.

Improved Localization. After identification of interest point candidates, the true maximum location has to be recovered from the discretized surface entropy function. Starting from a candidate's location, we apply the mean-shift mode searching approach: We integrate surrounding surface entropy samples via a Gaussian window in order to estimate the gradient of the surface entropy density. Then, the position of the candidate is shifted along this gradient direction. This procedure is repeated up to three times.

Occlusion Handling in Depth Images. Surface entropy is supposed to be high where multiple different layers join together. In depth images, however, one cannot always measure all joining surfaces explicitly due to occlusions, resulting in a reduced entropy. This peculiarity of the measuring system should be compensated. Therefore, we detect jump edges in the depth image. Since we know that there must exist another hidden surface behind each foreground edge, we approximate it by adding artificial measurements in viewing direction up to a distance that meets the biggest used local entropy scale (cf. Fig. 3). While we use such points for the detection of interest points, we do not include this artificial information into the descriptor. We also discard detected interest points in the background at occlusions, since they are not stable w.r.t. view point changes.

4 Local Shape-Texture Descriptor

Since our surface entropy measure detects interest points at location where the surface exhibits strong local variation, we design a shape descriptor that captures local surface curvature. When RGB information is available, we also describe the local texture at an interest point. We aim at a rotation-invariant description of the interest points in order to match features despite of view pose changes. For each individual cue, we select a reasonable distance metric and combine them in a distance measure for the complete feature.

4.1 Shape

Surfel pair relations (see Fig. 4) have been demonstrated to be a powerful feature for describing local surface curvature [38,26]. Given two surfels (q_1, n_1)

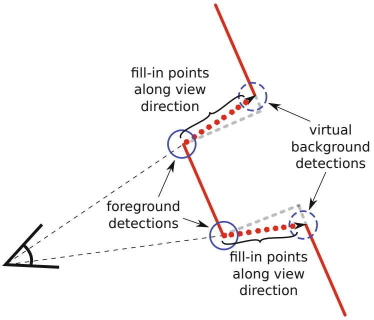

fill-in points along view direction
virtual background detections
foreground detections
fill-in points along view direction

Fig. 3. Occlusion handling. In depth images, structure may be occluded (dashed gray). At depth discontinuities, we therefore add artificial measurements (red dots) from foreground towards the background. Any "virtual background" detections are discarded, since they are not stable w.r.t. view point changes.

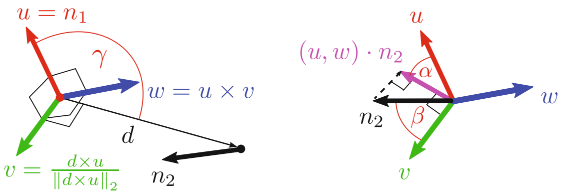


Fig. 4. Surfel pair relations describe rotation-invariant relative orientations and distances between two surfels

and (q_2, n_2) at points q_1 and q_2 with surface normals n_1 and n_2, we first define a reference frame (u, v, w) between the surfels through

$$\begin{aligned} u &:= n_1, \\ v &:= \frac{d \times u}{\|d \times u\|_2}, \text{ and} \\ w &:= u \times v, \end{aligned} \tag{7}$$

where $d := q_2 - q_1$. In this frame, we measure relative angles and distances between the surfels by

$$\begin{aligned} \alpha &:= \operatorname{arctan2}(w \cdot n_2, u \cdot n_2), \\ \beta &:= v \cdot n_2, \\ \gamma &:= u \cdot \frac{d}{\|d\|_2}, \text{ and} \\ \delta &:= \|d\|_2 . \end{aligned} \tag{8}$$

By construction, surfel pair relations are rotation-invariant and, hence, they can be used for a view-pose invariant description of local shapes.

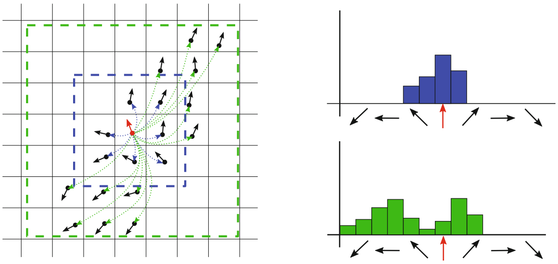


Fig. 5. Shape descriptor in a simplified 2D example. We build histograms of surfel pair relations from the surfels in a local neighborhood at an interest point. We relate surfels to the central surfel at the interest point. Histograms of inner and outer volumes capture distance-dependent curvature changes.

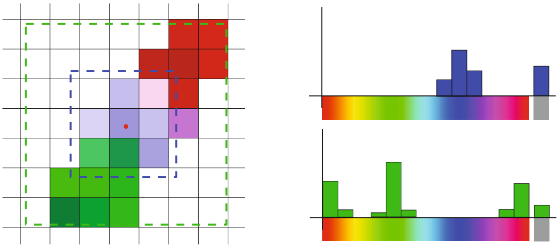


Fig. 6. Color descriptor. We extract hue and saturation histograms in an inner and outer local volume at an interest point.

In order to describe curvature in the local vicinity of an interest points, we build histograms of surfel pair relations from neighboring surfels (see Fig. 5). Each surfel is related to the surfel at the interest point being the reference surfel (p_1, n_1). We discretize the angular features into 11 bins each, while we use 2 distance bins to describe curvature in inner and outer volumes. We choose the support size of the descriptor in proportion to the histogram scale.

4.2 Color

A good color descriptor should allow interest points to be matched despite illumination changes. We choose the HSL color space and build histograms over hue and saturation in the local context of an interest point (see Fig. 6). Our histograms contain 24 bins for hue and one bin for unsaturated, i.e., “gray”, colors. Each entry to a hue bin is weighted with the saturation s of the color. The gray bin receives a value of $1 - s$. In this way, our histograms also capture information on colorless regions.

Similar to the shape descriptor, we divide the descriptor into 2 histograms over inner and outer volumes at the interest point. In this way, we measure the spatial distribution of color but still retain rotation-invariance.

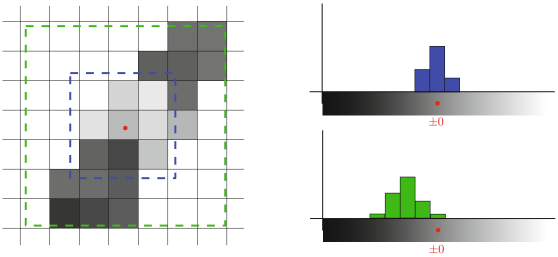


Fig. 7. Luminance descriptor. We describe luminance differences towards the interest point in histograms over local inner and outer volumes.

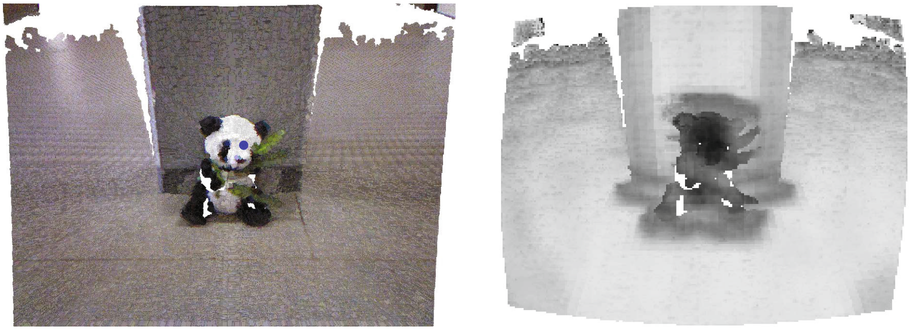


Fig. 8. Shape similarity w.r.t. the marked point (blue dot) measured using the Euclidean distance on our shape descriptors

4.3 Luminance

Since the color descriptor cannot distinguish between black and white, we propose to quantify the relative luminance change towards the color at the interest point (see Fig. 7). By this, our luminance descriptor is still invariant to ambient illumination. We use 10 bins for the relative luminance and, again, extract 2 histograms in inner and outer volumes.

4.4 Measuring Descriptor Distance

The character of the individual components of our descriptor suggests different kinds of distance metrics. We combine the distances $d_s(q_1, q_2)$, $d_c(q_1, q_2)$, and $d_l(q_1, q_2)$ between two points q_1 and q_2 using the arithmetic mean

$$d(q_1, q_2) := \frac{1}{3} \sum_{i \in \{s,c,l\}} d_i(q_1, q_2). \tag{9}$$

Shape Distance: For the shape descriptor, we use the Euclidean distance as proposed for FPFH features in [26]. We measure the arithmetic mean of the

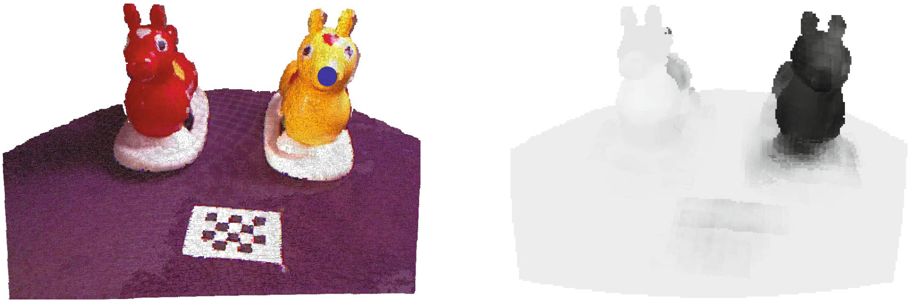


Fig. 9. Color similarity w.r.t. the marked point (blue dot) measured using the saturated Earth Mover's Distance ($\widehat{\text{EMD}}$) on our color descriptors

Euclidean distance of the angular histograms in the inner and outer volumes. Fig. 8 illustrates this distance measure in an example scene.

Color Distance: Since the HSL color space is only approximately illumination invariant, the domains of our color histograms may shift and may slightly be misaligned between frames. Hence, the Euclidean distance is not suitable. Instead, we apply an efficient variant of the Earth Mover's Distance (EMD, [25]) which has been shown to be a robust distance measure on color histograms.

The EMD between two histograms P and Q measures the minimum amount of mass in a histogram that needs to be “moved“ between the histograms to equalize them. Formally, the EMD is defined as

$$\text{EMD}(P, Q) = \frac{\min_{f_{ij}} \sum_{i,j} f_{ij} d_{ij}}{\sum_{ij} f_{ij}}, \tag{10}$$

where f_{ij} is the flow and d_{ij} is the ground distance between the bins P_i and Q_j. Pele and Werman [23] propose $\widehat{\text{EMD}}$, a modified EMD with saturated ground distance that is applicable to unnormalized histograms. They demonstrate that the $\widehat{\text{EMD}}$ can be implemented several magnitudes faster than the standard EMD but still retains its benefits. In our application, we saturate the ground distances at a distance of two bins. Fig. 9 illustrates our color distance in an example.

Luminance Distance: We also use the saturated $\widehat{\text{EMD}}$ to compare luminance histograms. See Fig. 10 for an example of our distance measure.

5 Experiments

5.1 Experiment Setup

We evaluate our approach on RGB-D images from a Microsoft Kinect and compare it with the NARF interest point detector and descriptor. We recorded

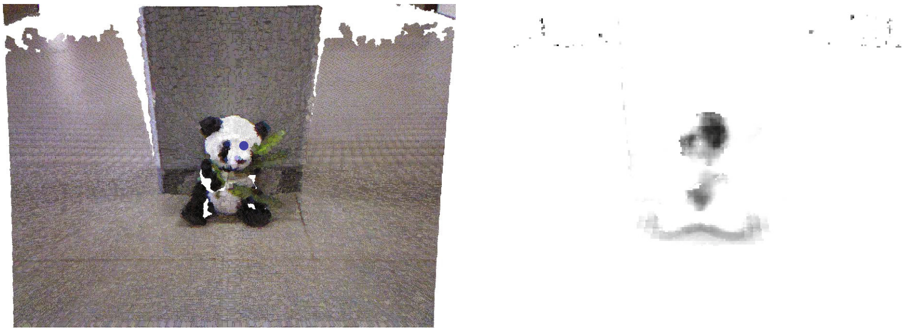


Fig. 10. Luminance similarity w.r.t. the marked point (blue dot) measured using the saturated Earth Mover's Distance ($\widehat{\text{EMD}}$) on our luminance descriptors

Table 1. Average run-time in seconds per frame for SURE and NARF detection and feature extraction

dataset	SURE 640x480	NARF 640x480	NARF 320x240	NARF 160x120
box	0.19	160.18	1.95	0.27
rocking horses	0.2	133.36	3.25	0.36
teddy	0.2	164.43	2.09	0.26
clutter	0.2	179.20	3.24	0.27

4 scenes, 3 containing objects of various size, shape, and color, and one cluttered scene with many objects in front of a wall. The objects are a box (ca. 50x25x25 cm), toy rocking horses (height ca. 1 m), and a teddy bear (height ca. 20 cm). Image sequences with 80 to 140 VGA images (640×480 resolution) have been obtained by moving the camera around the objects. We estimate the ground truth pose of the camera using checkerboard patterns laid out in the scenes. Furthermore, we evaluate the NARF descriptor on three resolutions of the datasets, at the original 640×480 and downsampled 320×240 and 160×120 resolutions. In each image of a sequence, we extract interest points on 3 histogram scales (SURE) or support sizes (NARF). We chose the scales 12, 24, and 48 cm.

5.2 Repeatability of the Detector

We assess the quality of our interest point detector by measuring its repeatability across view-point changes. We distinguish here between "simple repeatability" and "unique repeatability". Table 2 shows the average number of interest points found by the detectors. SURE finds a similar amount of features like NARF on 160×120 resolution.

We associate interest points between each image pair in the sequence using the ground truth transform. Each interest point can only be associated once to

Table 2. Average number of interest points for the SURE and NARF detectors

dataset	SURE 640x480	NARF 640x480	NARF 320x240	NARF 160x120
box	11.8	32.5	18.2	14.9
rocking horses	35.2	121.6	72.4	44.6
teddy	6.8	43.0	26.9	15.3
clutter	47.5	93.4	48.4	26.5

an interest point in the other image. We establish the mutually best correspondences according to the Euclidean distance between the interest points. Valid associations must have a distance below the histogram scale (SURE) or support size (NARF) of the interest point. “Unique repeatability” only accepts an association between interest points, if the match is unambiguous. This means, that the matched interest points must be the only possible match within the support size/histogram scale, otherwise the association is discarded.

From Fig. 11 we see that SURE and NARF yield similar repeatability on the box and the teddy datasets. The NARF detector shows here a better performance in the smaller resolutions, while performing worse in full resolution. On the rocking horses and the cluttered scene, SURE performs worse than NARF. However, about 50% resp. 25% of the interest points are still matchable across 90° view angle change. In Fig. 13 SURE performs better than NARF in terms of “unique repeatability”. The NARF detector allows several interest points being “close” to each other, i.e., in a distance smaller than their respective support sizes. A SURE interest point will be discarded if it lies within the histogram scale of another interest point and its entropy is lower compared to its neighbor. In that way, we ensure that a SURE interest point sticks out of its environment and can be uniquely matched by descriptor.

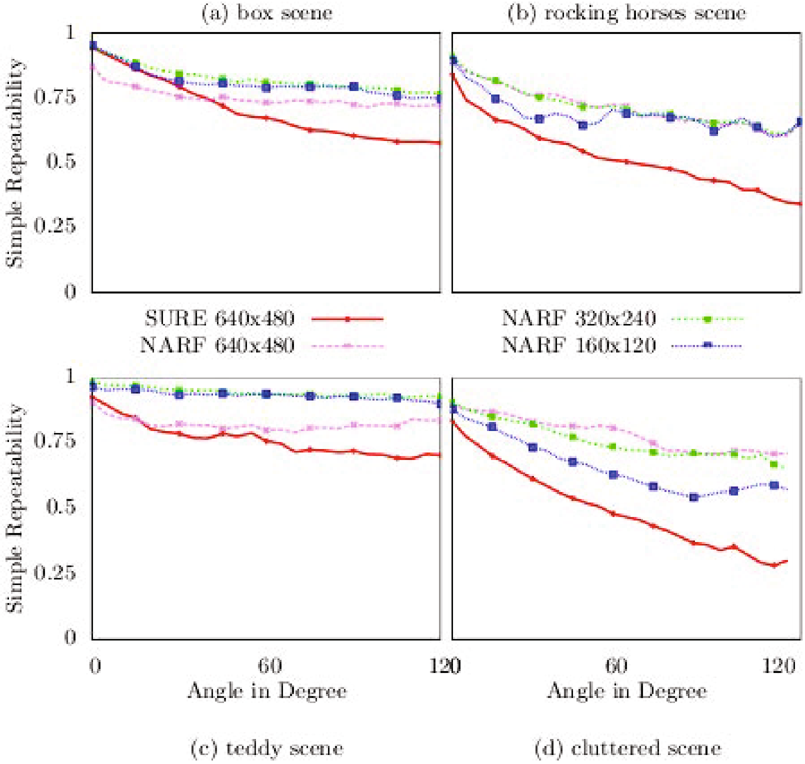
In Fig. 12 we also demonstrate the effect of our occlusion handling mechanism. If no artificial points are added along depth discontinuities, repeatability drops earlier with view angle change which is naturally expected.

5.3 Matching Score of the Descriptor

We also evaluate the capability of the detector-descriptor pair for establishing correct matches between images. We define the matching score as the fraction of interest points that can be correctly matched between images by the descriptor.

The results in Fig. 14 clearly demonstrate that SURE performs better than NARF in matching individual interest points. Its descriptor does not seem to be distinctive enough to reliably find correct matches. SURE, however, focuses on prominent local structure that is well distinguishable with our descriptor.

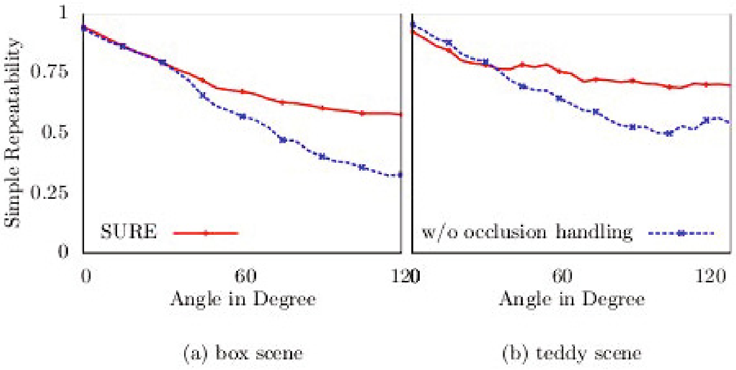
We also evaluate the matching score of the individual descriptor components of SURE in Fig. 15. In the teddy scene, very little color is present and the shape descriptor dominates color and lumincance. The clutter scene shows that the combination of these three descriptors performs considerably better than each of the descriptors alone.



(c) teddy scene

(d) cluttered scene

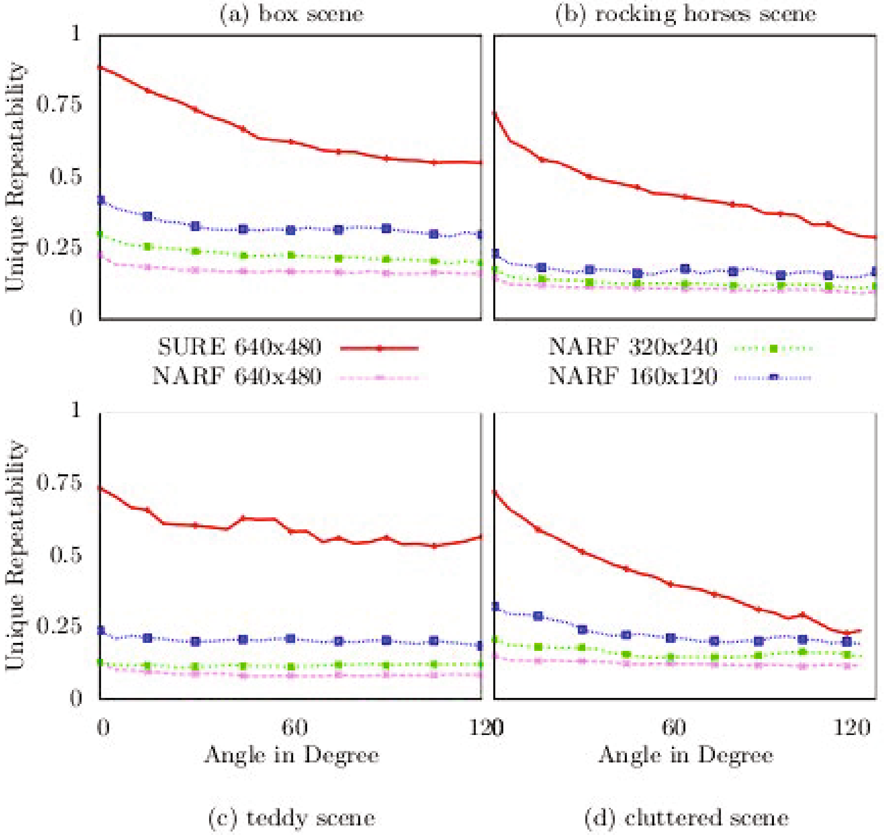
Fig. 11. Simple Repeatability in four different scenes comparing the SURE detector and the NARF detector. The NARF detector was applied in three different resolutions.



(a) box scene

(b) teddy scene

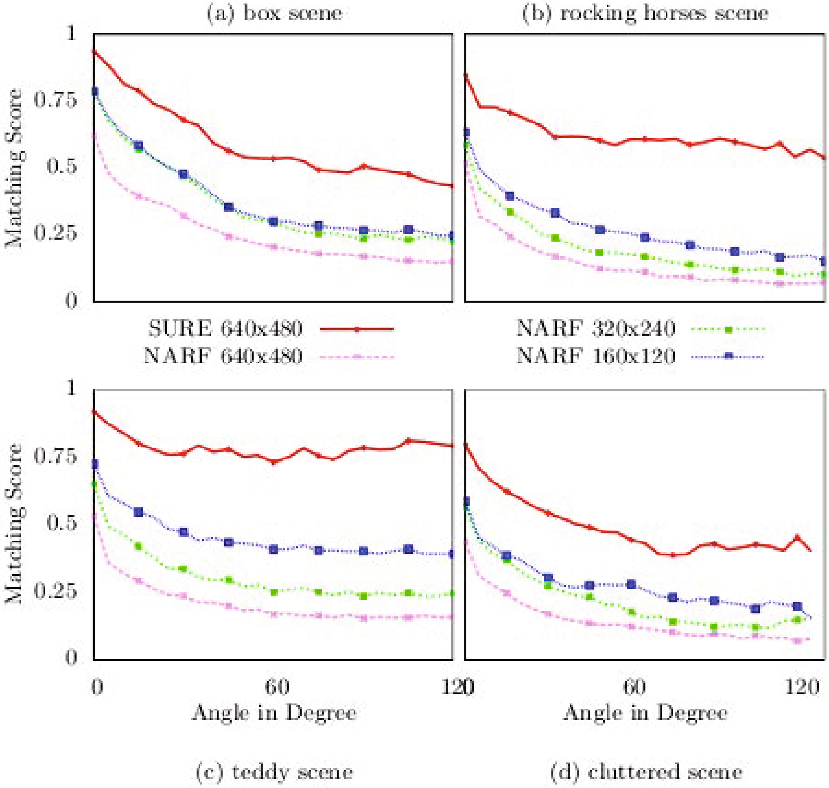
Fig. 12. Effect of occlusion handling on the repeatability of SURE



(c) teddy scene

(d) cluttered scene

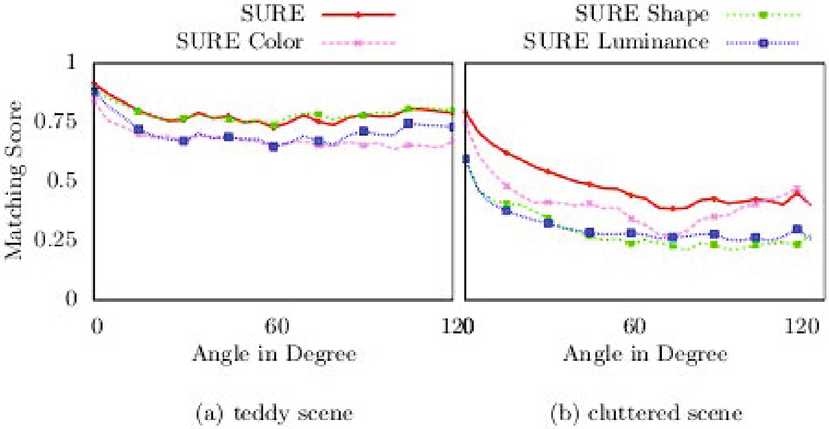
Fig. 13. Unique repeatability in four different scenes comparing the SURE detector and the NARF detector. Unique repeatability only accepts an association between interest points, if the match is unambiguous. This means, that the matched interest points must be the only possible match within the support size/histogram scale, otherwise the association is discarded.



(c) teddy scene

(d) cluttered scene

Fig. 14. Matching Score comparing SURE Feature Descriptor with the NARF Descriptor on four datasets



(a) teddy scene

(b) cluttered scene

Fig. 15. Matching Score comparing the different SURE Descriptors

5.4 Run-Time

Table 1 shows the run-time of NARF and SURE (detection and feature extraction). SURE outperforms NARF clearly on any of the processing resolutions of NARF, while SURE makes full use of the available data.

6 Conclusions

We proposed SURE, a novel pair of interest point detector and descriptor for 3D point clouds and depth images. Our interest point detector is based on a measure of surface entropy on normals that selects points with strong local surface variation. We designed a view-pose-invariant descriptor that quantifies this local surface curvature using surfel pair relations. When RGB information is available in the data, we also incorporate colorful texture information into the SURE descriptor. We describe color and luminance in the HSL space and measure distance using a fast variant of the Earth Mover's Distance to gain an illumination-invariant description at the interest point.

In experiments, we could demonstrate that the SURE detector achieves similar repeatability like the NARF descriptor. When matching features by descriptor, our SURE features outperform NARF regarding matching score. SURE also performs faster than NARF on 640×480 images.

In future work, we will further improve the run-time of SURE on depth and RGB-D images by exploiting the connectivity information in the image. We will also investigate automatic scale selection to further improve the repeatability and localization of the interest points.

References

1. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
2. Belongie, S., Malik, J., Puzicha, J.: Shape context: A new descriptor for shape matching and object recognition. In: NIPS, pp. 831–837 (2000)
3. Brown, M., Lowe, D.: Automatic panoramic image stitching using invariant features. Int'l Journal of Computer Vision 74, 59–73 (2007)
4. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: BRIEF: Binary Robust Independent Elementary Features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 778–792. Springer, Heidelberg (2010)
5. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning (2003)
6. Förstner, W., Dickscheid, T., Schindler, F.: Detecting interpretable and accurate scale-invariant keypoints. In: 12th IEEE International Conference on Computer Vision (ICCV 2009), Kyoto, Japan (2009)
7. Frahm, J.-M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.-H., Dunn, E., Clipp, B., Lazebnik, S., Pollefeys, M.: Building Rome on a Cloudless Day. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 368–381. Springer, Heidelberg (2010)

8. Frome, A., Huber, D., Kolluri, R., Bülow, T., Malik, J.: Recognizing Objects in Range Data Using Regional Point Descriptors. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3023, pp. 224–237. Springer, Heidelberg (2004)
9. Johnson, A., Hebert, M.: Using spin images for efficient object recognition in cluttered 3d scenes. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 21(5), 433–449 (1999)
10. Kadir, T., Brady, M.: Saliency, scale and image description. Int'l J. of Computer Vision 45(2), 83–105 (2001)
11. Kadir, T., Zisserman, A., Brady, M.: An Affine Invariant Salient Region Detector. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3021, pp. 228–241. Springer, Heidelberg (2004)
12. Lee, W.T., Chen, H.T.: Histogram-based interest point detectors. In: Int'l Conf. on Computer Vision and Pattern Recognition, CVPR (2009)
13. Lo, T.W.R., Siebert, J.P.: Local feature extraction and matching on range images: 2. 5D SIFT. Computer Vision and Image Understanding 113(12) (2009)
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (2), 91 (2004)
15. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proc. of the British Machine Vision Conference (2002)
16. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. Int'l Journal of Computer Vision (2004)
17. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions of Pattern Analysis and Machine Intelligence 27(10) (2005)
18. Moreels, P., Perona, P.: Evaluation of feature detectors and descriptors based on 3d objects. Int'l Journal of Computer Vision 73, 263–284 (2007)
19. Mori, G., Belongie, S., Malik, J.: Efficient shape matching using shape contexts. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 27(11), 1832–1837 (2005)
20. Novatnack, J., Nishino, K.: Scale-dependent 3D geometric features. In: Proc. of the IEEE Int. Conf. on Computer Vision, ICCV (2007)
21. Novatnack, J., Nishino, K.: Scale-Dependent/Invariant Local 3D Shape Descriptors for Fully Automatic Registration of Multiple Sets of Range Images. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 440–453. Springer, Heidelberg (2008)
22. Pauly, M., Keiser, R., Gross, M.: Multi-scale feature extraction on point-sampled surfaces. In: Eurographics (2003)
23. Pele, O., Werman, M.: Fast and robust earth mover's distances. In: Proc. of the Int. Conference on Computer Vision, ICCV (2009)
24. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: An efficient alternative to SIFT or SURF. In: International Conference on Computer Vision (2011)
25. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. Int. J. of Computer Vision 40, 99–121 (2000)
26. Rusu, R.B., Blodow, N., Beetz, M.: Fast point feature histograms (fpfh) for 3d registration. In: The IEEE Int. Conf. on Robotics and Automation, ICRA (2009)
27. Se, S., Lowe, D., Little, J.: Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. Int'l Journal of Robotics Research 21(8), 735–758 (2002)
28. Shah, T.R.: Automatic reconstruction of industrial installations using point clouds and images. Ph.D. thesis, TU Delft (2006)

29. Steder, B., Grisetti, G., Burgard, W.: Robust place recognition for 3D range data based on point features. In: Proc. of the IEEE Int. Conf. on Robotics and Automation, ICRA (2010)
30. Steder, B., Ruhnke, M., Grzonka, S., Burgard, W.: Place recognition in 3D scans using a combination of bag of words and point feature based relative pose estimation. In: Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, IROS (2011)
31. Steder, B., Rusu, R.B., Konolige, K., Burgard, W.: NARF: 3D range image features for object recognition. In: Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS) (2010)
32. Stückler, J., Behnke, S.: Interest point detection in depth images through scale-space surface analysis. In: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA (2011)
33. Tombari, F., Salti, S., Di Stefano, L.: A combined texture-shape descriptor for enhanced 3d feature matching. In: Proc. of the IEEE International Conference on Image Processing, ICIP (2011)
34. Tombari, F., Salti, S., Di Stefano, L.: Unique Signatures of Histograms for Local Surface Description. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 356–369. Springer, Heidelberg (2010)
35. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: A survey. Foundations and Trends in Computer Graphics and Vision 3(3), 177–280 (2007)
36. Tuzel, O., Porikli, F., Meer, P.: Region Covariance: A Fast Descriptor for Detection and Classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part II. LNCS, vol. 3952, pp. 589–600. Springer, Heidelberg (2006)
37. Unnikrishnan, R., Hebert, M.: Multi-scale interest regions from unorganized point clouds. In: Workshop on Search in 3D, IEEE Int. Conf. on Computer Vision and Pattern Recognition, CVPR (2008)
38. Wahl, E., Hillenbrand, G., Hirzinger, G.: Surflet-pair-relation histograms: a statistical 3D-shape representation for rapid classification. In: Proc. of the Int. Conf. on 3-D Digital Imaging and Modeling (2003)

Dense Map Inference with User-Defined Priors: From Priorlets to Scan Eigenvariations

Paloma de la Puente[1,⋆] and Andrea Censi[2]

[1] Intelligent Control Group, Universidad Politecnica de Madrid, C/ Jose Gutierrez Abascal, 2. 28006 Madrid, Spain
paloma.delapuente@upm.es

[2] Engineering and Applied Science, California Institute of Technology, Pasadena, MC 107-81, 91125, CA, USA
andrea@cds.caltech.edu

Abstract. When mapping is formulated in a Bayesian framework, the need of specifying a prior for the environment arises naturally. However, so far, the use of a particular structure prior has been coupled to working with a particular representation. We describe a system that supports inference with multiple priors while keeping the same dense representation. The priors are rigorously described by the user in a domain-specific language. Even though we work very close to the measurement space, we are able to represent structure constraints with the same expressivity as methods based on geometric primitives. This approach allows the intrinsic degrees of freedom of the environment's shape to be recovered. Experiments with simulated and real data sets will be presented.

1 Introduction

The Simultaneous Localization and Mapping (SLAM) problem is usually formulated in a Bayesian framework [16]. This paper concerns the use of prior distributions for the map: how to rigorously specify them and how to create an inference engine that works with multiple user-defined priors.

To see what role the prior plays in the problem, let us introduce some notation. Let $\boldsymbol{q}$ be the robot pose, let $\boldsymbol{m}$ be a variable representing the map, and let $\boldsymbol{z}$ be the measurements (including odometry and exteroceptive sensors), which follow the known sensor model $p(\boldsymbol{z}|\boldsymbol{q},\boldsymbol{m})$. SLAM can be formulated as the problem of estimating $p(\boldsymbol{q},\boldsymbol{m}|\boldsymbol{z})$, the joint distribution of pose and map conditioned to the measurements. We focus on the case of mapping with dense sensors and maps; if the map consists of landmarks, then most of the following remarks are not relevant. More specifically, we describe the formulation that uses the Rao-blackwellization technique [4], where one approximates the target distribution as $p(\boldsymbol{q},\boldsymbol{m}|\boldsymbol{z}) \simeq p(\boldsymbol{q}|\boldsymbol{z})p(\boldsymbol{m}|\boldsymbol{q},\boldsymbol{z})$, thereby factorizing SLAM in two subproblems: estimating the pose of the robot given the measurements ($p(\boldsymbol{q}|\boldsymbol{z})$), and mapping

[⋆] Paloma de la Puente's work was supported in part by Comunidad de Madrid and European Social Fund (ESF) programs and also by the Spanish Ministry of Science and Technology, DPI2010-21247-C02-01 - ARABOT (Autonomous Rational Robots).

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 94–113, 2012.

with known poses ($p(\boldsymbol{m}|\boldsymbol{q},\boldsymbol{z})$). Let us focus on the latter. Given the sensor model, we can compute the posterior using Bayes' theorem: $p(\boldsymbol{m}|\boldsymbol{q},\boldsymbol{z}) \propto p(\boldsymbol{z}|\boldsymbol{q},\boldsymbol{m})p(\boldsymbol{m})$.

Therefore, if we want to compute the posterior distribution of the map $\boldsymbol{m}$ given the observations, we need to know the prior $p(\boldsymbol{m})$. We remark that, had we formulated SLAM as a maximum-likelihood problem (find $\boldsymbol{m}$ that maximizes $p(\boldsymbol{z}|\boldsymbol{q},\boldsymbol{m})$), the knowledge of $p(\boldsymbol{m})$ would not be strictly necessary. That, however, would only work for finite-dimensional problems. In fact, if the underlying map is an arbitrary surface, the maximum-likelihood problem is ill posed, because the solution is any curve that perfectly interpolates the readings. To obtain a more reasonable solution, we always need some kind of regularization, which is the prior. Therefore, we conclude that, to make the SLAM problem with dense sensors and maps well posed, we have to specify a prior $p(\boldsymbol{m})$.

Other than to make the mathematical formulation correct, the knowledge of the prior helps in reducing the uncertainty of the estimate. For example, constraints such as collinearity are very powerful in reducing the map uncertainty. In general, any assumption about the environment that the user can provide helps in making the filter more efficient. Yet, to our knowledge, incorporating generic prior information in filters has never been done before, and that can be attributed to the representation used, which generally presents some limitations.

For instance, let us consider SLAM methods that represent maps using occupancy/evidence grids. Firstly, the grid resolution introduces some kind of spatial regularization, and makes it impossible to represent precise geometric primitives such as line segments. The other limitation is that each cell is assumed to be independent: this makes it impossible to effectively use the prior information because geometric constraints between different parts of the environment result in long-range correlation of cells occupancy.

A popular alternative to occupancy grids is using a map composed of geometric primitives (segments, circles, splines, etc.). In that case, the prior is implicit in the representation: representing a map by segments automatically gives non-segments map a zero prior. Using geometric primitives presents two major advantages: they provide explicit information about the geometrical nature of the environment, and the resulting maps are much more compact. With proper bookkeeping, the correlation between different parts of the environment can be precisely represented. However, they lack in flexibility. For example, in most realistic environments —except perhaps completely engineered factory floors— there will be parts of the environment that cannot be described by the prior. Moreover, often one wishes to impose "soft constraints": for example, rather than imposing that all walls are exact line segments, probably a better prior is that they are likely to be straight, or that they are of a bounded variation from straight; all these details should be figured out by the user. This flexibility cannot be accommodated by existing feature-based methods.

1.1 Contribution

We began this work by asking the question of whether it is possible to decouple the concept of prior from a particular representation. Instead of the prior being hidden in the representation, can it be made completely explicit and under the direct control of the user? Can we have an inference engine that works with multiple priors?

In Section 2, we start by defining a new representation. A range-finder provides an array of numbers measuring the distance to the obstacles. We augment that by associating to each measurement the corresponding surface normal. This gives us the flexibility of occupancy grids with the precision of geometric primitives. Just like large environments can be represented by a collection of patches, one can represent any environment by a collection of augmented scans; still, in this paper, we focus on processing the data from a single scan, with an ego-centric perspective.

In our system, the prior is entirely provided by the user, who describes the structure constraints and the model likelihood as a function of readings and normals, in a particular domain-specific language (Fig. 2 on page 99).

The inference engine, described in Section 3, takes two inputs: the noisy raw distance readings from a laser sensor, and the user-specified prior. The output is the posterior distribution for the local map, represented as a Gaussian distribution on the space of readings and normals. This is obtained by a two-step process. In the first step, we solve a nonlinear optimization problem to obtain the mode of the distribution. In the second step, described in Section 4, we use the knowledge of the structure constraints to shrink the measurements covariance, by projecting it onto the allowed submanifold. Fig 1 shows a geometric interpretation of the process.

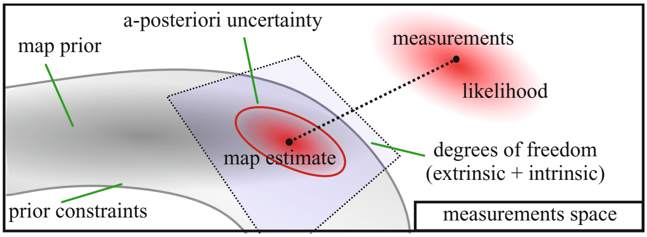


Fig. 1. Computing the posterior distribution of the map in the measurement space has a clear geometric interpretation. The prior $p(\boldsymbol{m})$ defines a (thin) surface in the measurement space. The initial measurements define a thick ellipse of uncertainty that gets projected and constrained to the prior surface.

We believe our method presents a novel approach for data segmentation and preprocessing in a flexible manner, being able to reduce the uncertainty of noisy measurements and providing information about the environment's geometrical nature. Section 5 includes some experiments showing how it works. In our opinion the proposed framework has a lot of potential for future research: it may also be very useful for scan-matching techniques with laser data and it is very promising for overall map optimization as well, helping us build better models of the physical world with a unique system in different situations.

1.2 Related Work

So far, prior information about the environment has been used explicitly only in feature-based SLAM methods. For example, Chong and Kleeman [3] employ collinearity constraints to enhance the state estimation with a Julier-Uhlman Kalman Filter. Rodríguez-Losada *et al.* [14] alleviate the inconsistency due to the linearization errors introduced by the Extended Kalman Filter by enforcing parallelism or orthogonality constraints. Parsley and Julier [12] propose a framework for the integration of prior information coming from different sources to improve the quality of feature based SLAM. Nguyen *et al.* [11] apply orthogonality constraints to build accurate simplified plane based 3D maps. Beevers and Huang [1] show that imposing a-priori known relative constraints also leads to consistency and efficiency improvements for particle filters. Other recent contributions are based on the graph-SLAM approach [7,17]. In all these works, a particular geometrical model for representation is used, and they only support equality constraints.

We know of no previous work using a dense representation and allowing for the use of different priors provided by the user. Modelling the scans as a Gaussian process [13] does allow to impose a prior distribution, corresponding to a smoothness constraint, but it cannot capture structured priors such as polygonal environments.

1.3 Notation

Let $\boldsymbol{q} = (\boldsymbol{t}, \theta) \in \mathrm{SE}(2)$ be the robot pose. Assume, without loss of generality, that the range-sensor frame coincides with the robot frame. The sensor model for the range-sensor measurements $\tilde{\boldsymbol{\rho}} = \{\tilde{\rho}_i\}_{i=1}^n$ is defined by $\tilde{\rho}_i = \rho_i + \epsilon_i$, where ρ_i is the true distance to the obstacle, and ϵ_i is additive Gaussian noise with covariance $\Sigma_{ij} = \mathrm{cov}\{\epsilon_i, \epsilon_j\}$, not necessarily diagonal. The true distance to the obstacle can be written as

$$\rho_i = r(\boldsymbol{m}, \langle \boldsymbol{t}, \theta + \phi_i \rangle), \tag{1}$$

where the angle ϕ_i is the direction of each reading in the scan, and the function $r : \mathcal{M} \times \mathrm{SE}(2) \to \mathbb{R}_+ \cup \{\infty\}$ is the "ray-tracing" function that returns the distance to the closest obstacle from a certain pose. The function r depends on the map $\boldsymbol{m} \in \mathcal{M}$. For now, we do not specify anything about $\boldsymbol{m}$, just that it represents the underlying map of the environment.

1.4 Problem Statement

Formally, we divide the problem of approximating $p(\boldsymbol{m}|\boldsymbol{z}, \boldsymbol{q})$, where $\boldsymbol{z} = \tilde{\boldsymbol{\rho}}$, in two sub-problems. First, we solve the maximum-a-posteriori problem to obtain the mode of the distribution. Since $p(\boldsymbol{m}|\boldsymbol{z}, \boldsymbol{q}) \propto p(\boldsymbol{z}|\boldsymbol{m}, \boldsymbol{q}) p(\boldsymbol{m})$, this can be posed as follows:

Problem 1. Find $\boldsymbol{m}$ that maximizes

$$\log p(\boldsymbol{z}|\boldsymbol{m},\boldsymbol{q}) + \log p(\boldsymbol{m}).$$

The first term is simply the measurements likelihood; the second term is the map prior. After we have found the mode of the distribution, we obtain a Gaussian approximation to $p(\boldsymbol{m}|\boldsymbol{z},\boldsymbol{q})$ by projecting the initial covariance onto the prior constraints (Fig. 1).

This process is conducted in a representation space very close to the measurement space, as described in the next section.

2 Defining Map Priors with Priorlets

The environment prior is specified by the user in a domain-specific language; a representative set of user-supplied prior definition files is shown in Fig. 2. Providing a flexible way to parametrize environment priors posed two challenges. The first mathematical challenge is choosing a unified representation that allows for the description of a multitude of priors. The second challenge is that this representation must also be user-friendly.

2.1 Representation: Distances $\boldsymbol{\rho}$, Normals $\boldsymbol{\alpha}$, Topology $\boldsymbol{\mathcal{T}}$

For what concerns the representation, our solution is parametrizing $p(\boldsymbol{m})$ by three finite-dimensional quantities.

True distance to the obstacle $\boldsymbol{\rho}$*:* The quantities $\{\rho_i\}_{i=1}^n$ were already defined as part of the sensor model in equation (1). They represent a zeroth-order approximation of the environment shape.

Surface normals $\boldsymbol{\alpha}$*:* The surface normals represent a first-order approximation of the environment shape, and will play an important role in defining the priors. The surface normal α_i can be written similarly to ρ_i as a function of the derivative of the ray-tracing function[1]. We define $\boldsymbol{x} \triangleq (\boldsymbol{\rho},\boldsymbol{\alpha})$ and we write compactly:

$$\boldsymbol{x} = (\boldsymbol{\rho},\boldsymbol{\alpha}) = \boldsymbol{r}(\boldsymbol{m},\boldsymbol{q}). \tag{2}$$

Environment topology $\boldsymbol{\mathcal{T}}$*:* We assume that the environment is partitioned into *surfaces,* and each *surface* is partitioned into one or more *regions.* For each two consecutive readings in the scan, there are three possible topology cases:

1. They belong to the same *surface* and the same *region.*
2. They belong to the same *surface*, but different *regions.*
3. They belong to different *surfaces.*

Having this fine distinction allows to precisely define the prior's constraints. To keep track of the topology information, we define a variable $\boldsymbol{\mathcal{T}} = \{\mathcal{T}_k\}_{k=1}^n$, where each $\mathcal{T}_k \in \{\mathsf{sameRegion}, \mathsf{differentRegion}, \mathsf{differentSurface}\}$ describes the relation between a pair of consecutive points.

[1] An explicit expression for the normal α_i as a function of the ray-tracing function r is $\alpha_i = \pi/2 + \arctan\left(\frac{\partial}{\partial\phi_i} r\left(\boldsymbol{m}, \langle \boldsymbol{t}, \theta+\phi_i\rangle\right)\right)$, but we are not going to need it in this paper.

```
name: Polygonal prior
order: 2
max_curvature: 0
p_1 = [cos( φ_1 ); sin( φ_1 )] * ρ_1; # define cartesian coords
p_2 = [cos( φ_2 ); sin( φ_2 )] * ρ_2; # as shortcuts
priorlet same_region:
   α_1 == α_2
  (p_2 - p_1)' * [cos( α_1 ); sin( α_1 )] == 0
```

(a) User-supplied definition for polygonal prior

```
name: Rectangular prior
specializes: Polygonal prior
priorlet different_region:
  ( α_2 == α_1 - pi/2 ) || ( α_2 == α_1 + pi/2 )
```

(b) User-supplied definition for rectangular prior

```
name: Rectangular prior (relaxed)
specializes: Polygonal prior
priorlet different_region:
   tolerance = 3; # 3deg tolerance
   cos( α_2 - α_1 ) <= cos(deg2rad(90+tolerance))
  -cos( α_2 - α_1 ) <= -cos(deg2rad(90-tolerance))
```

(c) User-supplied definition for relaxed rectangular prior

```
name: Rectangular prior (relaxed - alternative)
specializes: Polygonal prior
priorlet different_region:
  model_likelihood cos( α_2 - α_1 )^2
```

(d) User-supplied definition for alternative relaxed rectangular prior

```
name: Circular prior
order: 3
max_curvature: 10 # min radius = 0.1 m
# two oriented points define a circle. This is the radius.
r12 = sin(( α_2 - α_1 )/2) / norm(p_1 - p_2);
r23 = sin(( α_3 - α_2 )/2) / norm(p_3 - p_2);
r13 = sin(( α_3 - α_1 )/2) / norm(p_3 - p_1);
priorlet same_region:
  r12 == r23 # the three oriented points
  r23 == r13 # lie on the same circle
```

(e) User-supplied definition for circular prior

```
name: Circular prior (with prior on radius)
specializes: Circular prior
priorlet same_region: # it is likely that the radius is around 2.0
  model_likelihood (r13 - 2.0)^2
```

(f) Circular prior, with prior information for the radius

```
name: Splines prior
order: 2
max_curvature: 10
priorlet same_region:
  model_likelihood ( α_2 - α_1 )^2
```

(g) User-supplied definition for spline prior

Fig. 2. The environment prior is specified by the user with a domain-specific language. These are examples of actual source code interpreted by the inference engine (apart from some omissions in the interest of clarity). Using UNICODE, the special variables `alpha_i`, `rho_i`, `phi_i` can also be typed with Greek letters; this was inspired by Sun's Fortress language.

2.2 Expressing Priors as Functions of $\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{T}}$

We can express the prior as a function of the readings $\boldsymbol{\rho}$, normals $\boldsymbol{\alpha}$, and topology $\boldsymbol{\mathcal{T}}$ instead of as a function of the infinite-dimensional map $\boldsymbol{m}$. Assuming that it is possible, we rewrite Problem 1 as follows.

Problem 2. Find $\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{T}}$ that maximize

$$\log p(\tilde{\boldsymbol{\rho}}|\boldsymbol{\rho}) + \log p(\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{T}}).$$

Now we are dealing with a finite-dimensional optimization problem: the infinite-dimensional map "$\boldsymbol{m}$" has disappeared from the formalization. The limitation is that we can only define shape priors by their 0th (ρ) and 1st order (α) Taylor expansions. In the same spirit, we could use successive derivatives (curvature, and so on); nevertheless, we found that this parametrization has good expressivity. This does not mean that we are limited to piece-wise linear shapes; in fact, we can define shapes such as circles (Fig. 2e) and splines (Fig. 2g).

2.3 Expressing $p(\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{T}})$ with Local Constraints and Energies

Now we have fixed the representation, but we still have to solve the challenge of allowing the user to specify a prior in an intuitive way. It is clear that we can express almost any shape using a function $p(\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{T}})$. In theory, we could ask the user to provide a symbolic expression for $p(\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{T}})$. This, however, would be burdensome: assuming, for example, that there are 180 readings in a scan, the user would need to provide a symbolic expression with 540 variables. Moreover, that expression would have to be changed if the number of readings changed.

Our observation was that one can define interesting priors by describing *local* constraints between consecutive points. For example, if the environment prior is polygonal, we want to impose that nearby points have the same normals if they belong to the same region: $\alpha_1 = \alpha_2 = \cdots = \alpha_n$. This can be expressed compactly by saying that $\alpha_i = \alpha_{i+1}$ if points i and $i+1$ belong to the same region (compare Fig. 2a, line 7). In addition to these, we need constraints on ρ_i to ensure that the points are aligned (Fig. 2a, line 8).

In the case of a rectangular prior, we have the additional constraint that $(\alpha_i - \alpha_{i+1}) = k\frac{\pi}{2}$ if the two points do *not* belong to the same region (see Fig. 2b, line 4). Similarly, one can define different relaxations for a rectangular prior (Fig. 2c-2d). We will not describe in detail the interpretation of all the expressions in Fig. 2, but they all correspond to simple geometric constraints.

Certain priors cannot be specified by considering only two successive points. For example, it takes three consecutive points to describe a circular prior (see Fig. 2e), because it takes three points to define a circle. The *order* of a prior is the number of consecutive points needed for describing it.

2.4 Formal Definitions of Priorlets

We use the term *priorlet* for a set of local constraints plus energies imposed on n consecutive points in the environment.

Definition 1. *A* priorlet *of order n is a tuple $\langle F, G, H \rangle$ described by three sets of functions $F = \{f_k\}$, $G = \{g_k\}$, $H = \{h_k\}$. The arguments of all these functions are n couples of (distance, normal angle) and they all return a scalar. The functions $\{f_k\}$ represent equality constraints, the functions $\{g_k\}$ represent inequality constraints, and the functions $\{h_k\}$ represent "energies" (negative log-likelihoods).*

The semantics of a priorlet is the specification of a small part of a larger optimization problem:

$$\min_{\rho_{1:n}, \alpha_{1:n}} \quad \cdots + \textstyle\sum_k h_k((\rho_1, \alpha_1), \cdots, (\rho_n, \alpha_n)) + \cdots \ ,$$

$$\begin{aligned} \textit{subject to} \qquad & f_k((\rho_1, \alpha_1), \cdots, (\rho_n, \alpha_n)) = 0, \\ & g_k((\rho_1, \alpha_1), \cdots, (\rho_n, \alpha_n)) \le 0. \end{aligned}$$

The philosophy is very close to that of factor graphs *[6]; the formalization, however, does not match perfectly because usually factor graphs do not include constraints.*

Definition 2. *A* user-defined environment prior *is a collection of three priorlets: a "*`same_region`*" priorlet, a "*`different_region`*" priorlet, and a "*`different_surface`*" priorlet, describing the constraints/energies for neighbouring points for the three topology cases.*

Recall that the variable $\boldsymbol{\mathcal{T}}$ specifies the environment partition in regions and surfaces. Given a particular choice of $\boldsymbol{\mathcal{T}}$, we know which priorlet to apply to each couple (or triplet) of consecutive points. Therefore, we can define three functions $h_{\boldsymbol{\mathcal{T}}}(\boldsymbol{\rho}, \boldsymbol{\alpha})$, $f_{\boldsymbol{\mathcal{T}}}(\boldsymbol{\rho}, \boldsymbol{\alpha})$, $g_{\boldsymbol{\mathcal{T}}}(\boldsymbol{\rho}, \boldsymbol{\alpha})$. These represent, respectively, the cumulative effect of all the energies, and the stacked equalities and inequalities given by the application of the priorlets to each neighbourhood of points (we do not write them explicitly to avoid drowning in a sea of indices). We can rewrite Problem 2 as follows.

Problem 3. Find $\boldsymbol{\mathcal{T}}, \boldsymbol{\rho}, \boldsymbol{\alpha}$ as the solution of the problem:

$$\max_{\boldsymbol{\mathcal{T}}, \boldsymbol{\rho}, \boldsymbol{\alpha}} \quad \log p(\tilde{\boldsymbol{\rho}} | \boldsymbol{\rho}) + h_{\boldsymbol{\mathcal{T}}}(\boldsymbol{\rho}, \boldsymbol{\alpha}),$$

$$\begin{aligned} \text{subject to} \qquad & f_{\boldsymbol{\mathcal{T}}}(\boldsymbol{\rho}, \boldsymbol{\alpha}) = \mathbf{0}, \\ & g_{\boldsymbol{\mathcal{T}}}(\boldsymbol{\rho}, \boldsymbol{\alpha}) \le \mathbf{0}. \end{aligned}$$

2.5 A Domain-Specific Language for Priorlets

We have given a formal description of priorlets that might appear overly complicated. In practice, the process of specifying a prior is intuitive, using a domain-specific language whose syntax we believe easy to understand even without a formal definition.

The user must minimally specify a name, and the order of the prior. Then she specifies the three priorlets (a same_region priorlet, a different_region priorlet, and a different_surface priorlet), by specifying equalities (==) and inequalities (<=) over the predefined variables rho_i, alpha_i, phi_i, with $1 \leq i \leq$ order. Using UNICODE input, the variables can also be indicated with Greek letters. At any point in the file, other variables can be introduced using "=" (Fig. 2a, line 4). The syntax for the expressions is the one used by MATLAB/Octave. A "||" operand is supported for specifying a logical or (Fig. 2b, line 4). The model likelihood (the h function) is introduced by the keyword model_likelihood. The user can subclass existing priors using the specializes keyword; for example, the rectangular prior specializes the polygonal prior (Fig. 2b, line 2). Finally, we let the user specify an explicit max_curvature parameter that is used in the inference process.

3 Inference with Generic Priors

Our goal has been to build an inference engine that works for arbitrary user-specified priors. Of course, we are doomed to be less efficient than an optimization method designed for a particular prior; however, we believe there is value in showing a completely general approach. In this section, we briefly recall the standard constrained-optimization methods that we use, we show how additional constraints can be added to the problem, and finally we discuss the two-level optimization procedure.

3.1 Homotopy Methods

The idea of homotopy methods [15] is to solve the constrained optimization problem by solving a *sequence* of *unconstrained* optimization problems. The *penalty function* method is useful for dealing with equalities or inequalities. Suppose the minimization problem to solve is

$$\min_{\boldsymbol{x}} \; h(\boldsymbol{x}), \qquad \text{subject to } f(\boldsymbol{x}) = 0,$$

and assume that we do not know a feasible point. We then consider a sequence of unconstrained minimization problems, where we add to the objective function a penalty function representing the distance from the feasible set:

$$\min_{\boldsymbol{x}} \; h(\boldsymbol{x}) + \lambda f(\boldsymbol{x})^2.$$

Similarly, the penalty function for an inequality $g(\boldsymbol{x}) \leq 0$ would be $\lambda \max\{0, g(\boldsymbol{x})\}^2$. As $\lambda \to \infty$, the solution of the unconstrained problem tends to the solution of the constrained one. Therefore, we can solve the constrained problem by solving a sequence of unconstrained optimization problems, starting from $\lambda = 0$ and progressively raising it. Proper convergence can be proved under appropriate conditions [15].

The *log-barrier* method is useful for dealing with inequalities. Suppose we have to solve the problem

$$\min_{\boldsymbol{x}} \; h(\boldsymbol{x}), \quad \text{subject to } \boldsymbol{x} \leq \overline{\boldsymbol{x}},$$

and assume that we start from a feasible point $\boldsymbol{x}_0 \leq \overline{\boldsymbol{x}}$. Then we solve the sequence of unconstrained optimization problems

$$\min_{\boldsymbol{x}} \; h(\boldsymbol{x}) - \frac{1}{\mu} \sum_i \log\left((\overline{x}_i - x_i)\right).$$

The log term represents a "barrier" that goes to infinity near the bounds. As $\mu \to \infty$, the solution of the unconstrained problem tends to the solution of the constrained one.

3.2 Additional Details

Outliers: We expect that the prior supplied by the user describes most of the environment, but there will always be points that are clearly outside the prior, caused, for example, by clutter in the environment. Therefore, we define another optimization variable, the set INMODEL of points that do respect the prior. Suppose that the likelihood of a point being described by the prior model is $\beta \in (0, 1]$, and, for simplicity, that each point is independent. Then the log-likelihood component $\log(p(\text{INMODEL}))$ can be represented in the cost function by a term $\gamma|\text{INMODEL}|$, with $\gamma = \log(\beta/(1-\beta))$ and $|\text{INMODEL}|$ indicating the number of points.

Upper and lower bounds on $\boldsymbol{\rho}$, $\boldsymbol{\alpha}$: It is possible to derive upper and lower bounds for the variables $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$. For $\boldsymbol{\rho}$, bounds are obtained by using the initial covariance information. During the optimization, each ρ_i is allowed to vary at most $4\sigma_i$ from the initial estimate $\tilde{\rho}_i$. Because of that, outliers and clutter produce constraints that are impossible to satisfy, and eventually those points are removed from the INMODEL set. As for the normals, it is possible to derive bounds for α_i based on the allowed variation of $\rho_{i-1}, \rho_i, \rho_{i+1}$ and the knowledge of the maximum curvature in the environment.

3.3 Optimization Overview

We rewrite again the form of the optimization problem, with the new variable INMODEL and the bounds on the state.

Problem 4. Find $\boldsymbol{\mathcal{T}}, \text{INMODEL}, \boldsymbol{x}$ as the solution of:

$$\max_{\boldsymbol{\mathcal{T}}, \boldsymbol{x} \in \text{INMODEL}} \log p(\boldsymbol{x}|\boldsymbol{\rho}) + h_{\boldsymbol{\mathcal{T}}}(\boldsymbol{x}) + \gamma|\text{INMODEL}|,$$

$$\text{subject to} \quad f_{\mathcal{T}}(\boldsymbol{x}) = 0, \tag{3}$$
$$g_{\mathcal{T}}(\boldsymbol{x}) \leq 0, \tag{4}$$
$$\underline{\boldsymbol{x}} \leq \boldsymbol{x} \leq \overline{\boldsymbol{x}}. \tag{5}$$

We have to optimize over discrete and continuous variables. The discrete variables are the set INMODEL and the topology $\boldsymbol{\mathcal{T}}$. The continuous variable is $\boldsymbol{x} = (\boldsymbol{\rho}, \boldsymbol{\alpha})$. We solve the problem using two nested levels: the outer level (Algorithm 1) optimizes over INMODEL and $\boldsymbol{\mathcal{T}}$, while the inner level (Algorithm 2) optimizes over $\boldsymbol{x}$, given a particular choice of INMODEL and $\boldsymbol{\mathcal{T}}$. We describe the inner level first.

3.4 Inner Loop: Optimizing x Given INMODEL, $\boldsymbol{\mathcal{T}}$

Algorithm 2 solves Problem 4 assuming that INMODEL, $\boldsymbol{\mathcal{T}}$ have been fixed. We apply a double homotopy transformation to find $\boldsymbol{x}$. We use a penalty function for constraints (3)-(4) and a log-barrier method for constraint (5). Using the log-barrier for the bounds ensures that those are always satisfied during each iteration. Instead, the constraints on the prior are satisfied only in the limit: we start from the measurements and eventually arrive to the surface defined by the prior (Fig. 1).

At each iteration, we take a Newton step with backtracking. All the necessary gradients and Hessians are computed in closed form using symbolic derivations from the user-specified constraints. Moreover, we "convexify" the Hessian if it is not positive-definite by setting negative eigenvalues to a small positive value (Algorithm 2, line 12); this turns the Newton method into gradient descent in the non-convex parts of the state space.

Note that Algorithm 2 might fail to return a feasible point; this will be interpreted by the outer level as a sign that the topology $\boldsymbol{\mathcal{T}}$ is wrong and must be relaxed.

3.5 Outer Loop: Optimizing INMODEL, $\boldsymbol{\mathcal{T}}$

Algorithm 1 optimizes over the set INMODEL and the topology $\boldsymbol{\mathcal{T}}$. Solving this problem exactly has combinatorial complexity, as we would have to try each possible grouping of points into surfaces and regions. To obtain an approximate solution, we use a heuristic approach based on relaxation.

We initialize INMODEL to contain all the points, and $\boldsymbol{\mathcal{T}}$ to result in the strictest set of constraints ($\mathcal{T}_i = \texttt{sameRegion}$). Iteratively, we call the inner level to find a corresponding $\boldsymbol{x}$. If Algorithm 2 finds a feasible $\boldsymbol{x}$, we are done. Otherwise, we try to relax the problem. If the problem is infeasible, some of the prior constraints ($g_{\mathcal{T}}(\boldsymbol{x}) \leq \mathbf{0}, f_{\mathcal{T}}(\boldsymbol{x}) = 0$) are not respected and the corresponding penalty functions are non-zero. We check which couple of nearby points gave the most contribution to the penalty function, and we relax the topology (line 14). If the corresponding $\mathcal{T}_k$ was sameRegion, we set it to differentRegion; if it was differentRegion, we set it to differentSurface. We observed that this simple algorithm was effective in finding region and surface boundaries.

In addition, we check whether some regions are too small, and we remove the corresponding points from INMODEL (line 16). This is useful for dealing with outliers.

Algorithm 1. Discrete Optimization of INMODEL, $\boldsymbol{\mathcal{T}}$

```
function [$\boldsymbol{x}$, $\boldsymbol{\mathcal{T}}$] = map_optimization($\tilde{\boldsymbol{\rho}}$,$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\rho}}}$,prior):            1
  % initialize by using all points, and the strictest topology                        2
  INMODEL = all; $\mathcal{T}_k$ = sameRegion;                                           3
  while True:                                                                         4
    [$\underline{\boldsymbol{x}}$,$\overline{\boldsymbol{x}}$] = geometric_bounds($\boldsymbol{\mathcal{T}}$,$\tilde{\boldsymbol{\rho}}$,$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\rho}}}$)                                          5
    % Estimate surface normals                                                        6
    [$\boldsymbol{\alpha}_0$, covalpha] = estimate_initial_alpha($\tilde{\boldsymbol{\rho}}$,$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\rho}}}$,$\boldsymbol{\mathcal{T}}$)                         7
    % Restrict optimization to the INMODEL set                                        8
    $\boldsymbol{x}_0$ = $\{(\tilde{\boldsymbol{\rho}}, \alpha_0)\}$ for $i \in$ INMODEL                                            9
    [feasible, $\boldsymbol{x}$, link_penalties] =                                          10
       inner_optimization(prior,$\boldsymbol{x}_0$,cov0,$\boldsymbol{\mathcal{T}}$,[$\underline{\boldsymbol{x}}$,$\overline{\boldsymbol{x}}$])                      11
    if feasible: break                                                                12
    % If not feasible, break the topology based on the penalties                      13
    $\boldsymbol{\mathcal{T}}$ = break_greedily($\boldsymbol{\mathcal{T}}$,link_penalties)                                    14
    % Remove points in small regions from the INMODEL set                             15
    [INMODEL,$\boldsymbol{\mathcal{T}}$] = remove_lonely_points(INMODEL,$\boldsymbol{\mathcal{T}}$)                          16
 return [$\boldsymbol{x}$, $\boldsymbol{\mathcal{T}}$]                                                                      17
```

Algorithm 2. Continuous optimization of $\boldsymbol{x}$

```
[feasible,$\boldsymbol{x}$,link_penalties] = inner_optimization(prior,$\boldsymbol{x}_0$,$\boldsymbol{\Sigma}_{\boldsymbol{x}_0}$,$\boldsymbol{\mathcal{T}}$,[$\underline{\boldsymbol{x}}$,$\overline{\boldsymbol{x}}$])
  % Obtain functions from prior and topology
  $f_{\boldsymbol{\mathcal{T}}}(\boldsymbol{x}), g_{\boldsymbol{\mathcal{T}}}(\boldsymbol{x}), h_{\boldsymbol{\mathcal{T}}}(\boldsymbol{x})$ = prior_to_constraints(prior, $\boldsymbol{\mathcal{T}}$)
  for $\lambda$=$\lambda_0$; $\lambda \leq \lambda_{\max}$; $\lambda = \lambda_{\text{mult}}\lambda$:
    for $\mu = \mu_0$; $\mu \leq \mu_{\max}$; $\mu = \mu_{\text{mult}}\mu$:
      % return if the point is feasible
      if $f_{\boldsymbol{\mathcal{T}}}(\boldsymbol{x}) < \epsilon$: return [true, $\boldsymbol{x}$]
      % compute gradient and Hessian of objective + penalties
      $J(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{x}_0, \boldsymbol{\Sigma}_{\boldsymbol{x}_0}) + h_{\boldsymbol{\mathcal{T}}}(\boldsymbol{x})$
         + log_barrier([$\underline{\boldsymbol{x}}$, $\overline{\boldsymbol{x}}$],$\mu$,$\boldsymbol{x}$) + $\lambda$ penalty($f_{\boldsymbol{\mathcal{T}}}(\boldsymbol{x})$, $g_{\boldsymbol{\mathcal{T}}}(\boldsymbol{x})$)
      % convexify the Hessian (do gradient descent if nonconvex)
      H = convexify($\nabla^2_{\boldsymbol{x}} J(\boldsymbol{x})$)
      newton_direction = $-$inv(H) $*$ $\nabla_{\boldsymbol{x}} J(\boldsymbol{x})$
      $\boldsymbol{x}$ = back_tracking($\boldsymbol{x}$, newton_direction)
  % the problem is infeasible: compute the penalty for each link
  link_penalties = compute_link_penalties($\boldsymbol{\mathcal{T}}$,$\boldsymbol{x}$)
  return [false, null, link_penalties]
```

4 Recovering the Degrees of Freedom

We have shown how to define generic priors (Section 2) and how to perform inference with them (Section 3). We have decoupled the environment prior from the environment representation: while the priors are most general, the representation is always the same. This approach certainly has its advantages in terms of generality and flexibility. However, we lose something with respect to a feature-based approach. The advantages of representing the map with geometric primitives is that the representation implicitly encodes the *constraints* and *degrees of freedom* of the environment. If we fit a circle to the environment, we implicitly state that 1) the points are constrained to lie on a circle *(*constraints*)*; and 2) the circle can change in radius and position (degrees of freedom). In this section, we show how we can perform a similar analysis even using augmented scans for the representation.

4.1 The Geometric Structure of the Map Space

In equation (1), we let the sensor model depend on the underlying true map "$\boldsymbol{m}$", interpreted as an abstract infinite-dimensional quantity belonging to a certain set $\mathcal{M}$. In order to derive well-grounded results, we have to formalize some intuitive ideas about $\mathcal{M}$ (some of these are commented in more detail elsewhere [2]).

It is intuitive that, for each map $\boldsymbol{m} \in \mathcal{M}$, there will be other elements in $\mathcal{M}$ that have the same *shape* but are rotated/translated to different *poses*. Thus, we can assume that all reasonable sets $\mathcal{M}$ are isomorphic to the product $\mathcal{S} \times \mathrm{SE}(2)$, where $\mathcal{S}$ is called the *shape space*. Given this factorization, we can write an element $\boldsymbol{m} \in \mathcal{M}$ as a couple $\langle \boldsymbol{S}, \boldsymbol{p} \rangle \in \mathcal{S} \times \mathrm{SE}(2)$. This factorization is the basis of many works in the shape-space analysis [8,9]. Based on that, we introduce a technical condition on the user-defined prior.

Definition 3. *A prior $p(\boldsymbol{m})$ is* pose-independent *if it only depends on the map shape $\boldsymbol{S}$ but not on the map pose $\boldsymbol{p}$:*

$$p(\boldsymbol{m}) = p(\langle \boldsymbol{S}, \boldsymbol{p} \rangle) = p(\boldsymbol{S}).$$

Intuitively, this means that, if the prior allows a certain shape, then it must allow the same shape, rotated, with equal probability; or, equivalently, that observing the environment does not give any information on the robot pose in an external frame. We also state a simple lemma on the ray-tracing function.

Lemma 1. *The observations do not change if robot and map are jointly roto translated: $\boldsymbol{r}(\langle \boldsymbol{S}, \boldsymbol{p} \rangle, \boldsymbol{q}) = \boldsymbol{r}(\langle \boldsymbol{S}, \boldsymbol{\delta} \oplus \boldsymbol{p} \rangle, \boldsymbol{\delta} \oplus \boldsymbol{q})$.*

4.2 Analyzing the Degrees of Freedom

Assume we have found a feasible solution $\boldsymbol{x}$. By analyzing the constraints given by the prior, we can recover the degrees of freedom in the solution. More formally, we consider infinitesimal variations $\delta\boldsymbol{x}$ and we examine which ones are allowed by the prior. Recall that $\boldsymbol{x}$ contains both scan readings and surface normals, therefore $\delta\boldsymbol{x}$ belongs to $\mathbb{R}^{2n}$, where n is the number of readings. We first give the mathematical results and then we comment on the derivation.

Proposition 1. *Suppose the prior is pose-independent (Definition 3). Then the space of the allowed variations $\delta\boldsymbol{x}$ to the solution can be factorized as ("$\sqcup$" indicates disjoint union):*

$$\mathbb{R}^{2n} = \mathit{Constr} \sqcup \mathit{Free} = \mathit{Constr} \sqcup (\mathit{Intr} \sqcup \mathit{Extr}),$$

where the subspaces are defined (and computed) as follows:

$$\begin{aligned} \textit{Free} &\triangleq \ker \nabla_{\boldsymbol{x}} f_{\mathcal{T}}(\boldsymbol{x}), && (6)\\ \textit{Constr} &\triangleq \mathbb{R}^{2n} - \textit{Free}, && (7)\\ \textit{Extr} &\triangleq \textit{span}\{\nabla_{\boldsymbol{q}} \boldsymbol{r}\}, && (8)\\ \textit{Intr} &\triangleq \textit{Free} - \textit{Extr}. && (9) \end{aligned}$$

The subspaces Free are the directions corresponding to the map variations allowed by the prior. The subspace Free is further divided in intrinsic *(Intr) degrees of freedom, due to the uncertainty in the map's shape; and* extrinsic *(Extr) degrees of freedom, due to the uncertainty in the map's pose.*

To explain the first division in the subspaces Constr and Free, we just need to consider the equality constraints in the prior, which are represented by the equation $f_{\mathcal{T}}(\boldsymbol{x}) = 0$. This equation defines a hyper-surface inside $\mathbb{R}^{2n}$ where $\boldsymbol{x}$ is constrained to lie. The tangent plane to this surface is given by directions orthogonal to the gradient $\nabla_{\boldsymbol{x}} f_{\mathcal{T}}$, and corresponds to the (infinitesimal) directions that are allowed by the prior. The subspace Constr is simply the complement of Free.

The further division of Free in intrinsic (Intr) and extrinsic (Extr) degrees of freedom is a more delicate topic. We have seen that the map $\boldsymbol{m}$ can be represented as a couple shape-pose $\langle \boldsymbol{S}, \boldsymbol{p} \rangle$. The subspace Extr identifies the variation in the readings due to the uncertainty in the pose $\boldsymbol{p}$; or, more precisely, due to the uncertain pose between map and sensor. We can state the following result.

Proposition 2. *If the prior is pose-independent, the subspace Extr* $\triangleq$ *span*$\{\nabla_{\boldsymbol{q}} \boldsymbol{r}\}$ *is contained in Free.*

Proof. Using (2), we write $\boldsymbol{x} = \boldsymbol{r}(\boldsymbol{m}, \boldsymbol{q}) = \boldsymbol{r}(\langle \boldsymbol{S}, \boldsymbol{p} \rangle, \boldsymbol{q})$. If the prior does not depend on $\boldsymbol{p}$, then $f_{\mathcal{T}}(\boldsymbol{x}) = 0$ implies

$$f_{\mathcal{T}}(\boldsymbol{r}(\langle \boldsymbol{S}, \boldsymbol{\delta} \oplus \boldsymbol{p} \rangle, \boldsymbol{q})) = 0, \qquad \text{for all } \boldsymbol{\delta} \in \mathrm{SE}(2). \tag{10}$$

Given Lemma 1, we obtain that $f_{\mathcal{T}}(\boldsymbol{r}(\langle \boldsymbol{S}, \boldsymbol{p} \rangle, \ominus \boldsymbol{\delta} \oplus \boldsymbol{q})) = 0$, for all $\boldsymbol{\delta} \in \mathrm{SE}(2)$. This means that

$$f_{\mathcal{T}}(\boldsymbol{r}(\langle \boldsymbol{S}, \boldsymbol{p} \rangle, \boldsymbol{q})) = 0, \qquad \text{for all } \boldsymbol{q} \in \mathrm{SE}(2).$$

Intuitively, this says that the updated readings still respect the prior no matter where the robot is placed in the environment. If a function ($f_{\mathcal{T}}$) is constant with respect to an argument ($\boldsymbol{q}$), the derivative with respect to that argument is 0. In our case, using the chain rule, we obtain:

$$\nabla_{\boldsymbol{q}} f_{\mathcal{T}} = \nabla_{\boldsymbol{x}} f_{\mathcal{T}} \cdot \nabla_{\boldsymbol{q}} \boldsymbol{r} = \boldsymbol{0}.$$

Therefore, $\nabla_{\boldsymbol{q}} \boldsymbol{r}$ is always orthogonal to $\nabla_{\boldsymbol{x}} f_{\mathcal{T}}$; that is, $\mathrm{span}\{\nabla_{\boldsymbol{q}} \boldsymbol{r}\} \subset \ker \nabla_{\boldsymbol{x}} f_{\mathcal{T}} =$ Free.

Extr = span $\{\nabla_{\boldsymbol{q}}\boldsymbol{r}\}$ are the possible variations in the measurements due to the sensor movement. By contrast, the directions in Intr are due to the variation in the map shape $\boldsymbol{S}$, and correspond to the intuitive notion of the degrees of freedom in the structure. To compute them, we use equations (6), (8), (9). Note that we used the concept of map factorization in $\boldsymbol{m} = \langle \boldsymbol{S}, \boldsymbol{p} \rangle$ only as a theoretical tool in deriving the results. In practice, we do not need to know anything about such abstract representation; the only quantities we have to compute are $\nabla_{\boldsymbol{q}}\boldsymbol{r}$ and $\nabla_{\boldsymbol{x}} f_{\boldsymbol{\mathcal{T}}}$, which lie in the very concrete measurement space. The procedure is completely automatic and allows to recover the degrees of freedom for any prior.

4.3 Covariance Shrinking

Other than for visualization purposes, we can use the degrees of freedom knowledge for computing the posterior uncertainty of the estimate. Assume that the covariance of the initial estimate $\boldsymbol{x}_0$ was $\boldsymbol{\Sigma}_{\boldsymbol{x}_0}$. If the prior has only constraints and not energies (i.e., there is no term $h_{\boldsymbol{\mathcal{T}}}(\boldsymbol{x})$), we can obtain the posterior covariance $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ simply by projecting $\boldsymbol{\Sigma}_{\boldsymbol{x}_0}$ onto the subspace Free. Let the matrix P_{Free} be a projector onto Free. Then the posterior covariance estimate is $\boldsymbol{\Sigma}_{\boldsymbol{x}} = P_{\text{Free}} \boldsymbol{\Sigma}_{\boldsymbol{x}_0} P_{\text{Free}}^T$. If there is a term $h_{\boldsymbol{\mathcal{T}}}(\boldsymbol{x})$, we have to account for the further reduction of uncertainty. Treating it as an additional observation, we obtain that $\boldsymbol{\Sigma}_{\boldsymbol{x}} = P_{\text{Free}}(\boldsymbol{\Sigma}_{\boldsymbol{x}_0}^{-1} + \nabla^2 h_{\boldsymbol{\mathcal{T}}}^{-1})^{-1} P_{\text{Free}}^T$. Similarly, one can recover the contribution to $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ due to extrinsic or intrinsic uncertainty by projecting onto Extr or Intr. This process has a solid geometric intuition; see also, for example, Chapter 3 of Paul Newman's thesis [10]. The reader should note that this linearized analysis has the usual limitations [5].

For priors with many constraints, the rank of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is very low. Thus, it is better to represent it by its non-null eigensystem, which can be interpreted as the allowed scan eigenvariations $\{\langle \boldsymbol{v}^m, \sigma_m \rangle\}_{m=1}^{\dim(\text{Free})}$, each representing a direction $\boldsymbol{v}^m = \langle \boldsymbol{\rho}^m, \boldsymbol{\alpha}^m \rangle$ and corresponding uncertainty σ_m in that direction.

5 Experiments

We have conducted experiments with both synthetic and real data.

Fig. 3 shows an example test case, with a simulated scan from a square environment, using the rectangular prior. The original noisy simulated data is depicted in Fig 3(a), whereas Fig 3(b) presents the corrected measurements and the proper division of the scan into different regions. Fig 4 shows the orientation angles for all the readings corrected by applying our method (x-like crosses), with the ground-truth represented by dots and the initial estimates represented by circles. The vertical crosses indicate the bounds. The solution gets so close to the ground-truth that they can hardly be distinguished in the plot, with an average error of 0.26° for several tests and the walls being well aligned. We obtain similar results with a variety of other simulated environments. We also

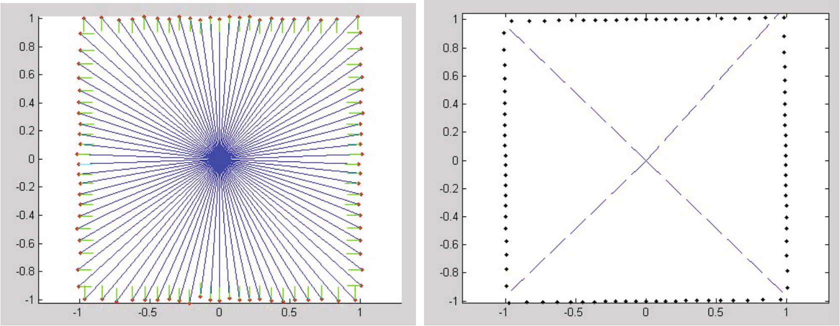


Fig. 3. (a): Noisy simulated scan. (b): Corrected measurements and topology $\mathcal{T}$ (dashed edges for different regions).

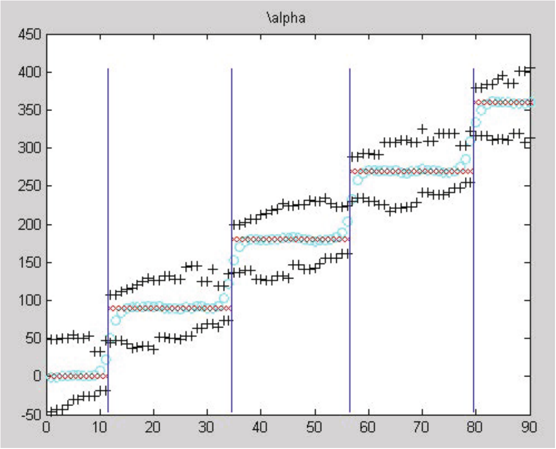


Fig. 4. Corrected orientation angles (x-like symbols), ground truth (small dots), initial values (circles) and extracted topology $\mathcal{T}$ (vertical edges for different regions)

conducted tests with a random number of outliers, Fig. 5 shows some examples. Other experiments had similar results, they are not presented here for lack of space. The quantitative analysis of these results requires some further work. It is not always easy to assess whether a measurement around the corner is properly considered an outlier or whether a point introduced as an outlier is an outlier indeed, it depends on its neighbors, on how the random outliers are distributed around the scan.

The whole process described in Section 3 can be seen in action with real data from a Hokuyo laser sensor in Fig. 6. Even if most of the environment is polygonal, the regions of the polygonal surfaces are interrupted by random

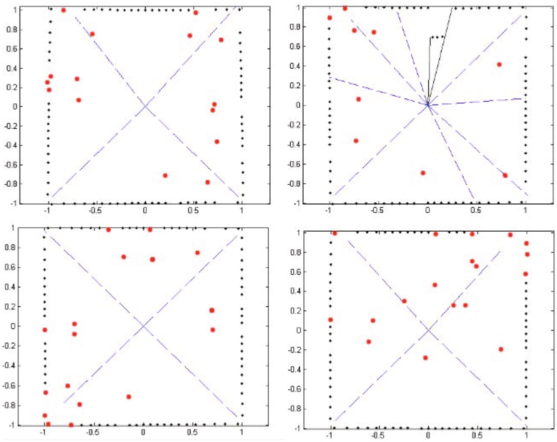
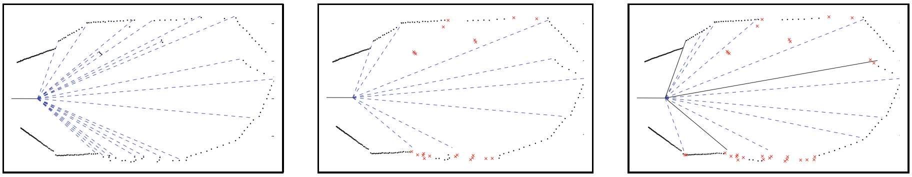


Fig. 5. Output of some experiments with outliers. Identified outliers, depicted as bigger dots, were eliminated from INMODEL



(a) Topology $\mathcal{T}$ after the first iteration

(b) Intermediate iteration

(c) Final topology

Fig. 6. The outer level optimization (Algorithm 1) works on the discrete variables, deciding which sensor readings can be described by the prior (variable INMODEL) and the division in regions/surfaces (variable $\mathcal{T}$). The pictures show the evolution of the topology. Solid lines indicate boundaries between surfaces; dashed lines indicate borders between regions. Dark crosses indicate readings outside the INMODEL set. (a): Some decisions on $\mathcal{T}$ can be taken based on the geometric constraints and the knowledge of the maximum curvature, specified in the prior. (b): The rest of the algorithm guesses where the region/surfaces boundaries are based on a greedy relaxation algorithm. Clutter and outliers tend to be isolated in small regions that are later removed. (c): The final result is feasible according to the prior; the outliers have been removed from the set INMODEL.

clutter and outliers (Fig. 6a). The first part of the relaxation introduces several breaks around outliers (Fig. 6b) producing a very fragmented topology. Then we remove the clutter from INMODEL and we can return to the simple correct topology for the rest of the points (Fig.6c).

Fig. 7 shows some more experiments with real data acquired at Principe Felipe Science Museum in Spain. Most outliers come from glass panels and people.

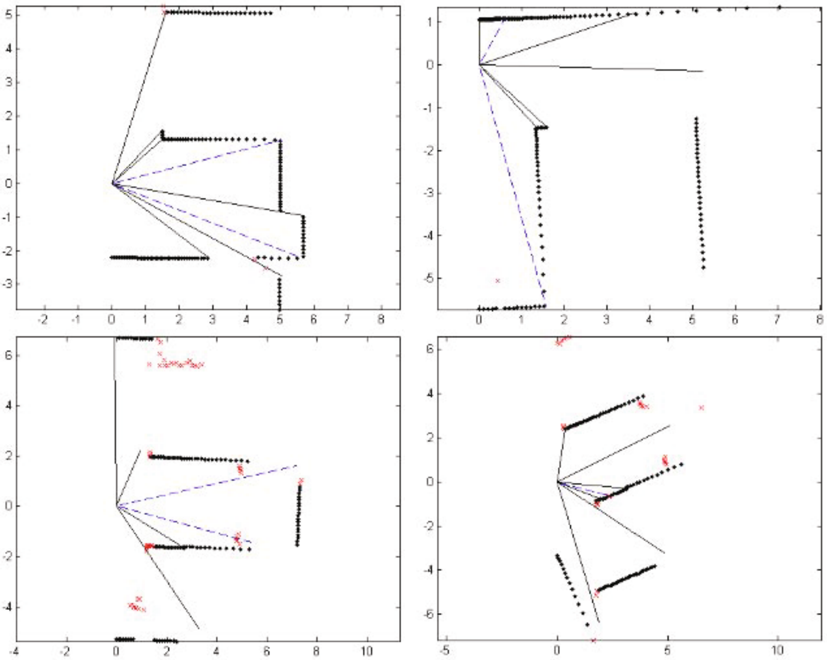
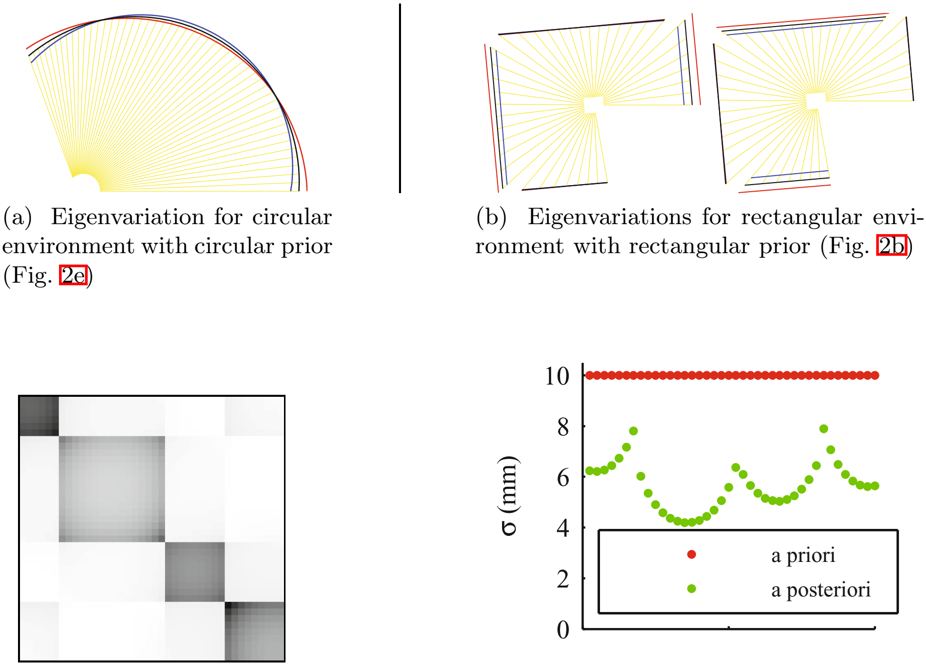


Fig. 7. More experiments with real scans. Data from Principe Felipe Museum (Valencia, Spain), highest floor

Regarding the degrees of freedom extraction, the system recognizes well, for instance, that a circular environment has only one intrinsic degree of freedom (the radius of the circle) (Fig. 8a). A rectangular environment has two degrees of freedom (Fig. 8b) using the rectangular prior (Fig. 2b), but 5 if we use the polygonal prior (Fig. 2a), because the walls orientation is not constrained. Our system recognizes well the degrees of freedom in more complicated situations. For example, an environment with two circles has three degrees of freedom (the radii and the distance between the centers); however, they quickly become hard to visualize. Moreover, in the figures we plot only the variation of the readings because the variations of the normals are hard to visualize as well.

Fig. 8 also shows an example of covariance shrinkage with a rectangular environment. We assume that the initial covariance of $\tilde{\boldsymbol{\rho}}$ is band-diagonal with slight correlation across neighbours. After the projection, the posterior covariance (Fig. 8c) correlates readings corresponding to the same surface or region. As shown in Fig. 8d, there is a dramatic uncertainty reduction.



(a) Eigenvariation for circular environment with circular prior (Fig. 2e)

(b) Eigenvariations for rectangular environment with rectangular prior (Fig. 2b)

(c) A-posteriori readings covariance

(d) Readings standard deviations

Fig. 8. After we have found the solution to the optimization problem, the inference engine uses the knowledge of the prior for extracting the intrinsic degrees of freedom (*scan eigenvariations*) and for *shrinking* the covariance by projecting it onto the Free subspace. (a): For example, the inference engine can recognize that a circular environment has one allowed scan eigenvariations. (b): In the case of a rectangular environment and rectangular prior (Fig. 2b), we find 2 allowed scan eigenvariations. These can be interpreted as the variations of width and height of the environment. (c): We can shrink the a-priori readings covariance by projecting it onto the constraints. In this case, we assume that the a-priori covariance (not shown) has slight correlation between consecutive readings. The a-posteriori covariance has very low rank, and distant readings become correlated because of the structure. (d): The shrinking can be visualized by plotting the diagonal elements of a-priori and a-posteriori covariance.

6 Conclusions and Future Work

We have shown how to build an inference engine that can use different priors with the same representation. The priors are defined by the user in a domain-specific language. The problem of approximating the map posterior is turned into a constrained optimization problem and a covariance projection over the unconstrained directions. This way, it is possible to apply structured priors (polygonal, rectangular, etc.) using a unified representation. Besides being able to reason in terms of regions and surfaces, one can recover the "structure" information under the form of scan eigenvariations, using the degrees-of-freedom analysis.

As part of future work, we plan to improve the greedy Algorithm 1 by introducing backtracking. The incorporation of automatic methods for the representation of the prior of an environment could be another future contribution. We are also interested in testing how preprocessing different sensor data with our method may help scan matching standard techniques. Finally, we are working on the integration of this algorithm into complete SLAM methods, by using the reduced degrees of freedom for global map optimization.

References

1. Beevers, K., Huang, W.: Inferring and Enforcing Relative Constraints in SLAM. In: Algorithmic Foundation of Robotics VII. Springer, Heidelberg (2008)
2. Censi, A.: On achievable accuracy for pose tracking (2009), http://purl.org/censi/2008/posetracking
3. Chong, K., Kleeman, L.: Sonar based map building for a mobile robot (1997)
4. Grisetti, G., Stachniss, C., Burgard, W.: Improved techniques for grid mapping with Rao-Blackwellized particle filters 23(1), 34–46 (2007)
5. Julier, S., LaViola, J.: On Kalman filtering with nonlinear equality constraints 55(6) (2007)
6. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press (2009)
7. Kümmerle, R., Steder, B., Dornhege, C., Kleiner, A., Grisetti, G., Burgard, W.: Large scale graph-based SLAM using aerial images as prior information. Journal of Autonomous Robots 30(1), 25–39 (2011)
8. Le, H., Kendall, D.G.: The Riemannian structure of Euclidean shape spaces: A novel environment for statistics. Annals of Statistics 21(3), 1225–1271 (1993)
9. Michor, P.W., Mumford, D.: Riemannian geometries on spaces of plane curves. J. of the European Math. Soc. 8, 1–48 (2004)
10. Newman, P.: On the Structure and Solution of the Simultaneous Localisation and Map Building Problem. Ph.D. thesis, U. of Sydney (1999)
11. Nguyen, V., Harati, A., Siegwart, R.: A lightweight SLAM algorithm using orthogonal planes for indoor mobile robotics (2007)
12. Parsley, M.P., Julier, S.J.: Towards the exploitation of prior information in SLAM. In: Proc. of the IEEE Int. Conf. on Intelligent Robots and Systems, IROS (2010)
13. Plagemann, C., Kersting, K., Pfaff, P., Burgard, W.: Gaussian beam processes: A nonparametric bayesian measurement model for range finders. In: Robotics: Science and Systems, RSS (2007)
14. Rodríguez-Losada, D., Matía, F., Jiménez, A., Galán, R.: Consistency improvement for SLAM-EKF for indoor environments (2006)
15. Ruszczynski, A.P.: Nonlinear Optimization. Princeton U. Press (2006)
16. Thrun, S., Burgard, W., Fox, D.: Probabilistic Robotics. MIT Press (2005)
17. Trevor, A.J.B., Rogers, J.G., Nieto, C., Christensen, H.I.: Applying domain knowledge to SLAM using virtual measurements. In: Proc. of the IEEE Int. Conf. on Robotics and Automation, ICRA (2010)

Towards Geometric Mapping for Semi-autonomous Mobile Robots

Georg Arbeiter, Richard Bormann, Jan Fischer, Martin Hägele, and Alexander Verl

Institute for Manufacturing Engineering and Automation, Fraunhofer IPA,
70569 Stuttgart, Germany
georg.arbeiter@ipa.fraunhofer.de

Abstract. Semi-autonomous mobile robots are a promising alternative for tasks that are too challenging for autonomous robots. Especially in an unstructured environment, full autonomy is still far from being realized. In order to enable the human operator to control the robot properly, visualization of the environment is crucial. In this paper, we introduce a pipeline for geometric mapping that uses narrow field of view RGB-D cameras as input source and builds a geometric map of the environment while the robot either is operated manually or moves autonomously. Geometric shapes are extracted from subsequent sensor frames and are clipped and merged in a geometric feature map. Evaluation is done both in simulation and on the real robot.

1 Introduction

Although performance of fully autonomous robots has improved greatly in recent years, they still fail frequently while solving tasks in unstructured environments. This is because of inaccurate sensors and actuators as well as non-robust algorithms. A promising alternative are semi-autonomous robots that try to fulfill common tasks autonomously until an unexpected situation occurs. In this case, a human operator can compensate for the lack of intelligence and accomplish the task manually. Optimally, the robot is able to learn from the human actions and thus increases its degree of autonomy over time.

For both the human operator and the robot, perception of the environment is inevitable. Whereas the robot needs information about its surroundings for localization, collision avoidance and planning of actions, the human operator needs visualization of both the current field of view of the robot and past sensor data in order to be able to understand the environment. Most of the robot's demands can be met with a point map representation whereas a geometric map is suitable for data transfer over network and visualization. This leads to the need of a hybrid environment model, consisting of a point and a geometric map.

In this paper, we propose a pipeline for geometric mapping of the environment with focus on semi-autonomous robots. Both a point and a geometric representation of the environment are created during processing. As sensors, RGB-D cameras are used. Being very cost efficient and having a high frame rate they are advantageous over previously used tilting laser scanners. However, the narrow field of view demands additional

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 114–127, 2012.

processing steps. The method we propose does most of the calculations on single sensor frames instead of the full map. Hence, we meet the requirements of a continuous data flow and are able to create the map while the robot is moving. This is a special need if a human operator is present as he or she needs immediate feedback of the robot's vicinity. As the movements commanded by the human are unforseeable and may be at a wide range of speed, higher robustness of the mapping is needed than for a fully autonomous mapping.

The first processing step is point cloud registration. The clouds are aligned to the robot's map coordinate system using *Iterative Closest Point* (ICP). This step is followed by iterative extraction of geometric features like planes. Each plane is passed to the geometric map afterwards. We propose a novel method for processing these extracted planes: The surfaces are transformed to a common coordinate system. Afterwards, 2-D polygonal clipping is applied followed by a merging step. The merged planes are adjusted corresponding to their relative pose. This increases the robustness against inaccurate plane extraction. Using a polygonal representation, we offer the chance to provide a clear visualization to the human user. Additionally, user input can be used to correct erroneous maps by selecting single shapes, deleting them or changing their position. This is not possible with a bare point cloud representation.

The key contributions of this paper are (1) a mapping pipeline for single frame processing of RGB-D data in order to create a geometric map, (2) a novel approach for generating a geometric map from extracted planes and (3) a geometric map representation that can easily be understood and modified by a human.

The remainder of this paper is structured as follows: Section 2 provides related work regarding geometric mapping, semi-autonomous behaviour and polygon clipping. in Section 3 we present the mapping architecture and algorithms used. The evaluation of the mapping and results are shown in Section 4. The paper concludes with a resume and an outlook on future work.

2 Related Work

Aggregation of geometric maps from point cloud data has been subject to many research activities in recent time. For example, Rusu et al. used a tilting laser scanner to acquire point clouds in [1] and performed planar segmentation using *Random Sample Consensus* (RANSAC) in order to find table surfaces. Further work by Rusu et al. shows semantic object labeling of planar surface structures in kitchen environments like cupboards, tables and drawers [2]. They also create polygonal represenations of extracted surfaces but do not approach the merging problem related to sequnetial mapping. In [3], Nüchter et al. used a combination of ICP and RANSAC for plane extraction in point clouds in order to create a semantic map. However, they stop at the stage of labeled point representations rather than creating a map consistent of geometric shapes. Another mapping pipeline was proposed by [4]. Henry et al. perform RGB-D SLAM using a Kinect camera. They use feature points from the color image, apply RANSAC and ICP based registration including loop closure and finally create a SURFEL representation without deriving geometric shapes. A method for volumetric mine mapping using occupancy griuds was presented in [5]. All these approaches are targeted at fully autonomous robots and do not propose a map representation suitable for human perception. Some of them create maps in stop-and-go fashion which is not applicable for humanly controlled robots.

Semi-autonomous behaviour in perception is not so extensively researched. Some approaches deal with human interaction in navigation. In [6,7], virtual objects are augmented in a camera stream and the user can control the robot to avoid obstacles. However, no 3-D mapping is used in order to improve the users immersion. Goodfellow et al. [8] introduced a system that presents the output of the perception module to the user in order to get feedback about the next action. Recently, Pitzer et al. presented an approach for shared autonomy in perception [9]. The user has to identify objects the robot is not able to recognize. Both approaches focus on the field of object detection and have a much stronger user involvement than our approach. Currently, there is no system known to the authors that is able to create geometric map representations from 3-D data in order to satisfy both the requirements of autonomous and tele-operated mode. Basic ideas of our work were already presented in [10], namely point cloud registration and processing of convex hull polygons. This work is extend within this paper.

Processing of polygons is a common task in computer graphics and gaming. A variety of different approaches ([11,12,13]) to polygon clipping can be found in literature. Comparison of several polygon clipping methods is available in [14]. Most of the methods are limited in the types of polygons they can handle. Also, computational speed varies greatly. A generic solution to 2-D polygon clipping was introduced by Vatti in [15]. His method is able to clip and merge almost any kind of polygons in an efficient way.

3 Methodology

Our mapping pipeline for geometric maps is designed for the use of narrow field of view RGB-D or time-of-flight cameras. The system processes one sensor frame after another and does not use the full map representations for most of the calculations. Only depth information is currently used. We carefully choose algorithms to achieve on-the-fly processing and keep the computational complexity of the system low since processing of sensor frames during robot movement is essential for a tele-operated robot.

3.1 System Architecture

The system architecture is shown in Fig. 1. The first step is point cloud registration. We use a variant called frustum ICP for alignment. ICP is applied on a downsampled sensor frame to reduce computation time. Also, not every sensor frame is processed but only key frames. The output of the registration component is a point map and an aligned key frame. In the feature extraction step, planes are extracted from each key frame in an iterative way in order to find all planes in the current point cloud. Concave hulls of the planes are passed to the geometric aggregation module. The hull polygons are clipped and merged into the geometric map. Finally, merged polygons are adjusted in pose w.r.t. the input polygons.

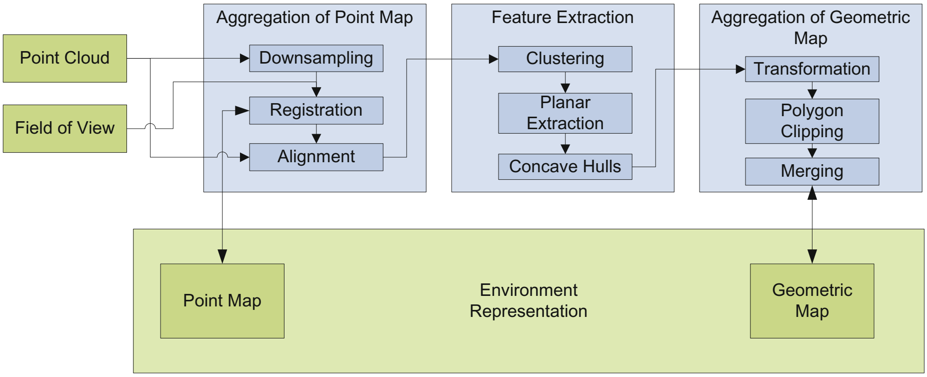


Fig. 1. Architecture of the mapping pipeline

3.2 Registration

Due to the narrow field of view of the camera sensors and the uncertainty of both camera data and robot position, registration of point clouds is inevitable. We rely on the ICP algorithm which is widely used for registration in robotics. However, to bound computational effort in long-term operation, we use the frustum ICP variant [16] that considers the current field of view of the sensor.

First, we select key frames according to the robot movement. This means that we only allow a new point cloud for registration if the robot moved to a certain extent since the last registration event. Each key frame is downsampled using a voxel filter and aligned to the existing point map. However, not the full map is used but only the part being in the current field of view of the sensor. The field of view is modeled as a frustum as described in [10] and for each point in the map an inside-outside test is performed using the normal vectors of the frustum planes. Once the final transformation is found, the original full resolution point cloud is transformed and passed to the feature extraction. Using key frames, we decrease the chance of mis-alignments that might occur if the robot is moved to fast by a human. Also, we evaluate the fittness score of the registration and reject a frame it is not high enough. In this case, we try aligning the next frame.

3.3 Feature Extraction

The next processing step is feature extraction. As features we consider basic geometric shapes like planes, lines or cylinders. In this paper, we use RANSAC to extract planes from each key frame.

The plane extraction is done in an incremental manner. First, Euclidean clustering is used to determine connected regions of the point cloud. In a second step, we try to fit planes to each cluster using RANSAC. If the plane found has at least a minimum number of inliers, it is considered for further processing, removed from the cluster and the fitting step is repeated. We do this as long as no new plane can be fitted or the cluster size falls below a user defined threshold.

It is also possible to set constraints in order to specify what kind of planes should be extracted. For example, if only horizontal planes are of interest for the geometric map, we only consider planes as valid if their normal vector is approximately parallel to the z axis.

In order to describe the extracted planes we use both the plane coefficients of the cartesian form and a concave hull. First, the inliers of the plane are extracted from the cluster and projected on the plane. For a point p the distance δ to a plane

$$pl : \boldsymbol{n} \cdot (\boldsymbol{x} - \boldsymbol{a}) = 0 \tag{1}$$

is defined as

$$\delta = \boldsymbol{n} \cdot (\boldsymbol{p} - \boldsymbol{a}). \tag{2}$$

The projection of p on the plane follows as

$$\boldsymbol{p_{pr}} = \boldsymbol{p} - \delta \boldsymbol{n}. \tag{3}$$

Second, a concave hull of the projected inliers is constructed using alpha shapes. The hull points are sorted so that it is possible to create a polygon. Afterwards, both the hull polygon and the plane coefficients are passed to the geometric map for further processing.

3.4 Aggregation of Geometric Map

As the same scene is observed multiple times from different points of view, many of the extracted planes represent the same objects in the environment and therefore overlap. The goal during aggregation of the geometric map is to merge all planes that describe the same plane in the environment. To achieve this, we use 2-D polygon clipping algorithms.

As polygon clipping is a common task in computer graphics, there are many approaches in literature. However, all of them only work for 2-D polygons. Thus, we have to transform polygons into a common coordinate system if we want to clip them. First, we define a similarity measure for the planes in order to determine candidates for merging. For two planes p_1 and p_2 with

$$p_i : a_i x + b_i y + c_i z + d_i = 0, \; i = 1, 2 \tag{4}$$

the normal vector is

$$\boldsymbol{n_i} = \begin{pmatrix} a_i \\ b_i \\ c_i \end{pmatrix} \tag{5}$$

The similarity measure is defined as

$$\|\boldsymbol{n_1} \cdot \boldsymbol{n_2}\| > t_1 \tag{6a}$$

$$d_1 - d_2 < t_2 \tag{6b}$$

if the normal vectors of the two planes point in the same direction. If not, one of the normal vectors has to be flipped. The conditions (6) limit the maximum angle and distance deviation. The parameters t_1 and t_2 are user defined and depend on the desired granularity of merging. If the two planes meet the similarity condition, they are transformed into a common coordinate system. We now consider to be p_1 the plane already residing in the map, whereas p_2 is a new plane coming from extraction. The weight δ is introduced for each plane in the map. It is increased after every successful merge to account for the fact that the confidence in this feature increases over time.

The clipping coordinate system of p_1 and p_2 is defined as follows: As the coefficients of the merge candidates are similar but not identical we define a virtual average plane

$$p_3 : \frac{(\delta a_1 + a_2)x + (\delta b_1 + b_2)y + (\delta c_1 + c_2)z + (\delta d_1 + d_2)}{\sqrt{(\delta a_1 + a_2)^2 + (\delta b_1 + b_2)^2 + (\delta c_1 + c_2)^2}} = 0 \tag{7}$$

The weighting factor δ yields a stronger influence of planes that have been merged multiple times before. Thus, the robustness against outlier planes is increased, e.g. if registration is not accurate or plane extraction fails.

The coordinate frame for this plane is defined so that $\boldsymbol{n_3}$ from (5) represents the z-axis $\boldsymbol{z_3}$. The x-axis $\boldsymbol{x_3}$ and y-axis $\boldsymbol{y_3}$ can be chosen freely, they only have to be located on $\boldsymbol{p_3}$ and form a right-hand coordinate system with $\boldsymbol{z_3}$. The origin of coordinate system is calculated by

$$x_o = y_o = z_o = \frac{-(d_1 + d_2)}{n_{3,x} + n_{3,y} + n_{3,z}} \tag{8}$$

which solves the plane equation. The transformation from the world coordinate frame to the p_3 frame can be derived in two steps. First, the rotation matrix is set to

$$\boldsymbol{R} = \begin{pmatrix} x_{3,x} & x_{3,y} & x_{3,z} \\ y_{3,x} & y_{3,y} & y_{3,z} \\ z_{3,x} & z_{3,y} & z_{3,z} \end{pmatrix} \tag{9}$$

using the axes of the p_3 frame. Second, the translation is found by

$$\boldsymbol{t} = \boldsymbol{R} \begin{pmatrix} x_o \\ y_o \\ z_o \end{pmatrix} . \tag{10}$$

The full transformation follows from (9) and (10) as

$$\boldsymbol{T_{w2p}} = \begin{pmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0} & 1 \end{pmatrix} \tag{11}$$

Now, all the points $\boldsymbol{pt_w}$ of p_1 and p_2 can be transformed to the common p_3 coordinate frame by

$$\boldsymbol{pt_p} = \boldsymbol{T_{w2p}pt_w} \tag{12}$$

For the 2-D polygon clipping, we are only interested in the x and y coordinates of the points. We therefore project the transformed points on p_3 using (3). The projection error is usually small since we only try to merge similar planes and can be ignored.

For clipping, we use the algorithm by Vatti [15]. This method is capable of clipping convex and concave polygons. Also, polygons with holes or self-intersecting polygons can be processed. It uses a generic approach to clip polygons of all kinds: The polygons are parsed in a scan line fashion. While doing this, the edges of the two polygons are marked as left or right bound of the new polygon and also as contributing or not contributing to the new polygon, depending on the occurrence of local minima or maxima. If a left and right bound intersection occurs, edge classification schemes are used to construct the merged polygon.

The algorithm can create both the union or the intersection of polygons. We use this to first check, whether two candidates intersect. If this is the case, the union of the two polygons is calculated and returned as the merged polygon. If the polygons do not intersect, merging is rejected. After merging, the points of the merged plane are finally transformed to the world coordinate system using $\boldsymbol{T_{w2p}}^{-1}$ and p_1 is replaced by p_3 in the map. Use of the virtual plane p_3 leads to an adjustment of the plane pose in the map over time. Hence, the pose of inaccurate planes can be improved over multiple merging steps.

4 Evaluation

In order to proof the functionality of our mapping concept, we do a performance evaluation both on simulated and real data. As test environment we choose the kitchen in our robot lab. The focus of the evaluation is on the geometric mapping part. The registration was already evaluated in [10].

4.1 Setup

As robot, Care-O-bot® 3 (Fig. 3a) is used [17] . It is equipped with a Kinect RGB-D camera on an agile head. Laser range finder based localization is used to provide an estimate for the robot pose. The robot is moved back and forth in front of the kitchen. Both the base and the neck are used to move the camera manually by a human user. This means, the motions of the robot are not planned in advance.

From the simulation, we obtained two datasets, the first of them was recorded while moving the robot in front of the empty kitchen. For the generation of the simulated data, we used the gazebo[1] simulator. It performs ray casting in a virtual environment to simulate a 3-D camera. In order to evaluate robustness, we added Gaussian noise at different magnitude. Set sim_{n0} is without noise whereas sim_{n005} and sim_{n02} have added noise with a standard deviation of 0.005m and 0.02m respectively. In the second dataset sim_{n0obj} we added non-planar objects to the scene in order to make plane extraction more challenging.

[1] `http://playerstage.sourceforge.net/gazebo/gazebo.html`

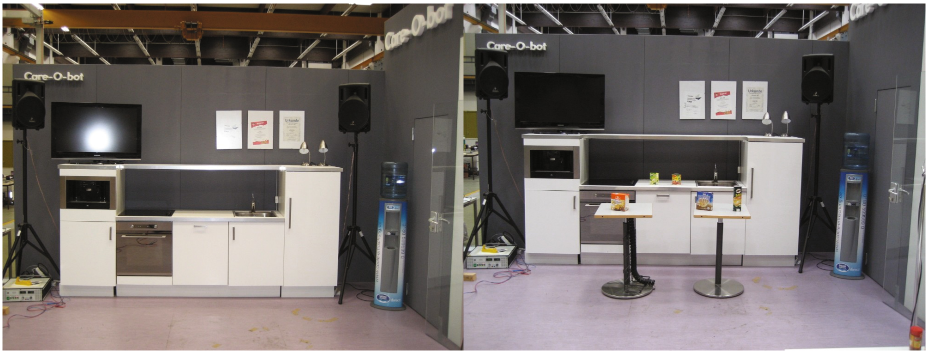


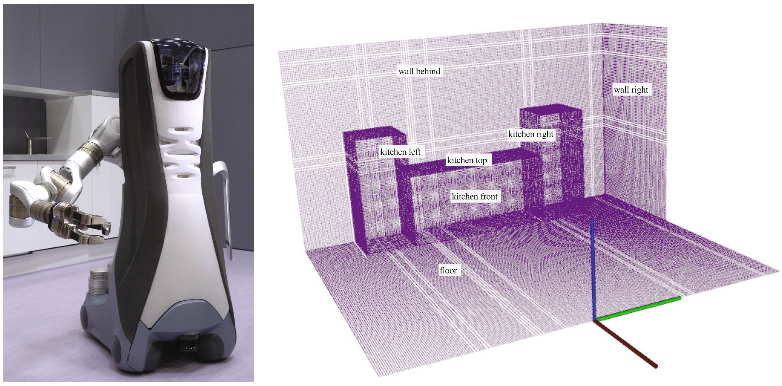
Fig. 2. Kitchen evaluation environment: (a) Set $real_1$, (b) Set $real_2$

With the real robot, two datasets were recorded, one similar to the simulated scene without objects $real_1$ (see Fig. 2a) and one with two additional tables in front of the kitchen and non-planar objects added (see Fig. 2b), $real_2$.

We evaluate the accuracy of the resulting geometric map and the level of plane reduction by merging. Also, we take a look at the mis-detection of planar surfaces if curved surfaces are present. For accuracy evaluation, we created a reference point map of the environment, based on manually measured data, see Fig. 3b. Seven ground truth planes were labeled by hand. The extracted planes are associated with and compared to the ground truth planes. We evaluate the deviation of the plane coefficients d_{coeff}, the angle and distance error d_{angle} and d_{dist} based on the coefficients and the point-to-plane errors RMS_p of the hull points between associated planes. The parameters in (6) are set to $t_1 = 0.95$ and $t_2 = 0.1$.

4.2 Results

The simulated data is used to proof the concept of our mapping algorithm and to test the robustness against noise. The results shown in table 1 can be interpreted as follows: The coefficient, angle and distance deviation grows with increasing noise. This was expected as the planes cannot be fitted as accurate with higher noise than with lower. Also, the point map quality suffers from a higher noise level. However, the point RMS error does not grow as much as one could expect. This is because planes are only merged if they are similar enough. With increasing noise, the deviation between the planes gets higher what leads to less merged planes but also less point errors. The comparison between the empty scene and the one with objects shows that the non-planar surfaces do not disturb the plane extraction and therefore do not influence map quality. On the whole, the values proof a good map at all levels of noise.



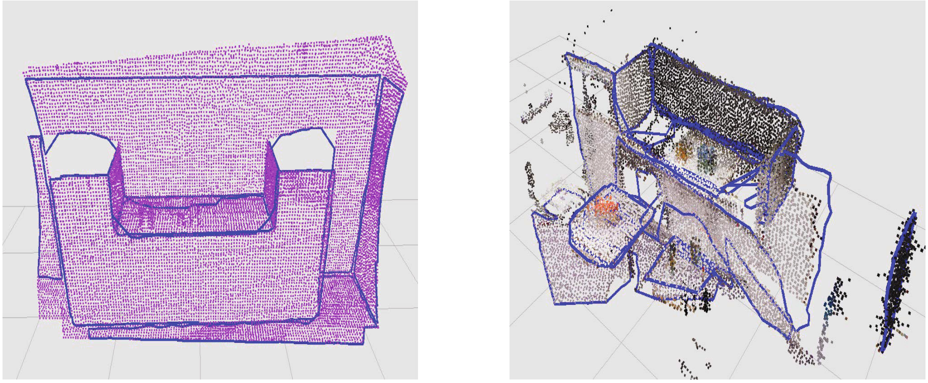
(a) Care-O-Bot® 3

(b) Ground truth

Fig. 3. Mobile service robot Care-O-Bot® 3 (a) and ground truth of the kitchen used for evaluation (b)

Table 1. Accuracy of geometric map for simulated data

set	d_{coeff}	d_{angle} (rad)	d_{dist} (m)	RMS_p (m)
sim_{n0}	0.0108	0.0067	0.0071	0.0054
sim_{n005}	0.0211	0.0057	0.0198	0.0246
sim_{n02}	0.1511	0.0580	0.1368	0.0211
sim_{n0obj}	0.0108	0.0056	0.0084	0.0048



(a)

(b)

Fig. 4. Point and geometric map from (a) simulation and (b) real data. Hull polygons of planes marked blue.

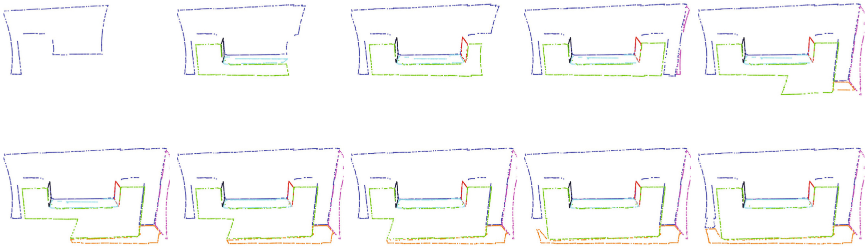


Fig. 5. Merging sequence

Fig. 4a shows the map for the case without noise. It can be seen that all planes are extracted correctly, that all planes belonging together are merged correctly and that the bounds of the planes correspond well with those of the point cloud. In Fig. 5 the merging sequence is shown. While the robot moves, the geometric map grows and so do the planes when additional parts come into view.

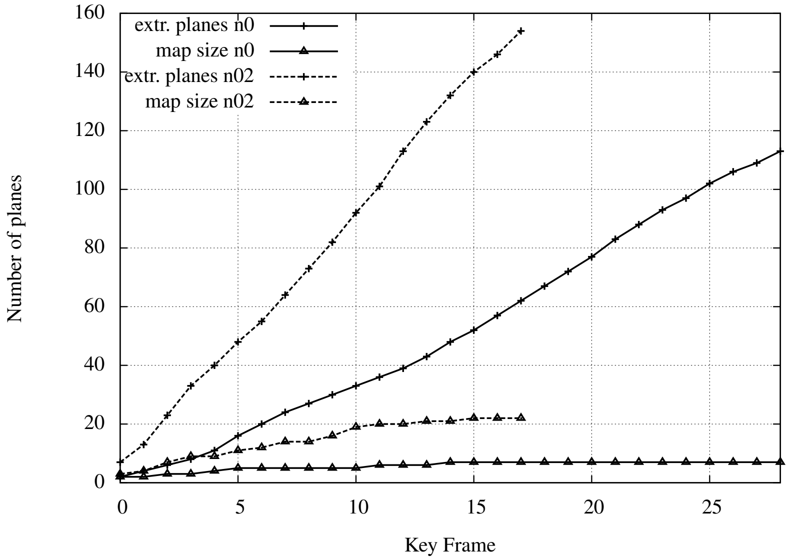


Fig. 6. Map size compared with number of extracted planes

Another interesting measure is the number of planes in the map compared with the number of incoming planes as it shows the power of the algorithm to reduce the number of multiply observed planes. Fig. 6 shows the number of planes over the number of key frames for the empty kitchen at 0 and 0.02 noise. It can be seen, that the number of planes in the map increases until all planes in the scene have been seen once and then stays constant. The plane reduction ratio is 16.14. At a higher magnitude of noise, more planes are extracted and cannot be reduced to the minimum number of planes. Nevertheless, the plane reduction ratio is still at 7.0. For the real robot, the empty kitchen scene and one with added tables and objects are evaluated. In table 2 can be seen that the deviations of the real scene are at a similar magnitude than those from the simulated scene at a noise level of $sim_n 02$. The performance of the mapping is similar in both scenes, also in the one with additional tables and objects. Hence we can conclude that our algorithm works robustly even in cluttered environments.

Fig. 4b shows the map generated from set $real_2$. It can be seen that most of the planes are placed well. Also, the objects on the tables are not detected as planes. However, merging of the planes is not as perfect as with the simulated data and there is an observable deviation of the kitchen front plane. We will investigate this issue in the following.

Table 2. Accuracy of geometric map for real data

set	d_{coeff}	d_{angle} (rad)	d_{dist} (m)	RMS_p (m)
$real_1$	0.1699	0.0692	0.1020	0.0421
$real_2$	0.1520	0.0658	0.1019	0.0317

The noise of the Kinect camera is comparable to the simulated scene but additionally, the Kinect shows distortion especially in regions that are further away or close to the border of the point cloud. This leads to decreased point map accuracy. The geometric map can never be more accurate than the point map as it uses aligned key frames as input. Thus, we take a closer look on the set $real_2$. Table 3 shows the RMS error per plane for the data set. The kitchen front has a significantly higher RMS error compared to the other planes. The explanation is that there is a relatively high distortion of the point cloud data as the plane is rather large. The other point is that the wall behind the kitchen is dominant when it comes to registration. Because of inaccurate range data, a good registration to the wall yields a poor registration of the kitchen front.

Table 3. RMS point error per plane for data set $real_1$

set	wall behind	floor	kit front	kit top	kit left	kit right
$real_1$	0.0160	0.006	0.0783	0.0149	0.0191	0.0055

Timings were measured on simulated data while performing a 360° scan of the kitchen environment. The total sequence has a duration of 90 s. The evaluation was run on a PC with Intel Core i7 @2.80 GHz and 6 GB of RAM. The timings were obtained in 100 runs, each run registering 52 key frames and extracting 152 planes followed by merging. Table 4 shows the results for each processing step averaged per key frame. It shows, that plane extraction is the most demanding step, whereas the merging needs almost no time. The procesing time for all steps is 0.194 s enabling the mapping system

Table 4. Computation time in s of the mapping per key frame.

Registration	Plane extraction	Merging	All
0.027	0.165	0.002	0.194

to run at approximately 5 Hz. As only key frames are processed, the system can run at full Kinect frame rate and fast robot movement.

Finally, we take a look on the data reduction potential using the proposed method. For the dataset sim_{n02}, 17 key frames are registered which results in a total of 5222400 raw points. Having three 32 bit values per point, the complete size of the data handled is 62.7 MB. Downsampling after registration reduces the size of the point map to 49572 points or 595 kB. Eventually, the resulting geometry map consists of 535 Points plus 4 parameter values for each polygon. Thies yields a total size of 6.5 kB. The huge potential in data reduction becomes clear if we take a look on the relative numbers: From the raw data points to geometric represenatation, the amount of data is reduced to 0.01% of the original size. Table 5 shows the absolute and relative numbers for data reduction.

Table 5. Data reduction

representation	raw data	point map	geometric map
number of points	5,222,400	49,572	535
amount of data (bytes)	62.7 MB	595 kB	6.5 kB
reduction to (%)	100	0.95	0.01

5 Conclusion

We presented a novel pipeline for geometric mapping with RGB-D data as input. The processing includes point cloud registration, extraction of planar surfaces, construction of concave hulls and aggregation of a geometric map. A 2-D polygon clipping algorithm is used to merge new features to the map.

We evaluated the performance of our mapping both in simulation and on real sensor data and showed that all relevant planes in the scenes were detected with sufficient accuracy. We also showed that the mapping is robust against noise.

For the future, several extensions and improvements are possible. For example, updating the map in a dynamic scene is an interesting and important topic. The current field of view could be used to replace both parts of the point and the geometric map if the environment changes. Another extension would be to use additional geometric shapes like lines or cylinders in order to describe non-planar parts of the environment. To achieve this, concurrent extraction of multiple feature types and an extended geometric map have to be created.

Acknowledgements. This research was partly financed by the EU FP7-ICT-247772 Multi-Role Shadow Robotic System for Independent Living and by the research program ”Effiziente Produktion durch IKT” of the Baden-Württemberg Stiftung, project ”ATLAS”.

References

1. Rusu, R.B., Blodow, N., Marton, Z.C., Beetz, M.: Close-range scene segmentation and reconstruction of 3D point cloud maps for mobile manipulation in domestic environments. In: Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1–6 (2009)
2. Rusu, R.B., Marton, Z.C., Blodow, N., Holzbach, A., Beetz, M.: Model-based and learned semantic object labeling in 3D point cloud maps of kitchen environments. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), St. Louis, MO, USA (2009)
3. Nüchter, A., Surmann, H., Hertzberg, J.: Automatic model refinement for 3D reconstruction with mobile robots. In: 3DIM 2003, pp. 394–401. IEEE Computer Society, Los Alamitos (2003)
4. Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. In: The 12th International Symposium on Experimental Robotics (ISER) (2010)
5. Thrun, S., Hahnel, D., Ferguson, D., Montemerlo, M., Triebel, R., Burgard, W., Baker, C., Omohundro, Z., Thayer, S., Whittaker, W.: A system for volumetric robotic mapping of abandoned mines. In: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA 2003, vol. 3, pp. 4270–4275 (2003)
6. Carff, J., Johnson, M., El-Sheikh, E., Pratt, J.: Human-robot team navigation in visually complex environments. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009, pp. 3043–3050 (2009)
7. Cherubini, A., Oriolo, G., Macri, F., Aloise, F., Babiloni, F., Cincotti, F., Mattia, D.: Development of a multimode navigation system for an assistive robotics project, pp. 2336–2342. IEEE (April 2007)
8. Goodfellow, I., Koenig, N., Muja, M., Pantofaru, C., Sorokin, A., Takayama, L.: Help me help you: interfaces for personal robots. In: Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction, pp. 187–188 (2010)
9. Pitzer, B., Styer, M., Bersch, C., DuHadway, C., Becker, J.: Towards perceptual shared autonomy for robotic mobile manipulation. In: ICRA, pp. 6245–6251 (2011)
10. Arbeiter, G., Hägele, M., Verl, A.: Incremental field of view dependent registration of point clouds and extraction of table-tops. In: 2011 IEEE/ICRA International Conference on Robotics and Automation (ICRA 2011), Workshop on Semantic Perception, Mapping and Exploration (2011)
11. Holwerda, K.: Complete boolean description (1998), http://boolean.klaasholwerda.nl/algdoc/top.html
12. Schutte, K.: Knowledge Based Recognition of Man-Made Objects. PhD thesis, University of Twente (1994)
13. Zalik, B., Gombosi, M., Podgorelec, D.: A quick intersection algorithm for arbitrary polygons. In: SCCG 1998 Conf. on Comput. Graphics and it’s Applicat., pp. 195–204 (1998)

14. Leonov, M.: PolyBoolean . comparison.... (1998), http://www.complex-a5.ru/polyboolean/comp.html
15. Vatti, B.R.: A generic solution to polygon clipping. Commun. ACM 35, 56–63 (1992)
16. Luck, J., Little, C., Hoff, W.: Registration of range data using a hybrid simulated annealing and iterative closest point algorithm. In: IEEE International Conference on Robotics and Automation, vol. 4, pp. 3739–3744 (2000)
17. Parlitz, C., Hägele, M., Klein, P., Seifert, J., Dautenhahn, K.: Care-o-bot 3 - rationale for human-robot interaction design. In: Proceedings of 39th International Symposium on Robotics (ISR), Seoul, Korea (2008)

Tutorial on Quick and Easy Model Fitting Using the SLoM Framework

Christoph Hertzberg*, René Wagner, and Udo Frese

SFB/TR8 – Spatial Cognition, Reasoning, Action, Interaction
Universität Bremen, Postfach 330 440, 28334 Bremen, Germany
{chtz,rwagner,ufrese}@informatik.uni-bremen.de
http://informatik.uni-bremen.de/agebv

Abstract. In many areas of experimental science ranging from robotics to psychophysical research, to evaluation of spatial sensor-data and surveying, model fitting is a ubiquitous subproblem. Often it is not the actual scientific goal but rather the "necessary evil" of calibrating the equipment. This tutorial introduces methodology and a library allowing to solve model fitting problems easily without requiring the user to have an in-depth understanding of this subject.

After a brief introduction to the theoretical background we guide the reader through using all main features of the SLoM C++ framework based on a stereo camera and inertial measurement unit (IMU) calibration example which is solved with less than 70 lines of non-problem specific code, and provide hints on applying SLoM to other classes of problems.

The reader is only assumed to have a working knowledge of C++ and a basic understanding of statistics and 3D geometry.

Keywords: Model Fitting, Tutorial, Least Squares, Optimization, Manifolds, Calibration, SLAM.

1 Least Squares Optimization in a Nutshell

Least squares optimization determines the most likely values of previously unknown (or only vaguely known) model parameters or *variables* from noisy measured data. For this to work, the measured data and the variables need to be linked in a way that can be expressed as a *measurement function*, a function that predicts the measured data given certain values of one or more variables. The error of the predicted data vs. the actually measured data can be used to adjust the variables to minimize the error. If this is done for all variables and all measurements simultaneously this optimization process yields a maximum likelihood solution, i.e., a variable assignment that is most plausible given the measured data as it minimizes the overall error.

As a concrete example, suppose we want to calibrate a digital camera. The variables to be determined are its intrinsic parameters, i.e., a set of numbers

[*] Corresponding author.

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 128–142, 2012.

describing its optics, e.g., the focal length. The measurement data consists of pixel coordinates of some markers detected in an image. The marker positions in the world are known. Then, the measurement function is simply a pinhole camera model which calculates at which pixel coordinates in the image each marker should be seen assuming the camera has certain intrinsic parameters and is located at a certain position and orientation (collectively called *pose* and serving as an auxiliary variable in this example). Now, to determine the maximum likelihood parameters all we need to do is plug an initial guess of all variables, the measurement function and measurement data obtained from different viewpoints into a least squares solver.

In mathematical terms, the above can be captured in a concise but self-contained form (previously presented in [21]) as follows. The model parameters (including auxiliary ones) that we want to fit to the measured data are the random variables $x_{1,\ldots,n}$. Each of the measurements $M_{1,\ldots,m}$ can be represented as a tuple [21]

$$M_i = (z_i, \Sigma_i, f_i, Y_i = \{x_j \,|\, \mathrm{dep}(z_i, x_j)\}). \tag{1}$$

The two most important bits here are z_i, the actually measured datum, and f_i, the so-called measurement function, which returns the expected datum $\hat{z}_i$, i.e., the datum we would expect to measure assuming a set of dependent variables Y_i have certain values. Comparing the two yields the error function to be minimized.

The covariance Σ_i describes the uncertainty of the measurement since, as noted above, the measured data is noisy. More specifically, the measurement errors are assumed to adhere to a normal distribution with mean 0 and covariance Σ_i, i.e., [21]

$$f_i(Y_i) \boxminus z_i \sim \mathcal{N}(0,\, \Sigma_i)\,. \tag{2}$$

You will probably wonder what the curious $\boxminus$ is all about and we will get to that later in the tutorial. For now, it will suffice to think of it as the same as a regular vector subtraction.

So far, we have only looked at individual measurements. We can now form the combined problem as follows. We stack all random variables x_i into the vector X and all individual error functions $f_i(Y_i) \boxminus z_i$ into the "big" combined error function F

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad\qquad F(X) = \begin{bmatrix} f_1(Y_1)\boxminus z_1 \\ \vdots \\ f_m(Y_m)\boxminus z_m \end{bmatrix}, \tag{3}$$

so that we can state the combined least squares problem as

$$\hat{X} = \underset{X}{\operatorname{argmin}}\, \tfrac{1}{2}\, \|F(X)\|^2_{\Sigma}\,. \tag{4}$$

The curious Σ in (4) denotes the normalization of all measurement errors according to their respective covariance, i.e., uncertainty. One can think of this as a weighting of the errors depending on the measurement precision. The take-home message here, however, is that in (4) we have brought a wide range of problems into a form that a least squares optimizer will understand, i.e., if we can describe

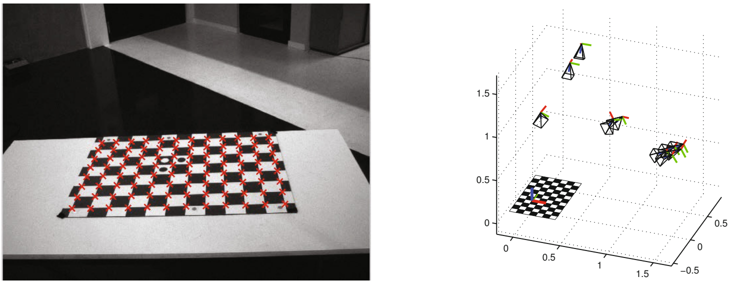


Fig. 1. Results from camera calibration. The left image shows checkerboard corners projected into the image using the estimated parameters. The estimated poses of the cameras w.r.t. the checkerboard are illustrated on the right. Note that for better presentation only the left cameras are displayed. Images are taken from [21].

a model fitting problem as in (1) there is a well-understood black box algorithm that solves it for us. The only other things we need to worry about are that all variables must be observable, i.e., changes of variables always lead to changes in at least one predicted measurement, and that we provide sufficient measurement data to constrain the variables to a unique solution.

Luckily, the above abstraction can be nicely implemented as a software library interface and we will see this in action in the next section.

2 Tutorial

This section will show how to solve model fitting problems using the Sparse Least squares on Manifolds (SLoM) C++-framework. We give actual C++-code, first to be concrete and second because most practical problems are conceptually simple and with SLoM this simplicity carries over to the actual code. SLoM is available as a sub-project of the Manifold ToolKit (MTK) from `http://openslam.org/MTK.html`. MTK uses the Eigen matrix library [9], allowing to write textbook-style matrix expressions. As for the term manifold, for now it will suffice to say that it refers to certain properties all variables have in MTK/SLoM. We will get back to this in §4.

We use the calibration of a stereo camera system with an inertial measurement unit (IMU) as a worked example. The calibration determines all parameters needed to interpret images and inertial measurements spatially. This example is manageable, it is a self-contained realistic application, and shows the main features of SLoM. In particular, such calibration problems tend to have a heterogeneous structure where the SLoM library helps most to avoid complex bookkeeping in the code. The program and an example data set are available from our website `http://www.informatik.uni-bremen.de/agebv/en/pub/hertzbergsc12`. In §3 we discuss the extension to other problems.

Camera calibration involves variables shared by all measurements, e.g., the intrinsic camera parameters which are the same across all calibration images assuming the camera hardware (its optics in particular) remains the same, and variables that are specific to individual (sets of) measurements, e.g., each calibration image was taken from a different camera pose. Depending on the calibration setup an arbitrary number of measurements can be involved. In this example, the measurements are the coordinates of the checkerboard corners detected in the image (Fig. 1) as well the gravity vector observed by the IMU. The checkerboard geometry is known, so the detected checkerboard corners determine all variables involving the camera, i.e., the intrinsic parameters (optics) and each camera pose. Also, the checkerboard is placed such that it is leveled horizontally, so the cameras observe the direction of up and down from the image whereas the IMU observes it from gravity. Several of these pairs determine the orientation of the IMU relative to the cameras.

We will show how to define variables and measurements, then how to initialize and start the optimization process using the SLoM framework.

2.1 Defining Variables

First of all, the variables which are to be optimized have to be defined. There are some shared variables such as the parameters describing the camera optics, the transformation between left and right camera as well as the 3D orientation between the IMU and the cameras. Another internal parameter is the accelerometer bias of the IMU, which we will assume stays constant during the measurements. Furthermore, we need to estimate the amount of gravity (which is different in different geographical locations).

3D orientations are parameterized by the rotation group $SO(3)$, transformations and poses by the Euclidean group $SE(3)$, both of which are readily implemented by MTK. We declare some `typedef`s for convenience:

```
1  typedef MTK::vect<2> vec2;      // 2D vector
2  typedef MTK::vect<3> vec3;      // 3D vector
3  typedef MTK::SO3< > SO3;        // 3D Orientation
4  typedef MTK::trafo<SO3> SE3;    // 3D Transform
5  typedef MTK::Scalar< > Scalar; // Scalar variable
6  typedef MTK::vect<9> CamIntrinsics; // Camera intrinsics
```

The camera intrinsics consist of a number of scalar values, such as the focal length and lens distortion parameters. We will store them into a single vector (line 6). As `CamIntrinsincs` is essentially a C++ `class`, we can inherit from it and add member functions implementing, e.g., the camera's measurement model:

```
7  struct Camera : public CamIntrinsics {
8    vec2 sensor2image(const vec3& point) const;
9  };
```

Next, we combine two `Cameras` and a transformation between them to a single variable defining a stereo camera:

```
10  MTK_BUILD_MANIFOLD(StereoCamera,
11    ((Camera, left))
12    ((Camera, right))
13    ((SE3, left2right))
14  )
```

Here the macro MTK_BUILD_MANIFOLD constructs a new compound manifold, named by the first macro parameter. The second parameter is a list containing the sub-components of the manifold. Each of these is given as a pair specifying the type and the name of the entry enclosed in double parentheses. The macro hides all necessities for SLoM to work with the new manifold.

Again, we could inherit from this class to implement the measurement model for the stereo camera. However, in this case we will do it outside this class later.

2.2 Defining Measurements

Our calibration process will involve two kinds of measurements. Visual measurements of both cameras and accelerometer measurements of the IMU.

As we combined the intrinsics of both cameras to a single variable, we do the same for the measurement. Thus each measurement will depend on a StereoCamera as well as an SE3 describing the pose of the left camera. As measurement data it includes the known position of the checkerboard corner on the plate (considered world frame here) and the corner's pixel coordinates in both camera images. The measurement is declared using the following macro:

```
15  SLOM_BUILD_MEASUREMENT(StereoMeasurement, 4,
16    ((StereoCamera, cam))
17    ((SE3, left2world))
18    ,
19    ((vec3, cornerInWorld))
20    ((vec2, leftMeas))
21    ((vec2, rightMeas))
22  )
```

The first parameter is the name of the measurement, then follows its dimensionality (in this case 4 as we measure two 2D feature positions). The third parameter is a list of dependent variables, again in a double-parenthesized list of types and names, and fourth a list of extra user data which is treated by SLoM as arbitrary constant data that is made available to the measurement model (see below) but not otherwise looked at. Note that the dependent variables need to be manifolds, whereas the extra user data can be of arbitrary types.

Next, we implement the measurement model, i.e., a function that, in terms of (1), computes $f_i(Y_i) \boxminus z_i$. By line 23 SLoM automatically generates the necessary function header, requiring the result to be stored in a real-valued vector ret of the dimension passed as the second parameter to SLOM_BUILD_MEASUREMENT above. It can easily be assigned using the = operator or Eigen's comma initializer as is done here.

```
23  SLOM_IMPLEMENT_MEASUREMENT(StereoMeasurement, ret){
24    vec3 cornerInLeft = left2world->inverse() * cornerInWorld;
25    vec3 cornerInRight = cam->left2right * cornerInLeft;
26    ret << cam->left.sensor2image(cornerInLeft) - leftMeas,
27           cam->right.sensor2image(cornerInRight) - rightMeas;
28  }
```

In the implementation of the measurements, variables and user data members can be accessed by name. Variables need to be dereferenced by the * or -> operator. The measurement function transforms the checkerboard coordinates to the left (line 24) and right (line 25) camera's coordinate system, then applies the appropriate camera projections and subtracts the measured coordinates (lines 26 and 27).

Next, we define the gravitation measurement. This will depend on the accelerometer's bias, the orientation of the accelerometer with respect to the left camera and the position of the left camera in the world. We also assume that we do not know the exact amount of gravitational acceleration, so we add another variable estimating g. Finally, we include the measured acceleration as data member.

```
29  SLOM_BUILD_MEASUREMENT(Gravity, 3,
30    ((vec3, acc_bias))
31    ((SE3, left2world))
32    ((SO3, left2imu))
33    ((Scalar, g))
34    ,
35    ((vec3, acc))
36  )
```

Implementing the measurement requires transforming the local measurement to world coordinates and subtracting the expected gravity:

```
37  SLOM_IMPLEMENT_MEASUREMENT(Gravity, ret){
38    vec3 acc_world = *left2world * left2imu->inverse() * (acc - *acc_bias);
39    ret = (acc_world - *g * vec3::UnitZ());
40  }
```

2.3 Insertion of Variables and Measurements

Once all variables and measurements have been defined, we can collect data and insert it into an `Estimator`. For brevity, we omit the process of obtaining the data and finding initial guesses for the camera poses. The templated `VarID` class is a handle to the actual variable, required to declare measurements and needed to obtain their content after optimization.

```
41  Estimator est; // Estimator, responsible for data management &
        optimization
42  std::vector<vec3> calib_points;  // Known calibration point positions
43
44  // Variables shared by multiple measurements:
45  VarID<StereoCamera> cam = est.insertRV(StereoCamera());
46  VarID<SO3> left2imu     = est.insertRV(SO3());
47  VarID<vec3> acc_bias    = est.insertRV(vec3());
48  VarID<scalar> grav      = est.insertRV(scalar(9.81));
49
50  for(int i=0; i<num_images; ++i){
51    SE3 pose;
52    std::vector<std::pair<vec2, vec2> > point_measurements;
53    // collect image points, get an initial guess for the left camera pose
54    /* LEFT OUT FOR BREVITY */
55
56   // left2world is only local, since we do not need its optimized value
57   VarID<SE3> left2world = est.insertRV(pose);
58   for(int j=0; j<num_points; ++j){
59     est.insertMeasurement(StereoMeasurement(cam, left2world,
60         calib_points[j],
61         point_measurements[j].first, point_measurements[j].second));
62   }
63   vec3 acc;   //insert gravitation measurement:
```

```
64   double acc_sigma = 1e-3;
65  est.insertMeasurement(Gravity(acc_bias, left2world, left2imu, grav, acc),
66                              SLOM::StandardDeviation(acc_sigma));
67  }
```

When inserting measurement, the last parameter (line 66) describes the uncertainty of the measurement. One can choose from multiple ways to represent uncertainty: either as the covariance (`SLOM::Covariance`), as the standard deviation (`SLOM::StandardDeviation`), as the information matrix, i.e., the inverse of the covariance (`SLOM::InvCovariance`) or as the inverse of the standard deviation (`SLOM::InvStandardDeviation`). Each method accepts either a single scalar (as in the example) or a vector describing a diagonal matrix. Covariances can also be passed as full (symmetric) matrix and standard deviations as lower triangular matrix, being the Cholesky factor of the covariance. If the uncertainty parameter is omitted, SLoM assumes the measurement has unit covariance (line 61).

2.4 Optimization and Obtaining the Results

Now that we have inserted all measurements, we can call the `optimize` function of the `Estimator` and read out the optimized values by dereferencing the `VarID` of each variable using the `*` or `->` operator. Note that MTK manifolds overload the streaming operators, so one can easily stream the result into files or to the console.

```
68  for(int i=0; i<100; ++i){
69    est.optimizeStep();
70  }
71  std::cout << "Camera intrinsics " << *cam << "\nleft2imu " << *left2imu
72  << "\nGravity " << *grav << "\nAccelerometer Bias " << *acc_bias << "\n";
```

3 Applying SLoM to Your Own Optimization Problems

Basic data types describing vectors, orientations and transformations are readily implemented. Therefore, SLoM requires no definition of custom manifolds to solve problems such as pose relation and landmark based simultaneous localization and mapping (SLAM). Furthermore, it is easy to combine multiple basic manifolds to combined variables using the `MTK_BUILD_MANIFOLD` macro, which covers more complex problems such as multi-camera bundle adjustment or humanoid robot calibration [21].

With all variables defined, applying SLoM to arbitrary optimization problems boils down to defining custom measurement functions. As shown in the tutorial (e.g., lines 15 to 28) this only requires listing the involved variables and required measurement data, and to implement the actual measurement function.

However, domain knowledge is crucial, i.e., you will first want to understand your problem very well and then expose as much of its structure to SLoM as possible. Usually, this means that whenever possible you want to operate on raw measured data. E.g., if you work with visual markers detected by a camera you want to pass each raw pixel coordinate to SLoM without any pre-processing. It is the job of the measurement function to predict these raw coordinates using a model that is as accurate as possible, i.e., if you intend to correct for radial distortion, you will want to add this as a parameter to the model and take it into account in the measurement function. You do not want to adjust the measured pixel coordinates by pre-processing. Similarly, you should always pass each individual raw measurement to SLoM – do not combine several measurements into one datum, i.e., use individual pixel coordinates as opposed to a some sort of score value you have pre-computed for an entire image.

Afterwards, acquiring the data for measurements is usually more laborious than feeding the data into SLoM and running the optimizer. The user does not need to care about the tedious task of data management – each variable is identified by a type-safe `VarID`, used to initialize measurements and to access the optimized data.

Most calibration problems require an initial guess not too far from the optimum, which if no explicit (approximating) formula exists must be obtained by measuring by hand or preferably re-using a previous calibration.

Also, especially in calibration problems, care has to be taken that the overall problem does not degenerate. This can happen if too few measurements are acquired or if the system contains unapparent gauge freedoms, i.e., non-determined degrees of freedom, such as a non-determined "free floating" start pose in SLAM. In this case one calls the problem rank-deficient and the standard solver will abort immediately. Another hint for a degenerating problem is if the residual sum of squares (i.e., $\|F(X)\|_{\Sigma}^{2}$ from (4)) keeps growing during the optimization process. This can be observed by looking at the results of `est.getRSS()` after each call to `optimizeStep()`.

Besides adding more measurements which constrain the variables better, a possible solution is to fix some variables, e.g.,

```
VarID<vec3> bias = est.insertRV(bias, false); // do not optimize bias
```

then use `bias` as shown in the tutorial and optionally, after optimization "un-fix" it by

```
est.optimizeRV(bias, true); // "un-fix" bias
est.optimizeStep();         // call optimizer again
```

Another solution is to use a more robust optimization algorithm such as Levenberg-Marquardt [15] instead of the default Gauss-Newton:

```
est.changeAlgorithm(new SLOM::LevenbergMarquardt());
```

If for certain variables, there is no measurement depending on it, i.e., the variable is not observable, SLoM automatically raises a warning. This will also make the standard solver fail immediately. Again, this can be circumvented by adding measurements depending on this variable or by fixing this variable (see above).

By alternately inserting variables/measurements and running the optimizer, SLoM is capable to perform online model fitting. In [10], we showed that using this approach it is possible to run full bundle adjustment online for a certain time. However, with increasing number of variables and measurements, the optimization time increases as well thus losing real-time capability, eventually. SLoM allows to fix variables and remove measurements (e.g., following a sliding window approach) to reduce the computation time by reducing the problem size – at the cost of a less optimal result.

4 What's with Those Manifolds?

Although surprising at first, you have already seen the most important property of manifolds – you have hardly taken notice of them at all. This is because we apply a little trick: Locally, a manifold behaves much like an $\mathbb{R}^n$ vector while globally its (topological) structure can be more complex. We can define [11] two encapsulation operators $\boxplus$ ("boxplus") and $\boxminus$ ("boxminus") for a manifold $\mathcal{S}$:

$$\boxplus : \mathcal{S} \times \mathbb{R}^n \to \mathcal{S}, \qquad \boxminus : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^n. \tag{5}$$

Here, $\boxplus$ adds a small perturbation vector to a manifold variable, while $\boxminus$ is its inverse, calculating the difference between two manifold variables (This is the same $\boxminus$ we left unexplained in §1). If the perturbations are small, the manifold suddenly looks like $\mathbb{R}^n$ if you use $\boxplus/\boxminus$ instead of the familiar vector $+/-$.

Why does this matter? In its original form, least squares optimization only works with $\mathbb{R}^n$ variables. However, if we want to deal with real-world spatial data this is insufficient since, most notably, there is no singularity-free representation of 3D orientations with just three parameters – there are always some orientations where a small change in orientation requires a very large change in the representation. Overparameterized (e.g., $\mathbb{R}^4$) representations are not a solution either since the optimizer would try to make use of the extra degrees of freedom which do not actually exist.

We overcome this dilemma by having the least squares optimizer only operate on the local vectorized view established by the $\boxplus$ and $\boxminus$ operators. It does not know about the global structure and still does the right thing as we show in [11] which also gives mathematical details, proofs and experiments.

Although the interested reader is encouraged to read the referenced paper, this is not absolutely necessary even though the above only provides a vague idea of manifolds, $\boxplus$ and $\boxminus$. This is due to another trick: It happens that the Cartesian product of two (and by induction arbitrarily many) manifolds yields another, compound manifold. More importantly, the $\boxplus/\boxminus$ operators of that compound manifold are simply the operators of its components (or sub-manifolds) applied component-wise. This allows `MTK_BUILD_MANIFOLD` to generate compound manifold classes automatically for you if you have implementations of the sub-manifolds. Luckily, MTK comes with implementations of all practically relevant manifold primitives: vectors ($\mathbb{R}^n$), 2D orientations ($SO(2)$), 3D orientations ($SO(3)$) and the (less commonly used) unit sphere (S^2). Thus, based on these, you can build virtually all specialized manifolds you will need without an in-depth understanding of manifolds, $\boxplus$ and $\boxminus$.

5 Related Work

Least squares optimization for model fitting goes back to Gauss [5]. Thus, we will focus on key ideas and readily available tools/libraries here.

The intuition that lead to the use of manifolds in the SLAM community goes back to Triggs et al. [19, p. 6–7] who suggested handling non-vector space states such as orientations by using a global over-representation R with local perturbations δR using a minimal parameterization. The first to explicitly use manifold properties were Ude [20] for least squares optimization and Kraft [13] for Kalman filtering. The extension to arbitrary manifolds and the de-coupling of the optimizer implementation from the variable representation was done in earlier work by the authors [11].

5.1 Problem Specific Frameworks

There are many tools and libraries for specific problems such as camera calibration [2,4,18], bundle adjustment [16] or pose adjustment [17,8,7,6]. They have in common that they are specialized to a specific task and are not easily extensible to more complex problems such as when the sensor setup is non-standard, sensors are added or measurements need to be combined differently.

5.2 Generic Frameworks

Other than SLoM, to our knowledge, there are currently three C++-frameworks aimed at solving arbitrary optimization problems. Namely g^{2}o [14], optimized for fast batch optimization and iSAM [12], focusing on online or incremental optimization problems and the very recently released Ceres solver [3].

While g^{2}o is slightly faster than SLoM, as it exploits the structure of the problem better, iSAM turned out to be slower than SLoM in a recent contest.[1] Very preliminary tests – re-implementing the examples of Ceres using SLoM – showed that Ceres is about as fast as SLoM. Be aware though that, especially for incremental/online optimization, a fair comparison of computation times is difficult, e.g., due to the fact that there is always a trade-off between precision and computation time. However, a more thorough comparison is beyond the scope of this paper.

When handling manifolds, g^{2}o adapts a similar concept as SLoM, whereas iSAM directly maps its variables to vector spaces, thus causing problems when singularities occur. Ceres adds the concept of “local parameterization” which essentially does the same as SLoM’s $\boxplus$ operator, however this is not bound to types but has to be handled individually when registering variables.

Both g^{2}o and iSAM provide basic variable-classes such as vectors, orientations and poses, as well as simple measurements, such as pose relations and basic landmark measurements. This makes them easy to adapt to SLAM problems

[1] For comparisions of computation times see: `http://slameval.willowgarage.com/workshop/talks/2011-RSS-SLAM-Evaluation.pdf`, pp. 17–20.

requiring just these types of measurements. However, if new measurements need to be defined or if variables shall be combined from sub-variables, the user has to implement this from scratch, conforming the internal requirements of the respective framework. As of now, Ceres entirely works on pointers to scalars but provides some convenience functions for often required operations. This requires the user to manually keep track of variable indices. Previous generations of our calibration tools relied on similar techniques and this turned out to be a frequent source of errors which ultimately led to the development of SLoM.

We believe that SLoM's library support for constructing arbitrary compound variable types (manifolds) from primitives and the fact that measurement functions can be directly implemented as C++ functions are the key distinguishing features of SLoM. We try to elaborate this by giving a brief comparison of the different APIs in the appendix.

Beyond the realm of C++, the MTKM framework [21] ports the idea of the SLoM to Matlab and is as such able to solve the same problems. MTKM avoids some syntactic noise which the C++ implementation of SLoM requires. However, due to the poor performance of Matlab's object orientated programming extensions MTKM runs slower by orders of magnitude. We believe that MTKM provides a valuable alternative to users more accustomed to Matlab and not requiring real-time performance (or when working with smaller problem sizes).

6 Conclusions and Future Work

We showed that using SLoM it is possible to solve calibration problems, requiring only basic knowledge of statistics and 3D geometry. Using the same approach most model fitting problems arising in practice can be optimized. By using the $\boxplus/\boxminus$ operators SLoM can solve optimization problems involving manifolds without the user needing to bother about their internal structure.

In future work, we intend to simplify the way measurements are defined even more. Using C++11 features, most importantly variadic templates [1, §14.5.3], it will be possible to define measurements by only defining the actual measurement function:

```
VectorNd measurement(const Var1& v1, const Var2& v2, const Dat1& d1) {
    // do some calculations using v1, v2, ...
    return VectorNd(...);
}
```

and register them to the `Estimator` as such:

```
    VarID<Var1> v1 = est.registerRV(Var1());  // register variables
    VarID<Var2> v2 = est.registerRV(Var2());  // ...
    est.registerMeasurement(measurement, v1, v2, Dat1(...));
```

Furthermore, there is a proof-of-concept implementation allowing the user to easily supply symbolical derivations. By providing derivations of basic functions this will become almost as simple as implementing the actual measurement. Finally, there is on-going research improving SLoM's performance, especially its speed doing online optimization.

Appendix: API-Comparison of SLoM with iSAM and g^2o

We will give a short comparison of the user interfaces of SLoM vs. iSAM and g^2o. We compare the code required to implement and initialize constraints between a 2D pose and a 2D landmark. We believe that for more complicated examples (such as the calibration example described in this tutorial) the differences will be the same or even more evident. For Ceres we did not find a readily implemented 2D SLAM example, from which we could excerpt code to make a fair comparison.

Note that the three libraries use different nomenclatures for the same concepts. What is called "variables" and "measurements" in SLoM is called "nodes" and "factors" in iSAM and "vertices" and "edges" in g^2o.

We start looking at the definition of measurements. Listing 1 shows how a factor is defined in iSAM. Compared to that Listing 5 does the same for SLoM, but avoids much boiler-plate code by the use of macros. In g^2o (Listing 3) some initialization requirements are automatically done by the `BaseBinaryEdge` base class, but most of the initialization is done manually when inserting the edge (Listing 4). In contrast, iSAM (Listing 2) and SLoM (Listing 6) essentially provide initialization in a single statement. Uncertainty information must be provided as inverse covariance in g^2o and inverse standard deviation (or "square root information matrix" as they call it) in iSAM. SLoM is more flexible at this point and even allows to omit passing uncertainty information if the measurement already has unit-covariance. We give a more detailed discussion for each code fragment in its caption.

Listing 1. Defining a 2D landmark measurement in iSAM. Notice line 9 requires manually registering nodes to the factor, and lines 12 and 13 require manually converting the generic nodes to the respective types. The implementation of the actual measurement function starts at line 14. Code excerpted from `https://svn.csail.mit.edu/isam/include/isam/slam2d.h`, LGPL v2.1.

```
 1  class Pose2d_Point2d_Factor : public FactorT<Point2d> {
 2    Pose2d_Node* _pose;
 3    Point2d_Node* _point;
 4  public:
 5    Pose2d_Point2d_Factor(Pose2d_Node* pose, Point2d_Node* point,
 6        const Point2d& measure, const Noise& noise)
 7      : FactorT<Point2d>("Pose2d_Point2d_Factor", 2, noise, measure),
 8        _pose(pose), _point(point) {
 9      _nodes.resize(2); _nodes[0] = pose; _nodes[1] = point;
10    }
11    Eigen::VectorXd basic_error(Selector s = LINPOINT) const {
12      Pose2d po(_nodes[0]->vector(s));
13      Point2d pt(_nodes[1]->vector(s));
14      Point2d p = po.transform_to(pt);
15      Eigen::VectorXd predicted = p.vector();
16      return (predicted - _measure.vector());
17    }
18  };
```

Listing 2. Insert a 2D landmark observation using iSAM. Uncertainty information is passed as square root information matrix `noise2`. Code excerpted from `https://svn.csail.mit.edu/isam/examples/example.cpp`, LGPL v2.1.

```
Pose2d_Point2d_Factor* measurement =
  new Pose2d_Point2d_Factor(pose_nodes[1], new_point_node, measure, noise2);
slam.add_factor(measurement);
```

Listing 3. Defining a 2D landmark measurement in g^{2}o (excerpt from tutorial code `https://svn.openslam.org/data/svn/g2o/trunk/g2o/examples/tutorial_slam2d/edge_se2_pointxy.h`, LGPL v3). The `BaseBinaryEdge` class implements most boiler-plate code required for implementing binary edges (i.e., between two vertices). Still lines 8 and 9 require manually converting the generic vertices to the respective types. Only the last line defines the actual measurement function.

```
 1  class EdgeSE2PointXY
 2    : public BaseBinaryEdge<2, Vector2d, VertexSE2, VertexPointXY>
 3  {
 4  public:
 5    EdgeSE2PointXY() : BaseBinaryEdge<2,Vector2d,VertexSE2,VertexPointXY>() {}
 6    void computeError()
 7    {
 8      const VertexSE2* v1 = static_cast<const VertexSE2*>(_vertices[0]);
 9      const VertexPointXY* l2=static_cast<const VertexPointXY*>(_vertices[1]);
10      _error = (v1->estimate().inverse() * l2->estimate()) - _measurement;
11    }
12  };
```

Listing 4. Insert a 2D landmark observation using g^{2}o. Note that `optimizer.vertex()` does not do type checking, therefore possible errors are detected only at runtime. The `setMeasurement` function is typesafe, however it allows to pass only a single observation object, thus more complex observations need to be combined manually. Code excerpted from `https://svn.openslam.org/data/svn/g2o/trunk/g2o/examples/tutorial_slam2d/tutorial_slam2d.cpp`, LGPL v3.

```
    EdgeSE2PointXY* landmarkObservation = new EdgeSE2PointXY;
    landmarkObservation->vertices()[0] = optimizer.vertex(p.id);
    landmarkObservation->vertices()[1] = optimizer.vertex(l->id);
    landmarkObservation->setMeasurement(observation);
    landmarkObservation->setInverseMeasurement(-1.*observation);
    landmarkObservation->setInformation(information);
    optimizer.addEdge(landmarkObservation);
```

Listing 5. Declare and implement a 2D landmark observation using SLoM. Note that all variables can be referenced by name and are strongly typed.

```
SLOM_BUILD_MEASUREMENT(LM_observation, 2, ((Pose, pose)) ((vec2, landmark)),
    ((vec2, coords))
)
SLOM_IMPLEMENT_MEASUREMENT(LM_observation, ret) {
  ret = ( pose->inverse() * (*landmark) ) - coords;
}
```

Listing 6. Insert a 2D landmark observation using SLoM. The measurement is constructed by an auto-generated constructor. All arguments of the constructor are strongly typed, leading to compile-time errors if wrong types are passed. Passing covariance information is optional (by default unit-covariance is assumed).

```
    est.insertMeasurement(
        LM_observation( poseID, landmark[id], observation),
        SLOM::InvCovariance(information)
    );
```

References

1. Working draft, standard for programming language C++ (2008), `http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2008/n2798.pdf`
2. OpenCV (2012), `http://opencv.willowgarage.com`
3. Agarwal, S., Mierle, K.: Ceres solver (2012), `http://code.google.com/p/ceres-solver/`
4. Bouguet, J.-Y.: Camera calibration toolbox for Matlab, `http://www.vision.caltech.edu/bouguetj/calib_doc/`
5. Gauss, C.F.: Theoria combinationis observationum erroribus minimis obnoxiae. Commentationes Societatis Regiae Scientiarum Gottingensis Recentiores 5, 6–93 (1821)
6. Grimes, M.K., Anguelov, D., LeCun, Y.: Hybrid hessians for flexible optimization of pose graphs. In: Proc. International Conference on Intelligent Robots and Systems, IROS (2010)
7. Grisetti, G., Kümmerle, R., Stachniss, C., Frese, U., Hertzberg, C.: Hierarchical optimization on manifolds for online 2d and 3d mapping. In: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA (2010)
8. Grisetti, G., Stachniss, C., Grzonka, S., Burgard, W.: A tree parameterization for efficiently computing maximum likelihood maps using gradient descent. In: Robotics Science and Systems (2007)
9. Guennebaud, G., Jacob, B., et al.: Eigen 3.1 (2012), `http://eigen.tuxfamily.org`
10. Hertzberg, C., Wagner, R., Birbach, O., Hammer, T., Frese, U.: Experiences in building a visual SLAM system from open source components. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China, pp. 2644–2651 (May 2011)
11. Hertzberg, C., Wagner, R., Frese, U., Schröder, L.: Integrating generic sensor fusion algorithms with sound state representations through encapsulation of manifolds. Information Fusion (2011) ISSN 1566-2535, doi:10.1016/j.inffus.2011.08.003
12. Kaess, M., Ranganathan, A., Dellaert, F.: iSAM: Incremental smoothing and mapping. IEEE Trans. on Robotics (TRO) 24(6), 1365–1378 (2008)
13. Kraft, E.: A quaternion-based unscented Kalman filter for orientation tracking. In: Proceedings of the Sixth International Conference of Information Fusion, vol. 1, pp. 47–54 (2003)
14. Kümmerle, R., Grisetti, G., Strasdat, H., Konolige, K., Burgard, W.: g2o: A general framework for graph optimization. In: Proc. of the IEEE Int. Conf. on Robotics and Automation (2011)
15. Levenberg, K.: A method for the solution of certain non-linear problems in least squares. The Quarterly of Applied Mathematics (2), 164–168 (1944)

16. Lourakis, M.I.A., Argyros, A.A.: The design and implementation of a generic sparse bundle adjustment software package based on the Levenberg-Marquardt algorithm. Technical Report 340, Institute of Computer Science - FORTH, Heraklion, Greece (August 2004), http://www.ics.forth.gr/~lourakis/sba
17. Olson, E., Leonard, J., Teller, S.: Fast iterative optimization of pose graphs with poor initial estimates. In: Proceedings 2006 IEEE International Conference on Robotics and Automation, pp. 2262–2269 (2006)
18. Svoboda, T., Martinec, D., Pajdla, T.: A convenient multi-camera self-calibration for virtual environments. PRESENCE: Teleoperators and Virtual Environments 14(4), 407–422 (2005)
19. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle Adjustment – A Modern Synthesis. In: Triggs, B., Zisserman, A., Szeliski, R. (eds.) Vision Algorithms 1999. LNCS, vol. 1883, pp. 298–375. Springer, Heidelberg (2000)
20. Ude, A.: Nonlinear least squares optimisation of unit quaternion functions for pose estimation from corresponding features. In: Proc. of the Int. Conf. on Pattern Recognition (1998)
21. Wagner, R., Birbach, O., Frese, U.: Rapid development of manifold-based graph optimization systems for multi-sensor calibration and SLAM. In: Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011), San Francisco, California (2011)

Are Left-Right Hemisphere Errors in Point-to-Origin Tasks in VR Caused by Failure to Incorporate Heading Changes?

Bernhard E. Riecke

Simon Fraser University, Surrey, BC, Canada
ber1@sfu.ca
http://iSpaceLab.com/Riecke

Abstract. Optic flow displays are frequently used both in spatial cognition/psychology research and VR simulations to avoid the influence of recognizable landmarks. However, optic flow displays not only lead to frequent misperceptions of simulated turns, but also to drastic qualitative errors: When asked to point back to the origin of locomotion after viewing simulated 2-segment excursions in VR, between 40% (Riecke 2008) and 100% (Klatzky et al., 1998) of participants responded as if they failed to update and incorporate the visually simulated turns into their responses. To further investigate such "NonTurner" behaviour, the current study used a wider range of path geometries that allow for clearer disambiguation of underlying strategies and mental processes. 55% of participants showed clear qualitative pointing errors (left-right hemisphere errors), thus confirming the reliability of the effect and the difficulties in properly using optic flow even in high-quality VR displays. Results suggest that these qualitative errors are not caused by left-right mirrored responses, but are indeed based on a failure to properly incorporate visually presented turns into point-to-origin responses. While the majority of these qualitative errors could be attributed to NonTurner behaviour as previously proposed, we identified a novel, modified NonTurner strategy that could reconcile prior findings. Finally, results suggest that Turners (which properly incorporate visually presented turns) might use online updating of the homing direction, whereas NonTurners resort to more effortful and cognitively demanding offline strategies. Better understanding these strategies and underlying processes and how they depend on stimulus and display parameters can help to inform the design of more effective VR simulations.

1 Introduction

How do we remain oriented while navigating through our environment? For both rotations and translations, the directions and distances between ourselves and surrounding objects of interest constantly changes when we move. Nevertheless, we often manage to remain oriented with seemingly little conscious effort, at least for shorter travels (May and Klatzky, 2000; Presson and Montello, 1994; Rieser, 1989). Whenever unique and recognizable features ("landmarks") are available, they provide a reliable means to remain oriented or recover orientation after disorientation. Hence, such **landmark-recognition based navigation** (or "**piloting**") is widely used whenever suitable landmarks are available (for extensive reviews, see Gallistel, 1990; Golledge, 1999; Loomis et al., 1999).

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 143–162, 2012.

Path integration is an alternative (and often complementary) approach for remaining oriented, and is based not on position-fixing, but on the continuous integration of velocity and acceleration information during travel (Loomis et al., 1999). Especially when landmarks are temporarily unavailable or unreliable (e.g., in fog or heavy snowfall, thick forest, or darkness), path integration plays a vital role and allows the navigator to remain oriented, at least for some time. For increasing time and distance of travel, however, path integration is prone to accumulating errors due to the integration process. Nevertheless, path integration can provide the basis for an automatic and robust continuous spatial updating mechanism that enables observers to remain oriented with little if any cognitive load or effort (Farrell and Robertson, 1998; Presson and Montello, 1994; Riecke, 2003; Rieser, 1989). It can thus serve as a reliable (as largely automated) backup mechanism should piloting ever fail. Moreover, path integration and spatial updating of our immediate environment can provide the scaffolding for learning landmarks and building up configural knowledge, even in animals as seemingly simple as desert ants (Müller and Wehner, 2010).

In order to disentangle the influences of piloting and path integration, the current study used an immersive, projection-based virtual reality setup. This enabled us to exclude all landmarks and focus solely on human visual path integration under full stimulus control and repeatability that is difficult to achieve in real-world settings. A typical and ecologically inspired experimental paradigm to study path integration in animals including humans is to require them to travel or point back to the origin of locomotion ("home") after an actual or simulated excursion (for reviews, see Etienne and Jeffery, 2004; Loomis et al., 1999; Maurer and Séguinot, 1995). One of the simplest yet nontrivial homing task is "triangle completion", where navigators are asked to return home after an excursion path consisting of a first straight segment s_1, a subsequent rotation by a given angle γ, and a final straight path segment s_2. Most animals including humans can perform such triangle completion fairly well as long as they are allowed to physically move, even in the absence of any landmark information (e.g., when blindfolded or when landmarks are removed). A similar experimental paradigm uses point-to-origin or turn-to-face-origin tasks at the end of the excursion instead of actual homing (Klatzky et al., 1998; Loomis et al., 1999). Although this does not allow for distance estimates, using pointing instead of locomotion to the origin allows for much shorter response times, as locomotion time is excluded as a potential confound. Experimentally, this enables us to more directly investigate different underlying mental processes and neural substrates, as the time for computing the homing response can be more tightly controlled, and participants do not have additional processing time during the return path (Gramann, 2012; Gramann et al., 2010; Riecke, 2008).

Path integration based on biomechanical and vestibular cues from blindfolded walking is generally believed to be sufficient for eliciting automatic spatial updating of self-to-surround relationships (Farrell and Robertson, 1998; Klatzky et al., 1998; Presson and Montello, 1994; Rieser, 1989). Can visual cues alone, in the absence of any supporting biomechanical or vestibular cues from physical motion, be sufficient to enable similar automatic spatial updating of our surrounding environment? Research suggests that providing a naturalistic, landmark-rich scene in immersive VR can indeed trigger spatial updating that is both **automatic** (in the sense that it occurs automatically and online during simulated self-motion and requires little conscious effort, attention, or deliberate intention) and **obligatory** (in the sense that it is difficult to intentionally suppress or ignore) (Riecke et al., 2007, 2005b). However, when landmarks were replaced by a simple optic flow stimulus, updating performance decreased and the stimulus could more easily be ignored (Riecke et al., 2007). Potentially related to this reduced availability of automatic spatial updating, participants in optic flow-based VR often seem to resort to offline strategies to solve the task at hand. For triangle completion or point-to-origin tasks, such offline strategies can include abstract geometric strategies, mental arithmetics, imagining top-down views or other configural strategies that rely on building up some kind of survey or configural representation of the travelled path and pointing targets (Riecke, 2008; Riecke et al., 2002). Usage of such offline strategies might contribute to the finding that homing or point-to-origin performance often correlates with general mental spatial abilities.

A particularly striking example of strategy switch and resulting qualitative errors has been reported in a seminal paper by Klatzky et al. (1998), when they compared a variety of different locomotion conditions. Using a modified point-to-origin paradigm, participants were asked to physically turn to face the origin as if they had actually walked the 2-segment trajectory and were now at the end of it. While participants performed relatively accurately in a blind walking condition, they showed qualitatively different response patterns when they did not physically move but instead only watched someone else walk the 2-segment path, listened to a verbal description of the trajectory, or watched optic flow fields of the excursion path on a head-mounted display (HMD). That is, whenever participants did not move, they responded as if they did not update their cognitive heading during the turn, but instead responded as if they were at the end of the excursion pathway, but still facing their original orientation, as illustrated in Figure 1. In their study, optic flow presented on a HMD with a field of view (FOV) of 44° × 33° was in general insufficient to elicit spatial updating that enables correct updating of simulated heading changes. Only when the visually simulated rotations were accompanied by matching physical rotations did participants properly incorporate the rotations into their point-to-origin response.

Later studies reported similar failures to properly update rotations that are only visually simulated via optic flow, although the percentage of such “NonTurners” never reached 100% but typically averaged around 50% (Gramann et al., 2005, 2011; Riecke, 2008). To avoid such failures to update visually presented rotation in VR, several researchers have resorted to providing advance feedback training that allowed participants to correct their initial errors (Gramann et al., 2005; Lawton and Morrin, 1999; Mahmood et al., 2009; Riecke et al., 2002; Wiener and Mallot, 2006). But even with advance feedback training, optic flow-based point-to-origin tasks never seem to reach the ease, intuitiveness, and low cognitive load of blindfolded walking tasks, where failures to update rotations are virtually unknown (Easton and Sholl, 1995; Farrell and Robertson, 1998; Klatzky et al., 1998; Loomis et al., 1999). This might, at least in part, be related to to the finding that biomechanical and vestibular cues from blind walking are sufficient to induce automatic and obligatory spatial updating of our immediate surroundings (Farrell and Robertson, 1998; Klatzky et al., 1998), whereas optic flow-based visual cues (i.e., without landmarks) are typically not, often resulting in increased response times and errors (Chance et al., 1998; Klatzky et al., 1998; Lawton and Morrin, 1999; Riecke, 2008; Riecke et al., 2007; Wiener and Mallot, 2006). In a way, this bears similarity to the well-documented difficulty in imagining perspective switches, where response times are fairly long and errors increase the more the to-be-imagined orientation differs from one's physical orientation (Easton and Sholl, 1995; Farrell and Robertson, 1998; May, 1996; Presson and Montello, 1994; Rieser, 1989).

In summary, whenever online automatic spatial updating is induced by the available sensorimotor cues (e.g., from blind walking), participants can and typically do rely on this updating process to maintain orientation and a sense of the homing direction during the excursion path. In situations where the available stimuli are insufficient to elicit online automatic spatial updating (e.g., for verbal descriptions and most optic flow-based displays), however, participants often seem to resort to offline and/or cognitively more demanding strategies such as configural updating or mental arithmetic. One the one hand, this can lead to increased response times and perceived task difficulty. On the other hand, it can lead to qualitative errors such as the failures to properly incorporate self-rotations as discussed above (Gramann et al., 2005; Klatzky et al., 1998; Riecke, 2008).

The current study was designed to further investigate the phenomenon of left-right hemisphere errors such as the failure to incorporate heading changes as proposed by Klatzky et al. (1998). In particular, we used a much wider range of excursion path geometries than prior studies (Avraamides et al., 2004; Gramann et al., 2005, 2010; Klatzky et al., 1998; Riecke, 2008) to be able to disambiguate between different potential underlying strategies and processes. These potential underlying strategies are discussed below and illustrated in Figure 1 for one specific path geometry.

Turner. Among the four strategies discussed here, the Turner strategy is the only one that does not lead to systematic left-right hemisphere errors. Turner behavior is the default expected behavior if participants properly update the (real, simulated, or verbally instructed) orientation changes during the outbound path (see Figure 1, left). Note that systematic and random errors can, of course, still originate from misperceptions of the path geometry and in particular the turning angle, or other systematic or random sources of errors, e.g., during encoding, mental computation or updating of the homing direction, or execution of the pointing response (Fujita et al., 1990; Riecke et al., 2002).

NonTurner. Klatzky et al. (1998) were the first to describe the apparent failure of participants to update heading changes in situations where the rotations were not physically performed. That is, participants responded as if they were still facing their original orientation, as illustrated in Figure 1, right. Klatzky et al. (1998) were, however, careful in stating that "It is possible that subjects also have an imagined heading that is updated but does not govern their response" (p. 297). Indeed, a follow-up study by Avraamides et al. (2004) showed that participants responded correctly if a verbal response (e.g., "left, 120 degrees") was used instead of the body-referenced response of physically turning to face the origin. The authors proposed that participants indeed successfully updated an imagined (or "cognitive") heading in all conditions, but somehow did not use this imagined heading for the bodily pointing response. This might be caused by a reference frame conflict between the updated imagined (or cognitive) heading and their physical (or "perceptual") heading, as discussed in more detail in (Avraamides and Kelly, 2008; Avraamides et al., 2004; Gramann, 2012).

Left-right inversion. Although most of the prior data on left-right hemisphere errors could be explained by such failure to properly incorporate heading changes into point-to-origin responses, Riecke (2008) observed several cases of left-right hemisphere errors which could not be explained by simple failures to properly update and incorporate heading changes into the pointing responses. In their study, the second path segment s_2 was either of the same length or shorter than the first segment s_1, a fact that participants were aware of. These path layouts predict that NonTurners should always point into the rear (posterior) hemisphere, but never into the frontal (anterior) hemisphere. Five of the 17 participants showing consistent left-right hemisphere errors, however, did consistently point into the frontal (anterior) hemisphere for larger turning angles. This led Riecke (2008) to propose that these participants might not have failed to update their heading properly, but instead produced left-right mirrored responses (cf. Figure 1), potentially because they were "initially uncertain about the correct response, or somehow puzzled or distracted by the visual simulation, and initially picked the left-right mirrored response and then continued to do this, resulting in consistent left-right swap errors" (p. 169). In fact, for 2-segment paths where $s_1 = s_2$ (which is most commonly used in the literature), NonTurner and left-right inversion strategies produce identical predictions. Only for unequal segment length do the predictions differ, as illustrated in Figure 1 and 3. This motivated us to include conditions where s_1 and s_2 are vastly different to allow for clearer disambiguation between potential strategies underlying left-right hemisphere errors.

NonTurner pointing to turning position x_1. Finally, careful re-analysis of the five proposed left-right inverter cases in Riecke (2008) suggests an alternative possible strategy that could equally explain those data, but has not been previously described or discussed to the best of our knowledge. That is, we propose here that those participants might also be NonTurners, but instead of pointing to the origin of locomotion as instructed, they consistently pointed to the turning position x_1, as indicated in Figure 1 and 3. Although it is yet unclear why participants might use such a simplified NonTurner strategy, it can easily explain why the five proposed left-right inverter participants in Riecke (2008) pointed in the frontal (anterior) hemisphere for larger turning angles (see Figure 3, middle and bottom plots).

1.1 Goals, Research Questions, and Hypotheses

The current study extends our earlier work (Riecke, 2008) and was designed to investigate a series of research questions and hypotheses as described below. In particular, the study was designed to further our understanding of potential underlying factors and mechanisms leading to systematic left-right hemisphere errors in point-to-origin tasks that do not allow for physical turning.

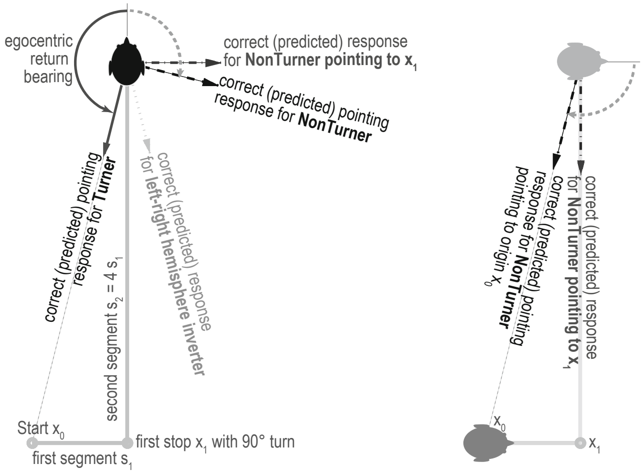

egocentric return bearing
correct (predicted) response for NonTurner pointing to x_1
correct (predicted) pointing response for NonTurner
correct (predicted) pointing response for Turner
second segment $s_2 = 4\ s_1$
correct (predicted) response for left-right hemisphere inverter
Start x_0
first stop x_1 with 90° turn
first segment s_1
correct (predicted) response for NonTurner pointing to x_1
correct (predicted) pointing response for NonTurner pointing to origin x_0
x_0
x_1

Fig. 1. Left: Top-down schematic illustration of predicted pointing responses for the different potential underlying strategies. **Right:** Illustration of NonTurner pointing strategy that does not incorporate the heading change into their pointing response, such that they act as if still facing the original orientation they had at the start position x_0.

Occurrence of left-right hemisphere errors. Similar to earlier studies using optic-flow-based point-to-origin tasks (Gramann et al., 2005, 2010, 2011; Riecke, 2008), we expect around 50% of participants to systematically show qualitative pointing errors, in that they systematically point into the left-right inverted hemisphere (e.g., for left turns they point into the right instead of the left hemisphere).

What processes underly left-right hemisphere errors? As detailed above, a central goal of this study was to disambiguate between the three proposed strategies that might underly left-right hemisphere errors: Left-right inversion, failure to update heading changes (NonTurner), or failure to update heading changes combined with pointing to the turning position x_1 instead of the origin (NonTurner pointing to x_1).

Are left-right hemisphere errors related to problems understanding task instructions and demands? Although previous research consistently showed the existence of left-right hemisphere errors in optic-flow-based point-to-origin tasks unless participants received explicit feedback training, it is conceivable that participants might have somehow misunderstood or misinterpreted the experimental task and procedure. If this were the case, than the occurrence and number of left-right hemisphere errors should decline if participants are provided with advance easy-to-understand task instructions. To this end, participants in the current study completed prior to the VR tests a real-world practice phase, in which they were blindfolded and led to walk along several 2-segment paths at the end of which they pointed back to the origin of locomotion using the identical pointing device as in the later VR experiment. We hypothesized that this task should be easy and lead to almost error-free pointings, and thus exclude all potential misunderstandings of experimental demands in the subsequent VR experiment.

Is the occurrence of left-right hemisphere errors related to general spatial abilities? If so, this would predict that NonTurners would on average show lower spatial abilities (tested here using a standard spatial abilities test as well as self-reported general spatial abilities) as compared to Turners that do not show such left-right hemisphere errors. In addition, we hypothesized that NonTurners might perceive the task as more difficult (which we assessed using post-experimental task difficulty ratings). While Riecke (2008) showed significantly lower spatial abilities test scores for participants showing left-right hemisphere errors, they found surprisingly no signifiant difference in terms of task difficulty ratings. The current study aims to test if these trends can be corroborated.

How do previous point-to-origin results extend to more extreme path geometries? In previous studies, the length of the first and second segment was typically either identical (as in (Gramann et al., 2005) or half of the trials in (Riecke, 2008)), or they were fairly similar in length such that participants might not have realized this or incorporated into their responses. In fact, our previous study (Riecke, 2008) revealed that participants could not reliably assess if the path length of the first and second segment were the same or differed by 50%. When asked to judge the relative length of s_1 versus s_2 in two post-experimental trials, 62.5% responded erroneously for an isosceles excursion path (where $s_1 = s_2$), and 16.7% mistook a path were the first segment was 50% longer than the second segment ($s_1 = 1.5 \times s_2$) as an isosceles path. The current experiment was designed to investigate if and how prior findings might extend to more uncommon path geometries where the first and second path segment have significantly different lengths. To this end, we compared the previously-used isosceles ratio of $s_1/s_2 = 1$ with two more extreme ratios of $s_1/s_2 = 1/4$ and $s_1/s_2 = 4$. Using these path geometries also allowed us to almost double the range of correct egocentric homing directions: Whereas isosceles paths with $s_1 = s_2$ yield correct egocentric pointing directions between 90°-180° (i.e., for left turns the origin will always be somewhere left and behind of the observer), using first segments that are considerably longer than the second segment (here: $s_1/s_2 = 4$) extends this range of correct egocentric pointing directions to almost 0°-180° (i.e., for left turns the origin will always be somewhere to the left, but could now also be in the frontal hemisphere).

2 Methods

Twenty participants (7 female) aged 20-32 years (mean: 24.3) completed the experiment for standard payment. All participants had normal or corrected-to-normal vision. Note that methods of the current experiment were held similar to our earlier study (Riecke, 2008) to allow for direct comparison.

2.1 Stimuli and Apparatus

Throughout the experiment, participants were seated 89cm from a flat projections screen (1.68m × 1.26m, corresponding to a field of view of about 84° × 63°), as illustrated in Figure 2. Visual stimuli were projected non-stereoscopically using a JVC D-ILA DLA-SX21S video projector with a resolution of 1400 × 1050 pixels. The virtual scene was designed to resemble a flat grass plane and provided ample optic flow and high contrast, but no landmarks. To exclude ambient sound that could have interfered with the task, participants wore active noise cancellation headphones (Sennheiser HMEC 300) displaying broad-band masking noise (an unobtrusive mix of river sounds). In addition, black curtains surround the whole setup to ensure that participants could neither see nor hear the actual surrounding lab. Pointing was performed using a modified gamepad, where the central knob was replaced by a 18cm long thin plastic rod to allow for more accurate responses (Riecke, 2008). The pointer was mounted above participants' lap to ensure correct alignment and ease-of-use.

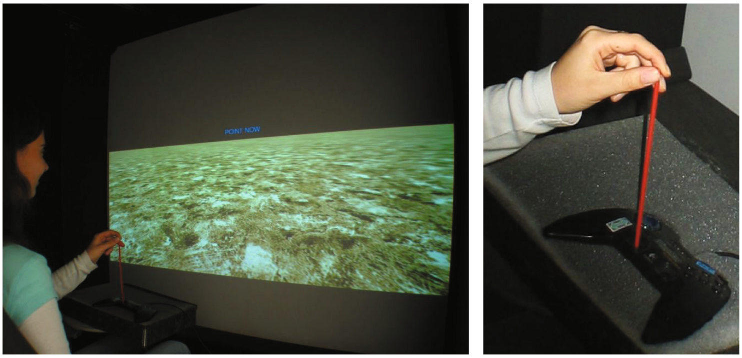


Fig. 2. Experimental setup: Participants with pointing device (modified gamepad) seated behind projection screen showing grass-like ground plane environment devoid of landmarks.

2.2 Procedure and Experimental Design

Participants' task was to point back to the origin of locomotion after visually displayed 2-segment trajectories. Trajectories consisted of a first segment s_1 (8m/s maximum velocity, with brief initial acceleration and final deceleration phase to avoid motion sickness), followed by a turn on the spot (30°/s rotational velocity), and a subsequent second segment s_2 (same velocity profile as s_1). The turning direction was alternated between trials to reduce the occurrence of potential motion aftereffects and motion sickness, but was not analyzed separately as it was not the focus of this study. Hence, the data were pooled over the turning direction for all analyses. Previous research had shown that participants in lab situations tend to resort to computationally expensive cognitive strategies (like mental trigonometry or algebra) to come up with the desired response, especially if response times are unlimited and performance feedback is provided (Gramann et al., 2005; Lawton and Morrin, 1999; Riecke et al., 2002; Wiener and Mallot, 2006). As we were interested in investigating participants' natural and intuitive spatial orientation/spatial updating in VR and reducing the influence of higher cognitive strategies, we instructed participants to point "as accurately and quickly as possible" and to point as if they had physically moved. Participants were never provided with any performance feedback to reduce potential effects of re-calibration and higher cognitive strategies. Using a within-participant design, each participant completed the following phases:

Demonstration Phase. Before starting the experiment, participants gave informed consent and received written and aural instructions. Participants then watched the experimenter perform three randomly selected trials while explaining the experimental procedure and pointing device. Care was taken that the pointing response of the experimenter was random such that participants did not model their responses.

Real-World Practice Phase. A real-world blind-walking point-to-origin pre-test was performed to serve as a baseline for the subsequent VR experiments. To this end, participants were blindfolded and donned the unplugged pointing device. They were led along nine different 2-segment paths, and at the end of each path asked to point back to the origin of locomotion using the pointing device. The experimenter visually judged the accuracy of the pointings. Unknown to the participants, path geometries were a subset of the geometries used in the subsequent VR experiment, in randomized order per participant (length of first segment s_1={1m, 2m, 3m) × turning angle γ={30°, 90°, 150°}; s_2 was adjusted such that the total path length was about 4m). Before the next trial they were led on a circuitous path to a new, randomly selected starting location. Participants responses were virtually error-free, and participants reported that the real-world pointing task was easy and intuitive to perform. Note that none of the participants showed any failures to properly update the rotations, confirming results by Klatzky et al. (1998). For the subsequent VR conditions, participants were instructed to treat the displayed visuals as if they originated from actual self-motion, and to respond as if they had actually moved, just like in this real-world practice phase. These instructions were chosen to ensure that all participants fully understood the experimental demands and in particular the pointing instructions.

2-Segment VR Practice Experiment. In order to reduce the impact of potential learning effects on the main experiment, all participants first performed a VR practice experiment, which used different turning angles than the subsequent main experiments to avoid direct transfer or memorization of turning angles. Each participant completed 14 trials, composed of a factorial combination of 3 lengths of the first straight segment s_1={6m, 15m, 24m) × 2 turning angles γ={60°, 120°} × 2 turning directions (left, right; alternating), plus 2 additional baseline trials without any rotation (γ=0°). s_2 was adjusted such that the total path length was always 30m.

2-Segment VR Main Experiment. Subsequently, participants performed 40 trials in the main 2-segment VR experiment, consisting of a factorial combination of 3 lengths of the first straight segment s_1={6m, 15m, 24m} × 3 turning angles γ={30°, 90°, 150°} × 2 turning directions (left, right; alternating) × 2 repetitions per condition (blocked), plus 4 randomly interspersed baseline trials without any rotation (γ=0°). As before, s_2 was adjusted such that the total path length was always 30m.

Mental Spatial Abilities Test and Debriefing. A standard paper-and-pencil mental spatial abilities test was used to investigate possible correlations between general mental spatial abilities and pointing performance as well as strategy choice (turner vs. Non-Turner) (Stumpf and Fay, 1983). A previous VR study (Riecke et al., 2002) demonstrated significant correlations between triangle completion performance and mental spatial abilities using the same test, such that we expected sufficient sensitivity for the current study. Subsequently, participants were debriefed, paid, and thanked for their participation.

3 Results and Discussion

Pointing data are summarized in Figure 3 and 4. In the real-world practice phase, all participants were able to point back to the origin of locomotion with negligible errors after being blindfolded and led along 2-segment excursion. In the virtual reality conditions, most participants still pointed fairly consistently, as indicated by the circular mean pointing vectors almost touching the unity circle in Figure 3 (Batschelet, 1981). Pointing directions showed, however, considerable between-participants variability as well as systematic pointing errors, especially for larger turning angles, potentially due to a misestimation of the visually presented turning angle.

3.1 Occurrence of Left-Right Hemisphere Errors

In addition to the errors described above, eleven of the 20 participants consistently showed qualitative (and not just quantitative) pointing errors in that they consistently pointed into the wrong (left-right inverted) hemisphere (see Figure 3). That is, for a 2-segment path including a left turn, they pointed to the right hemisphere instead of the left hemisphere and vice versa. Participants consistently showing such left-right hemisphere errors will be termed “NonTurners” in the following, as their behavior might be explained by a failure to properly integrate the visually presented turns into their egocentric pointing response (Avraamides et al., 2004; Gramann et al., 2005; Klatzky et al., 1998; Riecke, 2008). Conversely, participants generally pointing into the correct (instead of left-right inverted) hemisphere will be termed “Turners” here, as they respond as if they update and incorporate the visually presented turns at least qualitatively correct, even though they might misestimate the turning angle.

As pointing data is inherently noisy, we computed the ratio of trials with left-right hemisphere errors per participants to reliably and automatically categorize participants. Nine participants (with IDs 2, 4, 5, 7, 8, 10, 17, 18, and 19) were thus categorized as Turners, with a mean ratio of left-right hemisphere error trials of 7.8%. The remaining eleven participants were categorized as NonTurners, with a mean ratio of hemisphere error trials of 87.4%. Note that none of the Turners or NonTurners showed any left-right hemisphere errors in the prior real-world practice phase. This suggests that the NonTurners’ qualitative pointing errors in VR are neither based on a failure to understand the instructions nor a failure to use the pointing device properly, as the same instructions and pointing device were used in the real-world practice phase.

3.2 Statistical Analysis

Data were analyzed using separate repeated measures ANOVAs for the dependent measures response time and signed pointing error. Independent variables for in the ANOVAs included the within-participant factors turning angle γ and length of the first segment s_1,

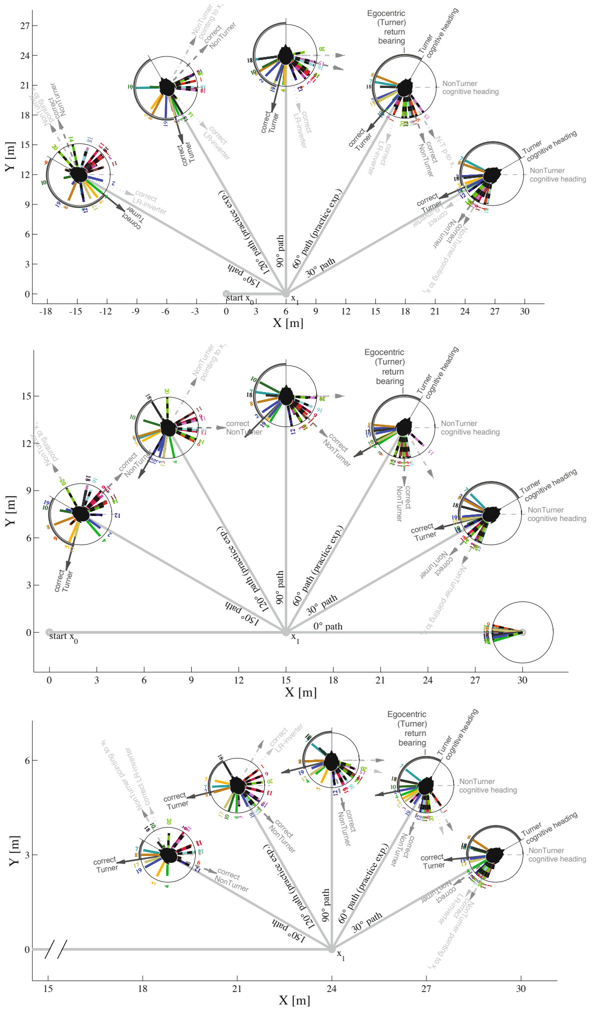


Fig. 3. Top-down schematic view of the outbound 2-segment paths (solid gray lines) for the three different values of s_1. Data from the practice experiment (60° and 120° turns) and the main experiment (30°, 90°, and 150° turns) are combined here for comparability. Circular mean pointing directions for each participant are indicated by solid bars for Turners and dashed bars for NonTurners. Numbers indicate participants numbers. The length of the circular mean pointing vector indicates the consistency of the individual pointing directions: Shorter mean pointing vectors indicate higher circular standard deviations of the individual pointing (e.g., participant 10), whereas mean pointing vectors close to the surrounding black unity circle indicate high consistency and thus low circular standard deviations of the individual pointings (e.g., participant 20; Batschelet, 1981). Correct homing vectors are plotted as a solid black arrow labeled "correct Turner". Predicted pointing vectors for participants that simply show left-right mirrored responses are labeled "correct LR-inverter", whereas predicted pointing vectors for "NonTurner" participants that act as if they did not update their cognitive heading such that they still face the original orientation (0°) are labeled "correct NonTurner".

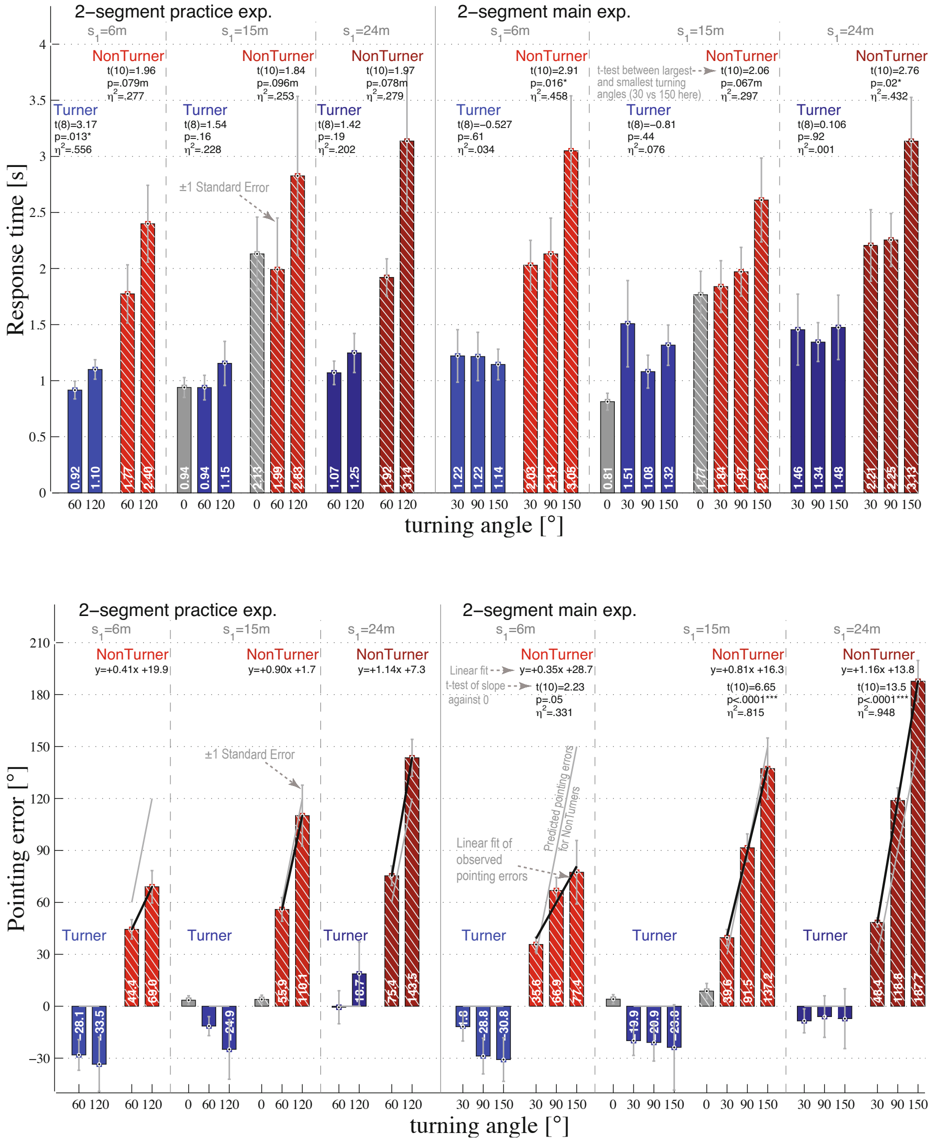


Fig. 4. Arithmetic means ± 1SEM of response time (top) and the signed pointing error (bottom) for the different experiments and path geometries. Turner (solid bars) and NonTurner (hatched bars) are plotted separately to show differences in response patterns.

Table 1. ANOVA results for practice experiment (top) and main experiment (bottom). Significant effects are typeset in bold, marginally significant effect in italics; * p<.05, ** p<.01, *** p<.001.

Practice experiment	Pointing error			Response time		
	F	p	η_p^2	F	p	η_p^2
LR-hemisphere errors	**F(1,18) = 77.9**	**p <.001*****	**.812**	**F(1,18) = 84.0**	**p <.001*****	**.824**
Length of first segment s_1	**F(2,36) = 24.7**	**p <.001*****	**.578**	F(21.3,1.18) = .821	p = .394	.004
s_1 × LR-hemisphere errors	F(2,36) = .670	p =.518	.036	F(21.3,1.18) = .255	p = .658	.014
Turning angle γ	**F(1,18) = 13.7**	**p =.002****	**.433**	**F(1,18) = 8.75**	**p = .008****	**.327**
γ × LR-hemisphere errors	**F(1,18) = 13.6**	**p =.002****	**.430**	*F(1,18) = 3.65*	*p = .072m*	*.168*
s_1 × γ	**F(2,26) = 3.85**	**p =.030***	**.176**	F(1.38,24.9) = .357	p = .625	.019
s_1 × γ × LR-hemisphere err.	F(2,36) = 1.13	p =.335	.059	F(1.38,24.9) = .392	p = .604	.021

Main experiment	Pointing error			Response time		
	F	p	η_p^2	F	p	η_p^2
LR-hemisphere errors	**F(1,18) = 35.7**	**p <.001*****	**.665**	**F(1,18) = 109.9**	**p <.001*****	**.859**
Length of first segment s_1	**F(2,36) = 38.5**	**p <.001*****	**.681**	**F(2,36) = 5.04**	**p = .012***	**.219**
s_1 × LR-hemisphere errors	**F(2,36) = 12.2**	**p <.001*****	**.404**	*F(2,36) = 2.65*	*p = .084m*	*.128*
Turning angle γ	**F(1.18,21.3) = 17.2**	**p <.001*****	**.488**	**F(1.52,27.3) = 6.19**	**p = .010****	**.256**
γ × LR-hemisphere errors	**F(1.18,21.3) = 23.5**	**p <.001*****	**.567**	**F(1.52,27.3) = 6.38**	**p = .009****	**.262**
s_1 × γ	**F(2.49,44.8) = 8.79**	**p <.001*****	**.329**	F(4,72) = .366	p = .832	.020
s_1 × γ × LR-hemisphere err.	**F(2.49,44.8) = 4.07**	**p = .017***	**.184**	F(4,72) = .637	p = .638	.034

and the between-participant factor left-right hemisphere errors (Turner vs. NonTurner, as analyzed above). The baseline condition of γ=0° was excluded from the ANOVAs and data were pooled over the turning direction (left/right) as this was not a focus of the current study. Greenhouse-Geisser correction was applied where needed. ANOVA results are summarized in Table 1.

3.3 Pointing Errors

As expected, overall pointing errors were significantly larger for NonTurners as compared to Turners (cf. Figure 4 and Table 1). Mean pointing errors for NonTurners were 83.1° (standard error: 7.3°) in the practice experiment and 89.2° (SE: 8.1°) in the main experiment, as compared to Turner pointing errors of -13.3° (SE: 8.1°) for the practice experiment and -17.5° (SE 8.9°) for the main experiment. That is, while NonTurners showed a considerable general underestimation of turning angles (as would be predicted if they indeed failed to update the turns), Turners showed a slight overall overestimation of visually presented turns. As indicated in Figure 4, NonTurners showed larger pointing errors for increasing turning angles (as predicted by failure to update rotations). This is corroborated by the linear fit slopes being significantly above zero for all lengths of s_1 (see t-test insets in Figure 4). Especially for s_1 = 15m and s_1 = 24m, NonTurners' overall pointing errors were remarkably close to the values predicted by a failure to update the rotation and pointing as if still being in the original (0°) orientation, depicted as solid gray lines in Figure 4. Pointing errors for Turners, however, showed no overall dependence on turning angles. Although there was large between-participant variability in pointing errors (cf. Figure 3), average pointing errors for Turners as well as NonTurners were fairly close to the predicted values.

3.4 Response Times

Mean response times were relatively low, both in the practice experiment (1.71s, SE: .28) and the main experiment (1.83s, SE: .18s). Turners showed significantly lower response times than NonTurners, both for the practice experiment (1.07s vs. 2.34s, respectively, F(1,18 = 11.6, p=.003**, η_p^2 = .393) and the main experiment (1.31s vs. 2.36s, F(1,18 = 9.04, p=.008**, η_p^2 = .334). For the practice experiment, both Turners and NonTurners showed a tendency towards increased response times for larger turning angles, as indicated by the significant main effect of turning angle γ on response time (see ANOVA results in Table 1), the lack of significant interaction between turning angle and LR-hemisphere errors, and the pair-wise t-tests between smallest and largest turning angles in Figure 4. For the main experiment, however, Turners showed no longer any tendency towards increased response times for larger turns, whereas NonTurners still showed longer response times for larger turns. This is supported by the significant interaction between turning angle γ and left-right hemisphere errors (see Table 1) and t-tests in Figure 4. This dichotomy corroborates the hypothesis that Turners and NonTurners use different underlying strategies to solve the pointing task.

3.5 Correlation between Behavioral and Post-experimental Data

Data from the post-experimental questionnaire and mental spatial abilities test are summarized in Figure 5. Although Figure 5 (a) suggests a tendency for NonTurners to score lower on the mental spatial abilities test (Stumpf and Fay, 1983) than Turners, this tendency did not reach significance. Note that this differs from previous findings by Riecke (2008), who observed significantly lower mental spatial abilities measures for NonTurners. This might be related to a different participant group used and/or insufficient statistical power due to only testing 20 participants in the current study. When asked to rate their everyday spatial orientation ability on a scale from 0-10, NonTurners scored somewhat lower than Turners (6.59 vs. 7.94). This trend did not reach significance, though, corroborating similar findings by Riecke (2008, Experimental series 2).

Similar to findings by Riecke (2008), there was a slight but non-significant tendency for males to perform higher on the mental spatial abilities test and the self-reported spatial orientation ability in the current study (cf. Figure 5 (a) & (b)). Note that we did not find the clear gender effects that are often reported for various spatial abilities (see reviews by Coluccia and Louse, 2004; Lawton and Morrin, 1999). Again, insufficient power and differences in participant population might both have contributed to the lack of gender effects in the current study. Turner and NonTurner did not differ significantly in terms of their age (t(18)=-1.54, p=.14, η^2=.116), amount of daily computer usage (t(18)=-.0218, p=.98, η^2=.10), or rated task difficulty (t(18)=-592, p=.56, η^2=.019). This corroborates our earlier findings (Riecke, 2008). Similarly, there was no significant influence of gender on any of these measures (all p>.17).

4 General Discussion and Conclusions

The current study was designed to investigate the phenomenon of left-right hemisphere errors that occur in point-to-origin tasks where participants do not physically execute the turn between the first and second segment (Avraamides et al., 2004; Gramann et al., 2005, 2010, 2011; Klatzky et al., 1998; Riecke, 2008).

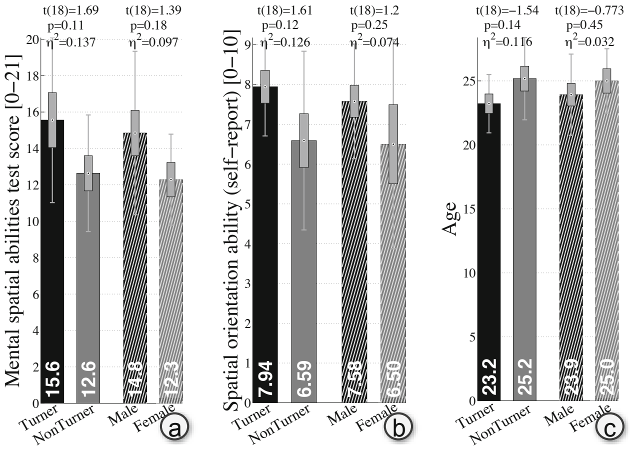


Fig. 5. Data from the post-experimental questionnaire. Boxes and whiskers denote ± 1SEM and ±1SD, respectively. Top insets show results from unpaired t-tests for Turner vs. NonTurner (left solid bars) and gender (right hatched bars).

4.1 Occurrence of Left-Right Hemisphere Errors

The general phenomenon of left-right hemisphere errors was confirmed in the current study, with 55% of the current participants showing such qualitative errors in their pointing responses. As detailed in Table 2, this percentage of left-right hemisphere errors was slightly larger than in (Riecke, 2008), and roughly comparable to Gramann et al. (2005, 2010, 2011). A recent study by Sigurdarson et al. (2012) showed that left-right hemisphere errors can occur even when visually simulated rotations are accompanied by matching physical rotations. This challenges the notion that physical rotations necessarily induce automatic and obligatory spatial updating (Klatzky et al., 1998; May and Klatzky, 2000; Presson and Montello, 1994; Rieser, 1989).

Table 2. Relative distribution of NonTurners amongst male and female participants

Study	Total # participants	% NonTurner	% NonTurner for males	% NonTurner for females
Current	20 (13 male)	55% (11/20)	31% (4/13 males)	100% (7/7 females)
Riecke (2008), Exp. 1	16 (half male)	38% (6/16)	13% (1/8 males)	63% (5/8 females)
Riecke (2008), Exp. 2	24 (half male)	46% (11/24)	33% (4/12 males)	58% (7/12 females)
Current + Riecke (2008)	60	47% (28/60)	27% (9/33 males)	70% (19/27 females)

4.2 What Processes Underly Left-Right Hemisphere Errors?

Using an unusually wide range of triangle geometries in the current study allowed us to use the behavioral (pointing) data to disambiguate between the potential strategies underlying left-right hemisphere errors. First of all, we found no direct support of left-right inversion strategies: As indicated in Figure 3 (top), left-right inversion would have predicted that participants in the $s_2 = 4 \times s_1$ conditions should always point into the far rear (posterior) hemisphere, with little dependence on the turning angle γ. This was not observed. Instead, participants showing left-right hemisphere errors pointed into the far posterior direction for small turning angles and increasingly towards more frontal (anterior) directions for increasing turning angles. While not compatible with left-right inversion, this behavior is compatible with both NonTurner strategies (cf. Figure 3 (top)). Note that participants might have misestimated turning angles (Riecke et al., 2005a), such that we refrain from a more quantitative analysis of the exact pointing angles when trying to disambiguate between potential underlying strategies.

Previous studies showed that participants in general use a chosen strategy quite consistently (Avraamides et al., 2004; Gramann, 2012; Gramann et al., 2005, 2010, 2011; Klatzky et al., 1998; Riecke, 2008). Hence, we assume here that participants did not switch strategy for the different path geometries. This is essential, as left-right inverter and NonTurner (pointing to the origin, not x_1) strategies yield identical predictions for the isosceles path geometries where $s_2 = s_1$. As indicated in Figure 3 (middle), analyzing the pointing data from the isosceles path geometries where $s_2 = s_1$ thus allows us to disambiguate between the NonTurner strategies where participants point to the turning position x_1 and the default NonTurner strategy where they point (as instructed) towards the origin of locomotion x_0. Whereas the latter (default NonTurner) strategy predicts that participants should never point into the frontal (anterior) hemisphere as long as $s_2 \leq s_1$, NonTurners pointing to x_1 would be expected to point into the frontal hemisphere for the largest turning angles (γ=120° and γ=150°). This was indeed observed for one participant (#20, depicted as green dashed line in Figure 3), who pointed into the frontal hemisphere for γ=120° and γ=150°. The remaining ten participant showed pointing behavior roughly consistent with predictions from the default NonTurner strategy, in that they did not point into the frontal hemisphere as long as $s_2 \leq s_1$.

A similar response pattern was observed for the trials where the second segment was much shorter than the first one ($s_2 = 1/4 \times s_1$), as indicated in Figure 3 (bottom): Whereas participant #20 pointed again into the frontal hemisphere for the largest turning angle, the remaining 10 participants always pointed into the rear (posterior) hemisphere, which is consistent with the default NonTurner behavior for $s_2 \leq s_1$.

To complement this visual inspection of the data with an algorithmic and thus less subjective and more easily reproducible approach, we mathematically compared participants' pointing directions with predictions from each of the four proposed strategies: Turner, NonTurner, NonTurner pointing to x_1, and left-right inverter, as illustrated in Figure 3. To this end, we defined an error measure as the absolute difference between observed and predicted pointing directions for each condition and strategy, and used that to algorithmically categorize each participant: e.g., if this error measure was lowest for Turner predictions for a given participant, (s)he was categorized as a Turner. Incidentally, this algorithmic categorization led to identical categorization as this visual inspection described above, thus corroborating the earlier analysis: Participants 2, 4, 5, 7, 8, 10, 17, 18, and 19 were categorized as Turner (as in subsection 3.1 above), participant 20 as NonTurner pointing to x_1, and the remaining participants were categorized as regular NonTurner, with no participant being categorized as left-right inverter.

In summary, the current data suggest that the vast majority of participants showing consistent left-right hemisphere errors indeed simply failed to properly update the visually simulated heading change and respond accordingly, as proposed previously (Avraamides et al., 2004; Gramann, 2012; Gramann et al., 2005, 2010, 2011; Klatzky et al., 1998; Riecke, 2008). While we did not find support for a left-right inversion strategy, one participant consistently seemed to use a different strategy that is inconsistent with the default NonTurner strategy. We hypothesize that this participants did not incorporate heading changes (just as NonTurners), but in addition pointed not to the origin of locomotion as instructed, but instead to the position x_1 where the turn took place. Careful reanalysis of the (Riecke, 2008) data suggests that this strategy (NonTurner pointing to x_1) can indeed explain the data from those 5 participants that pointed into the frontal hemisphere and thus could not simply be explained by a normal NonTurner strategy. Further, carefully designed experiments are needed, though, to corroborate these hypotheses.

Are left-right hemisphere errors related to problems understanding task instructions and demands? In the current study, all participants performed a real-world practice phase, where they were blindfolded and led along several 2-segment excursion paths before being asked to point to the origin of travel using the same pointing device that was used in the subsequent VR experiment. As expected, participants easily understood the task and showed negligible pointing errors. Thus, it seems rather unlikely that the left-right hemisphere errors might be related to participants misunderstanding task instructions and demands.

Is the occurrence of left-right hemisphere errors related to general spatial abilities? Whereas Riecke (2008) observed significantly lower spatial abilities test scores for NonTurners as compared to Turners, the current study showed only non-significant trends, albeit in the same direction. Further experimentation with more participants and thus higher statistical power are needed to investigate if NonTurner behavior is indeed associated with lower overall mental spatial abilities.

How do previous point-to-origin results extend to more extreme path geometries? As participants in Riecke (2008) could not reliably disambiguate between trajectories where the lengths of the first and second segment were identical or differed by 50%, we used a much wider range of relative lengths of $s_2/s_1 = \{1/4, 1/1, 4/1\}$. Post-experimental debriefing indicated that this allowed participants to clearly disambiguate the different ratios of s_2 versus s_1. In general, previous point-to-origin results extended to those more unusual path layouts, yielding similar overall percentages of NonTurners as in previous studies and similar overall pointing response patterns.

4.3 Online Updating versus Offline/After-the-Fact Computation of Homing Direction?

If participants use online updating of the visually presented turns as is typically observed for automatic spatial updating, response times should be fairly low and not depend on the turning angle, as all processing should have been completed during the excursion path (Farrell and Robertson, 1998; Presson and Montello, 1994; Riecke et al., 2007; Rieser, 1989). Such an online strategy might be based on participants continuously keeping track of the direction to the starting position using some kind of imagined homing vector, similar to the homing vector updating that is proposed for path integration-based triangle completion in many animals including humans (Loomis et al., 1999; Müller and Wehner, 1988). Conversely, if participants use after-the-fact computation of the homing direction, on would expect response times to be (a) overall larger compared to previous studies that reported automatic spatial updating as well as (b) increase for larger turns and thus more difficult computations, especially for turning angles beyond 90° where reference frame conflicts become more pronounced.

The current data showed qualitatively different response time patterns for Turners versus NonTurners. On the one hand, Turners exhibited overall low response times of 1.07s in the practice experiment and 1.31s in the main experiment. These values are comparable to previously reported values of around 1.6s (Farrell and Robertson, 1998) and 1.2s (Riecke et al., 2007) in physical motion conditions where automatic spatial updating was observed. Moreover, Turner response times in the main experiment showed no systematic increase for larger turning angles. Together, this suggests that Turner might have used some kind of online updating strategy to perform the point-to-origin task, or a fairly efficient offline strategy, or some combination of both.

On the other hand, NonTurners showed considerably longer response times (2.34s and 2.36s for the practice and main experiment, respectively) than Turners and prior studies reporting automatic spatial updating (Farrell and Robertson, 1998; Riecke et al., 2007). Moreover, NonTurners' response times significantly increased for larger turning angles, with effect sizes η_p^2 between 28% and 46%. Both findings suggest that NonTurners might be more prone to using effortful offline, after-the-fact computation of the correct homing direction: If all computation had already been performed during the excursion path, there should be no additional computation time required for the largest and most difficult-to-update turning angles, but this is just what we found. Such after-the-fact computation might be based on some kind of mental rotations, which typically leads to a linear increase of response times with turning angle (Shepard and Metzler, 1971). Alternatively, after-the-fact computation might occur by participants using a configural strategy, for example by imagining a top-down view of the path geometry (Riecke et al., 2002; Wiener et al., 2011). The current study was not designed to disambiguate between those or other possibilities, and further studies are needed to investigate this. The data do, however, suggest that Turner and NonTurner do not only use very distinct strategies leading to qualitatively different behavior, but also systematically vary in the amount of time and cognitive resources needed to determine the homing direction. This might also be related to general differences in mental spatial abilities between Turners and NonTurners (Riecke, 2008).

As cognitive resources are scarce, and robust and effortless spatial orientation and behavior requires low effort and cognitive load, we posit that VR simulations should strive to reduce the occurrence of NonTurner strategies and other effortful and resource-intensive strategies. Thus, using relatively simple experimental paradigms such at the rapid point-to-origin use here, we can systematically investigate the perceptual and behavioral effectiveness of different stimulus and display parameters and combinations. A recent point-to-origin study in VR showed, for example, that using naturalistic stimuli can largely reduce the occurrence of NonTurner behavior, although it still occurred in 17% of participants (Sigurdarson et al., 2012). Thus combining spatial cognition research with an eye towards potential applications can not only help to systematically improve VR simulations and thus provide more effective experimental setups, but also foster a deeper understanding of the fascinating underlying processes and strategies.

References

Avraamides, M.N., Kelly, J.W.: Multiple systems of spatial memory and action. Cognitive Processing 9, 93–106 (2008)

Avraamides, M.N., Klatzky, R.L., Loomis, J.M., Golledge, R.G.: Use of cognitive versus perceptual heading during imagined locomotion depends on the response mode. Psychological Science 15(6), 403–408 (2004)

Batschelet, E.: Circular statistics in biology. Acad. Pr., London (1981)

Chance, S.S., Gaunet, F., Beall, A.C., Loomis, J.M.: Locomotion mode affects the updating of objects encountered during travel: The contribution of vestibular and proprioceptive inputs to path integration. Presence - Teleoperators and Virtual Environments 7(2), 168–178 (1998)

Coluccia, E., Louse, G.: Gender differences in spatial orientation: A review. Journal of Environmental Psychology 24(3), 329–340 (2004)

Easton, R.D., Sholl, M.J.: Object-array structure, frames of reference, and retrieval of spatial knowledge. Journal of Experimental Psychology– Learning, Memory and Cognition 21(2), 483–500 (1995)

Etienne, A.S., Jeffery, K.J.: Path integration in mammals. Hippocampus 14(2), 180–192 (2004)

Farrell, M.J., Robertson, I.H.: Mental rotation and the automatic updating of body-centered spatial relationships. Journal of Experimental Psychology– Learning Memory and Cognition 24(1), 227–233 (1998)

Fujita, N., Loomis, J.M., Klatzky, R.L., Golledge, R.G.: A minimal representation for dead-reckoning in navigation: Updating the homing vector. Geographical Analysis 22(4), 326–335 (1990)

Gallistel, C.R.: The organization of learning. Learning, development, and conceptual change. MIT Press, Cambridge (1990)

Golledge, R.G.: Wayfinding behavior: cognitive mapping and other spatial processes. JHU Press (1999)

Gramann, K.: Embodiment of Spatial Reference Frames and Individual Differences in Reference Frame Proclivity. Spatial Cognition and Computation (2012) (online pre-print), doi:10.1080/13875868.2011.589038

Gramann, K., Muller, H.J., Eick, E.M., Schonebeck, B.: Evidence of separable spatial representations in a virtual navigation task. Journal of Experimental Psychology–Human Perception and Performance 31(6), 1199–1223 (2005)

Gramann, K., Onton, J., Riccobon, D., Mueller, H.J., Bardins, S., Makeig, S.: Human brain dynamics accompanying use of egocentric and allocentric reference frames during navigation. Journal of Cognitive Neuroscience 22(12), 2836–2849 (2010)

Gramann, K., Wing, S., Jung, T.-P., Viirre, E., Riecke, B.E.: Switching spatial reference frames for yaw and pitch navigation. Spatial Cognition and Computation 12(2-3), 159–194 (2012), doi:10.1080/13875868.2011.645176

Klatzky, R.L., Loomis, J.M., Beall, A.C., Chance, S.S., Golledge, R.G.: Spatial updating of Self-Position and orientation during real, imagined, and virtual locomotion. Psychological Science 9(4), 293–298 (1998)

Lawton, C.A., Morrin, K.A.: Gender differences in pointing accuracy in Computer-Simulated 3D mazes. Sex Roles 40(1-2), 73–92 (1999)

Loomis, J.M., Klatzky, R.L., Golledge, R.G., Philbeck, J.W.: Human navigation by path integration. In: Golledge, R.G. (ed.) Wayfinding Behavior: Cognitive Mapping and other Spatial Processes, pp. 125–151. Johns Hopkins, Baltimore (1999)

Mahmood, O., Adamo, D., Briceno, E., Moffat, S.D.: Age differences in visual path integration. Behavioural Brain Research 205(1), 88–95 (2009)

Maurer, R., Séguinot, V.: What is modelling for?– a critical review of the models of path integration. Journal of Theoretical Biology 175(4), 457–475 (1995)

May, M.: Cognitive and embodied modes of spatial imagery. Psychologische Beiträge 38(3/4), 418–434 (1996)

May, M., Klatzky, R.L.: Path integration while ignoring irrelevant movement. Journal of Experimental Psychology– Learning, Memory and Cognition 26(1), 169–186 (2000)

Müller, M., Wehner, R.: Path integration in desert ants cataglyphis fortis. Proceedings of the National Academy of Sciences 85(14), 5287–5290 (1988)

Müller, M., Wehner, R.: Path integration provides a scaffold for landmark learning in desert ants. Current Biology 20(15), 1368–1371 (2010)

Presson, C.C., Montello, D.R.: Updating after rotational and translational body movements: Coordinate structure of perspective space. Perception 23(12), 1447–1455 (1994)

Riecke, B.E., Schulte-Pelkum, J., Bülthoff, H.H.: Perceiving simulated Ego-Motions in virtual reality - comparing large screen displays with HMDs. In: Proceedings of the SPIE, San Jose, CA, USA, vol. 5666, pp. 344–355 (2005a)

Riecke, B.E.: How far can we get with just visual information? Path integration and spatial updating studies in virtual reality, Logos, Berlin, vol. 8 (2003), http://www.logos-verlag.de/cgi-bin/buch/isbn/0440

Riecke, B.E.: Consistent Left-Right reversals for visual path integration in virtual reality: More than a failure to update one's heading? Presence: Teleoperators and Virtual Environments 17(2), 143–175 (2008)

Riecke, B.E., Cunningham, D.W., Bülthoff, H.H.: Spatial updating in virtual reality: the sufficiency of visual information. Psychological Research 71(3), 298–313 (2007)

Riecke, B.E., Heyde, M.V.D., Bülthoff, H.H.: Visual cues can be sufficient for triggering automatic, reflexlike spatial updating. ACM Transactions on Applied Perception (TAP) 2, 183–215 (2005b); ACM ID: 1077401

Riecke, B.E., van Veen, H.A.H.C., Bülthoff, H.H.: Visual homing is possible without landmarks: a path integration study in virtual reality. Presence: Teleoperators and Virtual Environments 11, 443–473 (2002); ACM ID: 772746

Rieser, J.J.: Access to knowledge of spatial structure at novel points of observation. Journal of Experimental Psychology: Learning, Memory, and Cognition 15(6), 1157–1165 (1989)

Shepard, R.N., Metzler, J.: Mental rotation of 3-Dimensional objects. Science 171(3972), 701–703 (1971)

Sigurdarson, S., Milne, A.P., Feuereissen, D., Riecke, B.E.: Can phys-i-cal motions pre-vent disori-en-ta-tion in nat-u-ral-is-tic VR? Orange County, CA, USA (2012)

Stumpf, H., Fay, E.: Schlauchfiguren - Ein Test zur Beurteilung des räumlichen Vorstellungsvermögens. Hogrefe, Göttingen (1983)

Wiener, J.M., Berthoz, A., Wolbers, T.: Dissociable cognitive mechanisms underlying human path integration. Experimental Brain Research 208(1), 61–71 (2011)

Wiener, J.M., Mallot, H.A.: Path complexity does not impair visual path integration. Spatial Cognition and Computation 6(4), 333–346 (2006)

The Effects of Visual Granularity on Indoor Spatial Learning Assisted by Mobile 3D Information Displays

Nicholas A. Giudice and Hengshan Li

Spatial Informatics Program: School of Computing and Information Science
University of Maine, Orono, Maine 04469, USA
giudice@spatial.maine.edu, hengshan.li@umit.maine.edu

Abstract. There is growing interest in improving indoor navigation using 3D spatial visualizations rendered on mobile devices. However, the level of information conveyed by these visualization interfaces in order to best support indoor spatial learning has been poorly studied. This experiment investigates how learning of multi-level virtual buildings assisted by mobile 3D displays rendered at different levels of visual granularity effect subsequent unaided navigation tasks. The visual granularity levels include: a high fidelity model, low fidelity model, wireframe model and sparse model. Results showed that using the sparse model during learning led to the most accurate and efficient overall pointing and navigation performance and that between-floor judgments were less accurate when assistance during learning was unavailable. These findings demonstrate that more information is not necessarily better and provide new insights into the optimal information content to be included in mobile 3D visualization interfaces supporting indoor spatial learning and cognitive map development.

Keywords: indoor navigation, 3D visualizations, mobile information displays, naïve realism, visual granularity, immersive virtual environments.

1 Introduction

Current advancements in the computational resources, memory capacity, and high-resolution display technologies available on mobile devices means that complex environmental visualizations are becoming a viable solution for real-time navigation systems. However, most existing navigation interfaces are limited to 2D representations and work exclusively outdoors. By contrast, our interest here is in designing indoor navigation systems based on 3D building visualizations. Considering that on average, people spend 87% of their time in indoor spaces [1] and since indoor built environments often are comprised of complex and confusing 3D spatial structures [2], providing access to a 3D visualization of the space (i.e., a ground-level egocentric map representation) is postulated as being advantageous and more realistic for supporting spatial learning and cognitive map development as compared to their traditional 2D analogs. Indeed, the efficacy of 3D visualizations and map representations for aiding navigation through indoor environments is a topic of growing interest in

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 163–172, 2012.

both academic research [3-4] and for commercial applications, e.g. Google Maps and Nokia 3D indoor maps.

One practical question for these 3D visualization based navigation systems is how the realism of the 3D models affects human navigation performance? In outdoor environments, several authors from the geo-visualization and cartography communities have advocated the use of abstract rather than photorealistic 3D visualizations for more efficient inference making [5-6]. Empirical experiments addressing this issue support the view that users often have misplaced faith in realistic representations, termed "Naïve Realism" [7]. For example, people using spatial interfaces for naval applications prefer spatially realistic 3D icons of ships and planes on their displays vs. functional, symbolic icons. However, these realistic features were shown to actually decrease identification performance [7]. Similarly, users predicted they would need high-fidelity photorealistic 3D displays to find routes across outdoor terrain, whereas experimental results demonstrated that they actually performed the task better with lower fidelity displays [8]. Several studies have clearly shown that while photorealistic representations of maps appeal to users, they often have a negative impact on behavioral performance [9-10]. As was illustrated in Klippel et al. (2010), people trying to use Google street view for wayfinding purposes converged on a similar experience--that simply providing photorealism is not enough for accurate spatial learning and wayfinding [11].

However, few studies have been conducted to evaluate the effect of environmental realism of mobile interfaces supporting real time indoor navigation. In part, this is due to the lack of accurate indoor positioning for indoor environments and a dearth of real time indoor data models for use on mobile devices. Although relatively impoverished renderings are assumed to be as effective in aiding people's navigation through indoor spaces as photorealistic models, this assumption has not been extensively studied, although initial evidence has provided some empirical verification. For example, Kalia et al. (2008) found that richly rendered (photorealistic) indoor virtual models were not as efficient for spatial learning as a sparse model [12]. However, this study did not investigate different levels of visual granularity of 3D models, nor was it aimed at evaluating the efficacy of using a mobile navigation device to learn multi-level buildings, as is the goal here.

In this study, we experimentally evaluate four simulation fidelity conditions which manipulate the level of visual granularity of the environment which is provided to the user by a simulated mobile device during learning of virtual buildings. We aim to assess whether users' navigation performance after spatial learning with the mobile device differs as a function of the visual granularity of the interface, findings which will help specify the optimal information content to be used in future 3D displays for real-time indoor navigation systems.

The experiment was conducted using immersive virtual environments (VEs) rather than physical environments (PEs) as VEs best facilitate manipulation of building layout and information content, as well as tracking of movement behavior (see Fig. 1).

2 Methods

2.1 Participants

Twenty participants (10 female and 10 male, mean age=20.9, SD=2.0) were recruited from the University of Maine student body. All participants self-reported as having normal (or corrected to normal) vision. All gave informed consent and received monetary compensation for their time. There were three sessions for each subject, with each session lasting approximately one hour.

2.2 Materials and Apparatus

We used an SX111 HMD (NVIS, Inc), incorporating inertial tracking, a panoramic 111 degree field of view, and a high resolution 1260 x 1080 stereo display, which provides a highly immersive VR experience. Two Nintendo Wii remotes were used in the experiment. One was used by the experimenter to control the sequence of experimental phases, and the other was used by the participant to translate through the VE. Turning in the VE was done through physical body rotation.

Our environments were comprised of five two level buildings which were richly rendered in the VE. 3DS Max was used as the 3D modeling and rendering tool. The Vizard 3D rendering suite, by WorldViz Inc., was used as the VE platform supporting users' real-time navigation and recording their trajectory and test performance. As is illustrated in Fig. 1, two types of models were used in the experiment: virtual reality environment models and 3D visualization models. The former simulated the physical world in the VE and were made to be as photorealistic as possible in order to foster the experience of walking in the physical world. The latter included the 3D visualizations which were shown on the simulated mobile device during environmental learning.

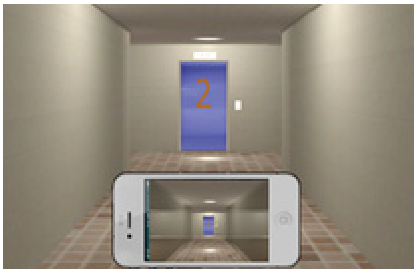
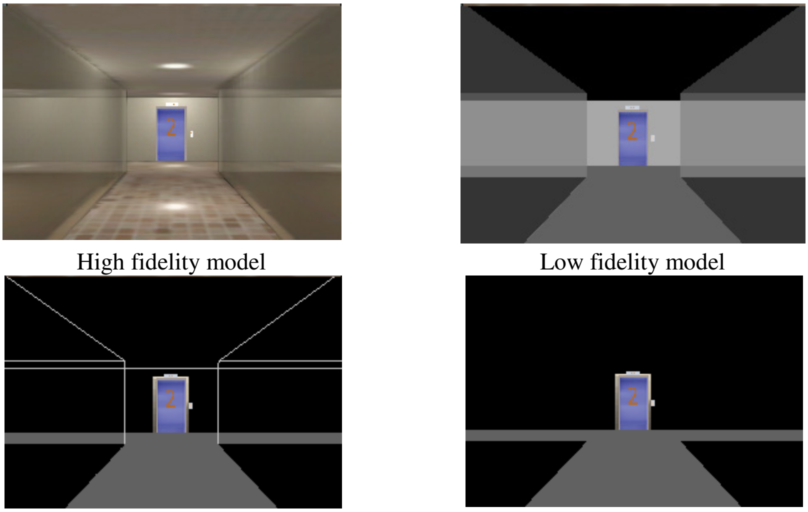


Fig. 1. Simulated mobile device in the VE

Four levels of visualization granularity represent a natural progression of degraded surface detail for environmental rendering, while preserving building topology. Each model is depicted in Fig. 2. The high fidelity model was rendered with photorealistic texture, natural light, and full color (The Mental Ray rendering plug-in was used to generate the model. The low fidelity model used grey scale color to represent the building and there was no rendering of texture or photorealistic light. The wireframe model only rendered the lines at each edge. The sparse model was the simplest representation as it only contained the floor plan of each layout without walls and ceilings.



High fidelity model

Low fidelity model

Wireframe model

Sparse model

Fig. 2. Four visualization fidelity models as shown on the simulated mobile device

Each level of the building was based on a 3 x 3 matrix of hallways, as illustrated in Fig. 3. Each hallway was subdivided into two corridor segments. We deleted two segments from the twelve possible corridor segments in the generic environment to create our experimental layouts. This procedure ensured that all the layouts were well matched in terms of number of nodes, segments, and intersections.

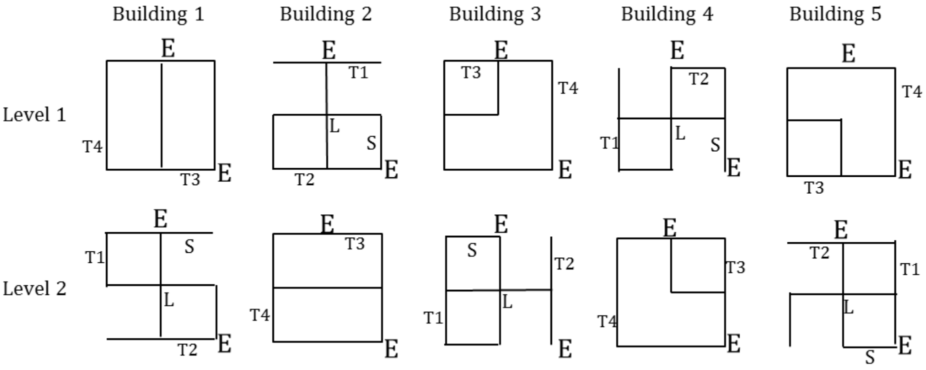


Fig. 3. Experimental building layouts

The two floors were connected by two elevators, which also served as salient landmarks for orientation in each of the experimental buildings ("E" represents the elevators in Fig. 3). From a top-down perspective, one elevator was always located at the top center and the other was located at the southeast corner. In Fig. 3, "L" represents the starting position during the learning period, which was located at the only 4-way intersection in the building. The starting position for the navigation tests, indicated by "S", was located near one of the two elevators to provide an orientation cue but was not visible from the starting learning point. There were two pictures on each floor which served as experimental targets, indicated by "T" in Fig. 3. Pictures were based on eight high imagery words: bottle, chair, clock, dog, fish, kite, table and tie. All routes between pictures were matched across building for route length and number of turns.

2.3 Procedure

A within subjects design was adopted, with twenty subjects running in all five visualization conditions. There were five phases in the experiment.

Phase 1: Practice. Subjects were familiarized with the apparatus and navigation behavior in the VE. All experimental tasks were explained and demonstrated before starting the experimental trials.

Phase 2: Route learning. In this task, participants learned the route to each picture with the assistance of the mobile device. From a north orientation at the learning start point, subjects were guided by arrows displayed on the mobile device to each target picture in each of the four visualization granularity conditions. After reaching the picture, which was hanging on the wall, they were asked to face the picture and remember its location. Subjects were then guided back along the same route to the learning start point and repeated the task for the next target. During the learning phase, the mobile device served as a navigation assistant as it provided increased visual access to the overall floor layout than was possible by simply looking around in the VE. In a fifth unaided control condition, the mobile device was not available during target learning; rather, guidance was done via arrows displayed on the ground. The outbound route for target learning was not necessarily the shortest route. Rather, we chose routes based on a trajectory that maximized environmental exposure. As such, if users looked around as they walked, as was the instruction, they could apprehend the entire building after traversal of the four learning routes. Overlap between routes was minimized to ensure no part of the building was over-learned.

Phase 3: Pointing criterion task. To test whether participants had successfully learned the four target locations from Phase 2 and could situate them in a globally coherent cognitive map of the building, they had to point to each target from the learning start point (target order was randomized by floor). The Phase 2 route learning and Phase 3 pointing task was done separately for each floor (floor order was counterbalanced). When making the pointing response, participants did not have access to the mobile device and no target was visible from the learning start point. Thus, accurate pointing required them to make Euclidean judgments from the learning start point to the target, with half of the targets located on a different floor. To meet criterion, participants needed to point to targets on each floor within a 15 degree tolerance. If they failed the first iteration, the Phase 2 learning and Phase 3 pointing tests proceeded until they either successfully met criterion or until they made four incorrect attempts. We recorded users' pointing time, angular error, and the number of iterations it took to pass the learning criterion test.

Phase 4: Re-exposure task. After the pointing test, participants once again walked from the start point to each of the four pictures (target order was randomized) with the assistance of the mobile 3D visualization interface in order to re-instantiate all targets in memory before starting Phase 5.

Phase 5: Unaided Navigation task. To perform this task, participants were positioned at the navigation start position as shown in Fig. 3. They were then given the name of one of the pictures and asked to navigate to it using the shortest route. This task was performed without assistance from the 3D visualization interface on the mobile device used during learning. The sequences of the pictures were pseudo-random to ensure two routes were within floor and two routes were between floors. Once they believe they had reached the picture, they pressed the button on the Wii mote to indicate its location and orientation. The sequence of pictures was counter balanced between conditions and participants. As subjects only traveled the route between the learning start point and each picture during the learning phase, determining the shortest route between pictures for this navigation task required accurate development and accessing of a “cognitive map” of the entire building. If the participant incorrectly indicated the picture’s location or orientation, they were guided to its correct location and orientation before starting the next trial. This corrective measure was done to prevent the accumulation of error between trials. They were then asked to follow the same sequence of steps for the next target picture. This was done for four routes in total. Two dependent variables for the navigation task were analyzed. The first was navigation accuracy, based on whether subjects successfully indicated the correct location and orientation of the picture. The second was navigation efficiency, based on whether the shortest route was executed (e.g., shortest route length over traveled route length).

3 Results

3.1 Pointing Task

A repeated measures ANOVA on pointing angle error was run with visualization (5 levels: four granularity conditions and the unassisted control) and floor (2 levels: within and between floor target trials) as the within subjects factors. The within-between floor factor was significant, F (1, 39) = 6.495, $p < .015$, $\eta^2 = 0.143$, with the within floor absolute pointing error being 4.3 degrees lower than the between floor pointing error (98.3% of all pointing trials were within the 15 degree tolerance after 2 iterations). There was no significant main effect of pointing error as a function of visualization condition, F (4, 156) = 1.138, $p < .341$, $\eta^2 = 0.028$. However, subsequent pairwise comparisons showed that pointing error for between floor trials was significantly higher than for within floor trials with both the unaided (control) condition ($p<0.015$) and the low fidelity model ($p<0.027$).

Table 1. Mean pointing error (SE in parentheses) for within floor and between floors

	Unaided	High Fidelity	Low Fidelity	Wireframe	Sparse
pointing error within floor	5.3 (.9)	9.3 (2.4)	5.6 (1.1)	7.4 (1.4)	7.4 (1.3)
pointing error between floors	11.3 (2.9)	12.8 (3.3)	12.1 (2.9)	7.9 (1.7)	11.8 (3.6)

A repeated-measures ANOVA on pointing iteration trials was run with the same two within-subjects factors. A significant effect was observed for floor, with more iterations needed to pass criterion for between floor judgments (m = 1.24, SE = 0.05) than for within floor judgments (m = 1.11, SE = 0.023), F (1, 39) = 6.193, p < .017, η^2 = 0.137. There was no significant main effect of iteration as a function of visualization condition, F (4, 156) = 0.624, p < .646, η^2 = 0.016.

Table 2. Mean iteration (SE in parentheses) for within and between floor pointing judgments

	Unaided	High Fidelity	Low Fidelity	Wireframe	Sparse
iteration within floor	1.08 (.04)	1.18 (.06)	1.03 (.03)	1.13 (.05)	1.13 (.05)
iteration between floors	1.28 (.10)	1.28 (.09)	1.35 (.12)	1.13 (.05)	1.18 (.07)

A repeated-measures ANOVA for pointing time was run with the same within-subjects factors. Only floor was significant, F (1, 39) = 10.79, p < .002, η^2 = 0.217, with pointing time taking 3.3 seconds longer for the between floor judgments than for the within floor judgments. There was no significant main effect of pointing time as a function of visualization condition, F (4, 156) = 0.506, p < .732, η^2 = 0.013. However, there was a significant interaction between visualization level and floor, F (4, 156) =2.754, p < .030, η^2 = 0.066. Subsequent pairwise comparisons showed that pointing time for between floor trials was significantly longer than for within floor trials with both the control condition and the low fidelity model, each p <0.005.

Table 3. Mean pointing time (SE in parentheses) for within floor and between floor judgments

	Unaided	High Fidelity	Low Fidelity	Wireframe	Sparse
pointing time within floor	7.59 (.77)	8.93 (1.02)	6.79 (.68)	8.53 (.96)	8.80 (.82)
pointing time between floors	12.21 (1.69)	11.86 (1.97)	13.60 (2.16)	9.87 (1.23)	9.83 (1.26)

3.2 Unaided Navigation Task

A repeated-measures ANOVA for target localization accuracy during the navigation task was run with the same two within-subjects factors of visualization and floor. There was a significant main effect of target localization accuracy as a function of visualization condition, F (4, 156) = 2.678, p < .034, η^2 = 0.064, with localization accuracy after learning with the sparse model (m = 86%, SE=3.6%) being reliably higher than after using the low fidelity model (65%, SE=5.7%), p<0.001. The within-between floor factor was also significant, F (1, 39) = 9.457, p < .004, η^2 = 0.195, α=0.05, with the navigation accuracy found for within floor performance (83%, SE=2.7%) being reliably higher than for between floor judgments (72%, SD=3.6%).

Table 4. Mean navigation accuracy (SE in parentheses) for within floor and between floors

	Unaided	High Fidelity	Low Fidelity	Wireframe	Sparse
navigation accuracy within floor	83% (6.1%)	83% (6.1%)	68% (7.5%)	90% (4.8%)	90% (4.8%)
navigation accuracy between floors	73% (7.1%)	70% (7.3%)	63% (7.8%)	70% (7.3%)	83% (6.1%)

A repeated-measures ANOVA for navigation efficiency was also run for the two within-subjects factors. There was a significant main effect of navigation efficiency as a function of visualization condition, F (4, 156) = 3.192, p < .015, η^2 = 0.076. Navigation efficiency with the sparse model (89%, SE=2.9%) was reliably better than the high fidelity model (73%, SD=4.9%) (p<0.008) and the low fidelity model (71%, SD=5.0%) (p<0.001). The within-between floor factor was not significant, F (1, 39) = 1.641, p < .208, η^2 = 0.040.

Table 5. Mean navigation efficiency (SE in parentheses) for within floor and between floors

	Unaided	High Fidelity	Low Fidelity	Wireframe	Sparse
navigation efficiency within floor	79% (5.4%)	76% (5.4%)	73% (6.9%)	87% (4.5%)	92% (3.1%)
navigation efficiency between floors	85% (5.4%)	71% (6.7%)	69% (6.6%)	73% (6.9%)	87% (4.9%)

4 Discussion

The most important findings of this study are that using the sparse model to assist learning led to the highest unaided target to target localization accuracy and route efficiency performance. These results provide evidence that use of a sparse model of layout structure is better than both of the highest fidelity models for assisting environmental learning of complex buildings. These findings are consistent with, and extend, previous research regarding the evaluation of the realism of 2D maps [7-10]. One explanation is that participants need to extract picture and layout information from high fidelity 3D visualizations to encode the relative positions of these pictures as well as their positions in the building, whereas this information is more directly specified from the sparse model. This synthesis and extraction process may yield additional cognitive effort during learning which resulted in the increased navigation error and decreased efficiency for information-rich displays compared to the displays rendered with lower visual granularity.

We interpret the absence of significant differences for any of the five presentation conditions in the pointing task (pointing time, pointing iteration trials, and pointing error) as further demonstrating that adding realism to the 3D models during learning is neither necessary nor advantageous for extraction of Euclidean relations between targets and accurate cognitive map development.

As for the within-between floor analyses, our results are consistent with previous literature for multilevel indoor navigation [13]. Subjects took longer times to point, required more iterations to meet criterion, exhibited greater errors, and had lower navigation accuracy when pointing and navigating to targets located on different floors than when they were on the same floor. These results suggest that it is more difficult for people to maintain the spatial relation of objects between floors, likely made more difficult when inter-floor layouts are not congruent. Given the known challenges for integration of vertical knowledge in cognitive maps, future experiments will investigate new mobile visualization interfaces for integrating multi-floor buildings during indoor navigation. Importantly, the finding that the control condition showed reliably worse between floor pointing performance than the aided conditions (but for the low-fidelity model), indicates that having assistance during learning (e.g., providing better visual access), may improve knowledge of inter-floor relations. Indeed, we believe that performance in the control condition was likely elevated for all metrics in this experiment as our decision to maximize floor coverage during the route learning phase likely provided sufficient opportunity to apprehend global spatial relations, thereby reducing the inherent benefit afforded by the mobile devices to depict layout configuration. It is likely that performance in the unaided condition would have been significantly worse if we had used a more realistic route learning paradigm that emphasized minimum route length rather than breadth, and was done in buildings with greater topological complexity.

Taken together, these results provide compelling evidence that there is no reliable advantage of 3D information displays rendered at a high level of visual granularity on learning and navigation of buildings and that in many cases, the best performance is obtained using a sparsely rendered spatial model. To our knowledge, our results are the first empirical demonstration showing the advantage of using sparse models on portable mobile devices as supporting real-time learning and navigation of complex indoor buildings. As illustrated by Smallman et al. (2005), good display design is more than slavishly adhering to realism [7]. Our research extends the theory of naïve geography to use of 3D real time indoor maps and provides new evidence for the basic principle of these displays that graphics should not provide more information than is needed by the user [7]. Our results also provide an empirical foundation to help guide the development of more efficient visualization interfaces to be implemented on future indoor navigation systems.

Acknowledgments. This research was supported by NSF grant number: IIS-0916219. We thank Bill Whalen and Chris Bennett for their comments on the manuscript and data analysis.

References

1. Klepeis, N., Nelson, W., Ott, W., et al.: The national human activity pattern survey (nhaps): a resource for assessing exposure to environmental pollutants. Journal of Exposure Analysis and Environmental Epidemiology 11(3), 231–252 (2001)
2. Giudice, N.A., Walton, L.A., Worboys, M.: The informatics of indoor and outdoor space: a research agenda. In: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness, pp. 47–53 (2010)
3. Mulloni, A., Nadalutti, D., Chittaro, L.: Interactive walkthrough of large 3d models of buildings on mobile devices. In: Web3D 2007: Proceedings of the Twelfth International Conference on 3D Web Technology, pp. 17–25 (2007)
4. Chittaro, L., Nadalutti, D.: Presenting evacuation instructions on mobile devices by means of location-aware 3D virtual environments. In: Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services (2008)
5. Haeberling, C.: Cartographic design principles for 3D visualization–A contribution to cartographic theory. In: Proceedings of the 22nd ICA International Cartographic Conference, A Coruña, Spain (2005)
6. Döllner, J.: Non-Photorealistic 3D Geovisualization. In: Multimedia Cartography, Berlin, Germany, pp. 229–240 (2007)
7. Smallman, H.S., John, M.S.: Naive Realism: Misplaced Faith in Realistic Displays. Ergonomics in Design: The Quarterly of Human Factors Applications 13, 6–13 (2005)
8. Smallman, H.S., Cook, M.B., Manes, D.I., Cowen, M.B.: Naïve realism in terrain appreciation. In: Proceedings of the 51st Annual Meeting of the Human Factors and Ergonomics Society, pp. 1317–1321 (2007)
9. Hegarty, M., Smallman, H.S., Stull, A.T.: Decoupling of intuitions and performance in the use of complex visual displays. In: Proceedings of the 30th Annual Conference of the Cognitive Science Society, pp. 881–886. Cognitive Science Society, Washington, DC (2008)
10. Hegarty, M., Smallman, H.S., Stull, A.T., Canham, M.S.: Naïve Cartography: How Intuitions about Display Configuration Can Hurt Performance. Cartographica: The International Journal for Geographic Information and Geovisualization 44(3), 171–186 (2009)
11. Klippel, A., Freksa, C., Winter, S.: You-are-here maps in emergencies – the danger of getting lost. Journal of Spatial Science 51, 117–131 (2006)
12. Kalia, A.A., Legge, G.E., Giudice, N.A.: Learning building layouts with non-geometric visual information: The effects of visual impairment and age. Perception 37(11), 1677–1699 (2008)
13. Holscher, C., Meilinger, T., Vrachliotis, G., Brosamle, M., Knauff, M.: Up the down staircase: Wayfinding strategies in multi-level buildings. Journal of Environmental Psychology 26, 284–299 (2006)

The Relationship between Coordination Skill and Mental Rotation Ability

Stefanie Pietsch and Petra Jansen

Institute of Sport Science, University of Regensburg

Abstract. Motor and mental rotation processes seem to share the same neural mechanism. Within this study we investigated whether there is a relationship not only between motor rotational ability and mental rotation, but also between coordination skill and mental rotation ability. All participants (42 males and 42 females) performed a standardized coordination test, a mental rotation test, and a speed of cognitive processing test. A multiple regression analysis revealed that both gender and coordination skill is a significant predictor for mental rotation performance. The investigation of motor training on mental rotation performance and vice versa in one experimental design is discussed.

Keywords: motor processes, mental rotation, sex differences.

1 Introduction

Mental rotation, the ability to imagine how an object appears when it is rotated from its first presentation, is one of the most investigated spatial processes in literature since the original work of Shepard and Metzler [1] more than 40 years ago. It has been intensively investigated in general psychology [2], neuropsychology [3], differential psychology [4], and developmental psychology [5], [6].

Since the studies of Wexler, Kosslyn, and Berthoz [7] and that of Wohlschlaeger and Wohlschlaeger [8] it is assumed that mental and motor rotations share the same neural processes which control the imagined as well as the physical rotation ("common-processing hypotheses"). Wiedenbauer, Schmid, and Jansen-Osmann [9] showed that manual rotation training with a joystick improves mental rotation performance in adults. In each of these studies motor rotation processes were investigated by the use of hand rotations, which is mostly an eye-hand coordination task.

This manual rotation benefit could also be shown for a more non-specific form of training, which incorporates both eye-hand coordination and inter-limb coordination: juggling. A group of adults who participated in juggling training once a week for three months demonstrated enhanced mental rotation performance compared to a control group who did not receive any training [10]. Moreau, Clerc, Mansy-Dannay, and Guerrien [11] showed improved mental rotation performance after specific sports training. After 10 months of wrestling or running training, only the wrestling group

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 173–181, 2012.

improved significantly on the mental rotation task. Furthermore, there are several quasi-experimental designs showing, for example, that sports and music students show a better mental rotation performance than students of education science [12]. Rotation experts have also demonstrated better mental rotation performance on a perspective transformation task compared to an object transformation task [13]. Moreau, Mansy-Dannay, Clerc, and Guerrién [14] found a clear effect of superior mental rotation performance in martial arts athletes, who constantly have to connect spatial and kinesthetic processes during their exercises, compared to runners, who focus on a cardiovascular fitness. Additionally, the study of Özel, Larue, and Molarino [15] showed that athletes have a better mental rotation performance compared to non-athletes. Both the training studies and the quasi-experimental study of Steggemann, et al. [13] investigated co-ordination skill by either conducting a long term coordinative motor training or by using participants that are rotational experts, such artistic gymnasts or trampoline performers. Due to the nature of quasi-experimental designs, a direct causal relationship could not be stated and results might be influenced by a third factor which caused participants to be good at both sports and mental rotation, such as different patterns of brain activation [16] or body weight. Weight seems to be an important factor since overweight children have demonstrated both impaired motor performance and impaired mental rotation performance compared to children of normal weight [17].

To our knowledge, there is currently only one correlational study showing a correlation between mental rotation and motor [18]. Eighty preschool children performed a paper-pencil mental rotation test, a non-verbal reasoning test, and a motor test. The motor test measured co-ordination ability, fine motor skills, balance, catching ability, jumping ability, speed of movement, and motor control. Mental rotation performance correlated with coordination ability, balance, jumping ability, fine motor skills, motor control, and the performance in the non-verbal reasoning test. A regression analysis showed that the variance was primarily explained by non-verbal intelligence and secondly by one of the four coordination ability tasks, "winding through a hoop", and one of the fine motor skill tasks, "collecting sticks bimanually". In this study it was not explained why only one item out of five items measuring coordination ability explained the mental rotation performance. This result might be caused by the fact that the five items of the coordination ability test (putting balls into buckets, winding through a hoop, jumping jacks, rolling on the ground, springing through a hoop) measure different aspects of coordination ability.

Main Goal of This Study

It is the main goal of this study to investigate the relationship between motor and mental rotation performance in adults. Such a study with adults is missing until now. The mental rotation task we used is that of Peters, Chisholm, & Laeng [20] and entails the following stages: a) perceptual processing and encoding of the objects and its orientation, b) the mental rotation itself, c) the judgment of parity, and d) the motor response. The castle-bomerang test was chosen as a motor test due to its similar stages [19]. This whole-body centered test includes the perceptual process of encoding different objects and sections of the spatial scene and the motor processes of practicing a rotational movement and changing directions. An additional similarity between the tests is that each must be performed under time pressure. Due to the similarities in the processing stages of both tests, a correlation is expected. To exclude the possibility that the results are influenced by a third factor, namely speed of cognitive processing, this variable was measured with the "Zahlen-Verbindungstest" (ZVT) and was included in the correlation analysis [21]. Furthermore, due to the well-known gender differences in mental rotation ability [4], gender was analyzed as a quasi-experimental factor and taken into account in the regression analysis.

2 Method

2.1 Participants

Eighty-four students, 42 males (mean age: 23.12, SD=2.23) and 42 females (mean age: 21.93, SD=1.93), from the University of Regensburg, Germany, participated. All participants gave their consent for publication. Participation was optional and termination of the test was allowed at any time, however no one took this option.

2.2 Material and Procedure

All participants had to first fill out a questionnaire which measured demographic data and sports participation (number of years and times per week spent practicing sports). Next, all students had to conduct a measurement of cognitive speed, the ZVT [21], which was equivalent to the trail making test by Reitan [22]. Participants were given four sheets of paper with the numbers 1-90 on each page. They had to connect the numbers in an ascending order as fast as possible. Maximum time allowed for each page was 30 seconds. The amount of numbers connected was translated into IQ-scores. The ZVT correlates with standard IQ-tests, r=.6 to r=.8 [23]. After completing the ZVT participants had to solve a mental rotation test (MRT)[20]. The test consisted of a short practice set with four tasks, for which answers were provided, and two test sets with 12 tasks each. These tasks were first developed by Vandenberg and Kuse [24] and redrawn by Peters et al. [18]. Each tasks consisted of five cube figures, one cube figure on the left side as a standard item and four cube figures on the right side as comparison items (see Figure 1).

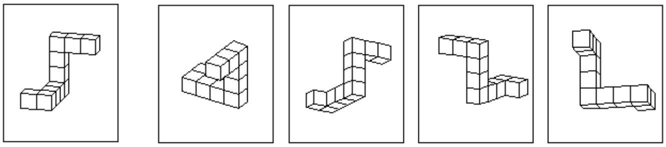


Fig. 1. One task in the Mental Rotation Test of Peters et al. (1995)

Two of the four comparison items were the same as the standard item and could be compared to this item by using mental rotation. Participants completed the practice sets without a time limit and had three minutes to complete each test set (six minutes in total). The standard scoring method followed Peters et al. [20] was followed: One point was given only if both items were marked correctly. The maximum possible score was 24. The split-half reliability was .80 [25].

After completing the two cognitive tests, participants had to complete the castle-boomerang test [19]. The Castle-boomerang test was first developed for testing the co-ordination ability of children and adolescents. Mean objectivity reads. 96, and mean reliability .86. Validity is declared and an experts rating confirms content validity. The castle boomerang test is additionally used for the hiring procedure for the German state police.

The motor test is presented in Figure 2.

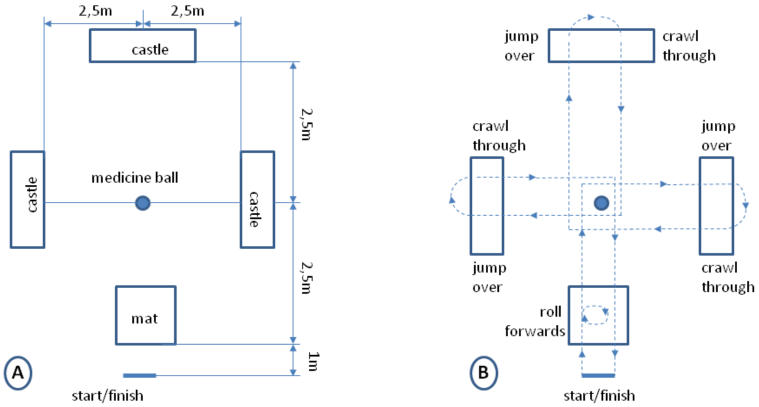


Fig. 2. Castle-Boomerang test: Figure 2A presents the arrangement of the barriers; Figure 2B presents the route the participant has to follow

This motor test measures whole-body centered coordinative skills under time pressure. It is the goal of this test to complete various motor tasks in an ordered sequence under a specified amount of time. Aerobic conditioning element could not be discarded but the impact was minimized.[1] The castle-boomerang test specifically investigates inter- and intra muscular co-ordination by measuring the reaction time to the start signal, orientation ability, differentiation performance, and the ability to coordinate the movement and also spatial orientation ability.

Figure 2A, three castles, a medicine ball, and a mat are arranged in a gym. The castles consist of a wooden frame and are concave, so it is possible to jump over and crawl through them. In Figure 2B the participant's movement through the course is

[1] The short distances as well as the plenty of turns assure that the possible maximal velocity (which is reached at earliest after 15 meters) and the skill to accelerate did not influence the performance. Due to the briefness of the test, aerobic physical fitness is not relevant (lactate values is only deteriorative after 20 sec.), and strength is not necessary because the height of the castle is only about 16 inch.

described: At the sign of the experimenter the participant starts at the Start/Finish line by a somersault on a mat. Afterward the participant runs to the left side of the medicine ball in the center of the room, turns 90°, runs and jumps over the castle 1, and then turns to crawl through this castle. The same procedure is followed for castles 2 and 3. After the last run around the marking in the middle of the room, participants run over the finish line. Time was measured to the deci-second. Co-ordination skill was quantified as the time needed to complete this motor task.

2.3 Statistical Analysis

First, gender differences were analyzed concerning cognitive speed of processing, coordination skill, and mental rotation performance. Second, a correlation analysis was conducted with the variables: ZVT performance, coordination skill, MRT performance, and the hours and years of Sports practice. Additionally, a correlational analysis was performed between the motor and the mental rotation test performance. Finally, a stepwise regression to predict mental rotation performance was performed.

3 Results

The univariate analysis of the factor "gender" only revealed a significant effect for the dependent variable "MRT performance", F(1,82)=22.92, p<.001, η^2=.22. Males (M=14.12, SD=4.44) solved more mental rotation tasks correctly than females (M=9.83, SD=3.78). There were no significant gender differences for the factors "ZVT performance", F(1,82)=1.98, n.s., (Males: M= 2.91, SD=0.52; Females: M=3.06, SD=0.50) or "coordination skill", F(1,82)=2.01, n.s. (Males: M= 18.39sec., SD=2.97sec; Females: M=19.14sec, SD=1.70sec).

Table 1 gives the correlations between the following variables: ZVT performance, coordination skill, MRT performance, and hours and years of sports practice.

Table 1. Correlation coefficients between the variables ZVT performance, coordination skill, MRT performance and the hours and years of sports practice

	Sum MRT	ZVT	Coordination skill	Years of sports	Hours of sports
Sum MRT	1	.04	-.382**	.014	-.088
ZVT		1	-.088	.340*	.303*
Coordination skill			1	-.304	-.103
Years of sports				1	.590**
Hours of sports					1

*signifies p<.05, **signifies p<.01.

Astonishing, neither years nor hours of sports practice correlated with the MRT performance. A correlation analysis performed separately for males and females showed that the correlation between MRT performance and coordination skill for males (r=-.405, p<.01) and females (r=-.27, p=.08) did not differ significantly (*Zdifference*=--.674, *n.s.*)

However, our former study [12] showed that these measurements of sports activity do correlate with a better performance on mental rotation tasks. This caused us to include both the variables "years of sports" and "hours of sports" in the regression analysis, as well as the factor "gender" and "co-ordination skill". A correlation between the performance in the ZVT and mental rotation performance could not be reported and due to this cognitive processing speed was not considered further as a significant influencing factor. Due to the significant correlation between "years of sports" and "hours of sports" a stepwise multiple regression was conducted. Because the mean variance inflation factor (1.525) was not substantially greater than 1, the regression was not biased by collinearity compare [26].

Table 2. Stepwise multiple regression for the mental rotation performance based on the following predictors: Gender, coordination skill, years and hours per week of sports practice

Predictor	Regression coefficient	ß	T	p
Gender	-3.836	-.418	-4.499	<.001
Coordination skill	-.599	-.317	-3.409	=.001
Years of sports		.037	.363	n.s.
Hours of sports		.111	1.034	n.s.

A significant predictor in the first model was "gender", which correlated with the MRT performance (R=.467) and explained 21.8% of the variance, F (2, 83) = 22.91, p<.001. Significant predictors included in the final model were "gender" and "coordination skill". Both variables correlated with the MRT performance, (R=0.563) and 31.6% of the variance is explained by both variables, F (2, 83) = 18.75, p<.001.

4 Discussion

Our results show that males perform better than females in the MRT, a result which is in line with many other studies [4]. We found no gender differences in cognitive processing speed or whole-body centered coordination skill.

This study showed that good motor coordination skill correlates with high mental rotation ability. This is very important because it shows that the relationship between physical and mental processes is not just limited to tasks with a high similarity, such as manual and mental rotation [8]. In this sense our study has extended the work of Wohlschlaeger and Wohlschlaeger [8]. Not only could common processes be theorized for rotational hand movements and mental imagery of objects, but also for whole-body processes and mental rotation processes. Nevertheless, hand movements and whole-body motor tasks share different components of the processes used during a mental rotation task. As such the similarities between the castle-boomerang test and the mental rotation test might be seen in their similar processing stages, namely the encoding of objects and the rotational movements (direction changes and somersault in the motor test and rotational imagery of the objects). The common constraint is the time limit. Due to the similar stages used when performing a mental rotation task and this coordination test, it seems plausible that both tasks correlate with each other. There are also stages, which are not comparable between both tasks, such as for example the conditioning elements, which are minimized but are present in the motor task, and the maintenance of the objects and the judgment of parity in the mental rotation task. The maintenance of the objects requires the involvement of working memory processes and the minimized conditional elements require energy processes. Whether there is a relationship between these stages remains speculative, however this assumption relates to a study of Sibley and Beilock [27] who showed that healthy adults with a low cognitive performance rate had a better performance on working memory tasks when they participated in an aerobic conditioning motor task.

To compare gross motor coordination ability in a more direct way with mental rotation ability, another type of mental rotation and motor test could be used. For mental rotation, the use of a chronometric test might be useful, so that the rotation speed itself as well as the encoding of the objects (intercept of the rotation speed function) could be differentiated. The castle-boomerang test might be changed to include only somersaults and other body-rotation gymnastic exercises without directional change elements within the room so that spatial navigation ability could be limited.

Even though the motor test could be redesigned for experimental reasons, it does have practical value. This test is often used in Germany as a qualifying examination for physically demanding jobs, such as a position as a police officer. If the result of this test relates to mental rotation performance this is an important result because the ability to imagine objects from different perspectives might be crucial even in law enforcement.

It is the nature of correlational analysis that the cause and effect of the results is not evident. Some studies show the influence of sports training on mental rotation performance [10], which leads to the causal interpretation in the direction of “motor training to mental training”. However, on the other side there are studies showing that mental training can improve motor performance [24-25]. Because of the disadvantages of a correlational analysis, a directional interpretation is not possible. We can’t exclude that participants with good spatial navigation ability have advantages in the castle-boomerang test. Furthermore, the jumping ability and body-height of the participants might influence the results in the castle-boomerang test as well as the proportion of fast twitch muscle fibers. These factors must be controlled in further studies.

The regression analysis showed that the mental rotation performance could be explained by both "co-ordination skill" and "gender", meaning that both factors predict the mental rotation performance. The null-finding of gender differences in mental rotation performance in the study of Jansen and Heil [18] is in line with studies in which a crucial time slot for gender difference to appear was seen at about 10 years of age [29]. However, some studies support the assumption that gender differences may appear in early childhood or even infancy [30]. On a very speculative basis, one might assume that the co-ordination ability and mental rotation ability is only related for the advanced mental rotation performer, in this case males. This is an assumption which deserves further attention and should be investigated in more detail and with more participants in further studies.

This study, as other studies before, shows that one specific spatial cognitive task - the mental rotation task - and gross motor ability, coordination skill, are related and might share common processes or common behavioral stages. There are studies indicating that motor processes influence mental (rotation) processes and that mental training influences motor processes. What is missing until now, and needs to be done in the near future, is the investigation of both theoretical directions in one experimental design.

References

1. Shepard, R.N., Metzler, J.: Mental rotation of three-dimensional objects. Science 171, 701–703 (1971)
2. Yuille, J.C., Steiger, J.H.: Nonholistic processing in mental rotation: Some suggestive evidence. Perception and Psychophysics 31, 201–209 (1982)
3. Jordan, K., Heinze, J., Lutz, K., Kanowski, M., Jäncke, L.: Cortical activations during the mental rotation of different visual objects. Neuroimage 13, 143–152 (2001)
4. Voyer, D.: Time limits and gender differences on paper-and-pencil tests of mental rotation: a meta-analysis. Psychonomic Bulletin & Review 18, 267–277 (2011)
5. Marmor, G.S.: Development of kinetic images: When does the child first represent movement in mental images? Cognitive Psychology 7, 548–559 (1975)
6. Newcombe, N., Frick, A.: Early education for spatial intelligence: Why, what and how. Mind, Brain, and Education 4, 102–111 (2010)
7. Wexler, M., Kosslyn, S., Berthoz, A.: Motor processes in mental rotation. Cognition 68, 77–94 (1998)
8. Wohlschlaeger, A., Wohlschlaeger, A.: Mental and manual rotation. Journal of Experimental Psychology: Human Perception and Performance 24, 397–412 (1998)
9. Wiedenbauer, G., Schmid, J., Jansen-Osmann, P.: Manual training of mental rotation. European Journal of Cognitive Psychology 19, 17–36 (2007)
10. Jansen, P., Titze, C., Heil, M.: The influence of juggling on mental rotation performance. International Journal of Sport Psychology 40, 351–359 (2009)
11. Moreau, D., Clerc, J., Mansy-Dannay, A., Guerrien, A.: Spatial abilities and motor performance: Assessing mental rotation processes in elite and novice athletes. International Journal of Sport Psychology 42, 525–547 (2011)

12. Pietsch, S., Jansen, P.: Different mental rotation performance in students of music, sports and education science. Learning and Individual Differences 22, 159–163 (2012)
13. Steggemann, Y., Engbert, K., Weigelt, M.: Selective effects of motor expertise in mental body rotation tasks: Comparing object-based and perspective transformation. Brain and Cognition 76, 97–105 (2011)
14. Moreau, D., Clerc, J., Mansy-Dannay, A., Guerrien, A.: Enhancing spatial abilities through sport practice. Journal of Individual Differences (2012)
15. Ozel, S., Larue, J., Molarino, C.: Relation between sport and spatial imagery: Comparison of three groups of participants. The Journal of Psychology 138, 49–63 (2004)
16. Sekiguchi, A., Yokoyama, S., Kasahara, S., Yomogidym, Y., Takeuchi, H., Ogawa, T., Taki, Y., Niwa, S.I., Kawashima, R.: Neural bases of a specific strategy for visuospatial processing in rugby players. Medicine & Science in Sports & Exercise 43, 1857–1862 (2011)
17. Jansen, P., Schmelter, A., Kasten, L., Heil, M.: Impaired mental rotation performance in overweight children. Appetite 56, 766–769 (2011)
18. Jansen, P., Heil, M.: The relation between motor development and mental rotation ability in 5-6 years old children. European Journal of Developmental Science 4, 66–74 (2010)
19. Bös, K.: Handbuch motorischer Tests. Hogrefe, Göttingen (2001)
20. Peters, M., Chisholm, P., Laeng, B.: Spatial ability, student gender and academic performance. Journal of Engineering Education 84, 60–73 (1995)
21. Oswald, W.D., Roth, E.: Der Zahlen-Verbindungs-Test ZVT. Hogrefe, Göttingen (1987)
22. Reitan, R.M.: Trail making test. Manual for administration, scoring, and interpretation. Indiana University Press, Indianapolis (1956)
23. Vernon, P.A.: Der Zahlen-Verbindungstest and other trail-making correlate of general intelligence. Personality and Individual Differences 14, 35–40 (1993)
24. Vandenberg, S.G., Kuse, A.R.: Mental rotations, a group test of 3-dimensional spatial visualization. Perceptual and Motor Skills 47, 599–604 (1978)
25. Geiser, C., Lehmann, W., Eid, W.: Separating “rotators” from “nonrotators” in the mental rotation test: A multigroup latent class analysis. Multivariate Behavioral Research 41, 261–293 (2006)
26. Bowerman, B.L., O’Conell, R.: Linear statistical models: An applied approach, 2nd edn. Duxbury, Belmont (1990)
27. Sibley, B.A., Bialock, S.L.: Exercise and working memory: An individual differences investigation. Journal of Sport & Exercise Psychology 29, 783–791 (2007)
28. Ranaganathan, V.K., Siemioniv, V., Liu, J.Z., Sahgal, V., Yue, G.H.: From mental power to muscle power – gaining strength by using the mind. Neuropsychologia 42, 944–956 (2004)
29. Gentili, R., Han, C., Schweighofer, N., Papaxanthis, C.: Motor learning without doing: trial-by-trial improvement in motor performance during mental training. Journal of Neurophysiology 104, 774–783 (2010)
30. Johnson, E.S., Meade, A.C.: Developmental patterns of spatial ability: An early sex difference. Child Development 58, 725–740 (1987)
31. Quinn, P.C., Liben, L.S.: A sex difference in mental rotation in young infants. Psychological Science 19, 1067–1070 (2008)

Collaborating in Spatial Tasks: Partners Adapt the Perspective of Their Descriptions, Coordination Strategies, and Memory Representations

Alexia Galati and Marios N. Avraamides

Department of Psychology, University of Cyprus
P.O. Box 20537, 1678 Nicosia, Cyprus
{galati,mariosav}@ucy.ac.cy

Abstract. The partner's viewpoint influences spatial descriptions and, when strongly emphasized, spatial memories as well. We examined whether partner-specific information affects the representations people spontaneously construct, the description strategies they spontaneously select, and the representations their collaborating partner constructs based on these descriptions. Directors described to a misaligned Matcher arrays learned while either knowing the Matcher's viewpoint or not. Knowing the Matcher's viewpoint led to distinctive processing in spatial judgments and a rotational bias in array drawings. Directors' descriptions reflected strategic choices, suggesting that partners considered each other's computational demands. Such strategies were effective as reflected by the number of conversational turns partners took to coordinate. Matchers represented both partners' viewpoints in memory, with the Directors' descriptions predicting the facilitated perspective. Thus, partners behave contingently in spatial tasks to optimize their coordination: the availability of the partner's viewpoint influences one's memory and description strategies, which in turn influence the partner's memory.

Keywords: perspective-taking, coordination, spatial memory, dialogue.

1 Introduction

People routinely share spatial information to coordinate in a variety of tasks, from giving directions to a visitor in an unfamiliar environment to jointly moving a piece of furniture across rooms. The selection of a perspective when producing or interpreting spatial descriptions has been systematically investigated, with findings identifying some of the cognitive, contextual, and communicative constraints influencing this selection process. However, this work usually focuses either on people's linguistic choices without directly examining the representations that support perspective-taking (e.g., Schober, 1993, 1995, 1999) or focuses on processing in noninteractive tasks (e.g., Carlson-Radvansky & Irwin, 1994; Carlson-Radvansky & Logan, 1997; Mou et al., 2004a) or in tasks where the interaction between (presumably) collaborating partners is constrained (e.g., Duran et al., 2011; Shelton & McNamara, 2004). Thus, it's not yet well understood how the perspectives that people spontaneously select, both for organizing spatial information in memory and for their descriptions, are

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 182–195, 2012.

influenced by partner-specific factors. It's also unclear how one partner's spatial representations and description strategies influence not only their coordination with another partner in the task, but also the memory representations the other partner constructs through that coordination. In this paper, we present some of our work that addresses these questions.

We begin by reviewing, in Section 2, research that identifies some of the factors that affect the perspective of speakers' descriptions and the efficiency of partners' coordination in spatial tasks. In Section 3, we review our recent work showing that knowing in advance the partner's misaligned viewpoint influences speakers' memory representations and their subsequent descriptions. In the remaining sections, we go beyond our earlier examination of how the availability of the partner's viewpoint influences speakers' memory representations and descriptions by using the same corpus to also investigate how these partner-specific factors influence the coordination between partners (Section 4) and the partners' memory representations that result from that coordination (Section 5). In Section 6, we summarize our findings, concluding that (1) people consider the task's cognitive demands on their partner to select the perspective of their descriptions and strategies that would maximize the efficiency of communication, (2) these strategies are indeed effective in facilitating coordination, and (3) the perspective of speakers' descriptions shapes the memory representations of their partners.

2 Coordinating in Collaborative Spatial Tasks

A confluence of findings suggests that people tailor their spatial descriptions in response to their conversational partner. For instance, in tasks in which pairs jointly reconstructed arrays, the degree of misalignment between partners affected the perspective of speakers' descriptions. Speakers were more likely to use partner-centered descriptions (e.g., "to your left" or "in front of you") than egocentric ones when describing arrays to partners who didn't share their viewpoint compared to partners who did (Schober, 1993; 1995). Moreover, constraints of the communicative situation can lead to attributions about the partner that also affect the perspective of speakers' descriptions: speakers describing arrays to an imaginary partner were more likely to use partner-centered descriptions and less likely to use egocentric ones compared to those describing arrays to a real partner (Schober, 1993).

Similarly, attributions about the partner's spatial ability, arising as the interaction unfolds, can also affect the perspective of speakers' descriptions (Schober, 2009). When partners were preselected to have matched or mismatched spatial abilities, high-ability speakers were more likely to use partner-centered descriptions whereas low-ability speakers were more likely to use egocentric ones. Additionally, as high-ability speakers formed attributions about their low-ability partners during the course of the interaction, they increased their partner-centered descriptions, whereas low-ability speakers decreased their use of partner-centered descriptions when describing to high-ability partners. The pairs' efficiency and accuracy also depended on their respective abilities. Pairs with two high-ability partners used fewer words than mixed pairs or pairs with two low-ability partners, and even though low-ability partners were generally less accurate in the task, their performance was better when paired with a high ability partner.

The accuracy and efficiency of coordination in spatial tasks doesn't only depend on the partners' cognitive constraints, such as their (combined) spatial abilities, but also on the affordances of the communicative situation, such as the visibility between partners or the shape of their shared space. In a task that involved reconstructing arrangements of lego blocks, pairs who could see each other were more accurate and efficient, since addressees could exhibit, poise, point at and orient blocks, and exchange feedback contingently, while speakers could also adapt their descriptions contingently in response (Clark & Krych, 2004). Even in a narrative task, the affordances of the communicative situation can shape how speakers encode spatial information: the relative locations of speakers and addressees influence the shape of their shared space and, as a consequence, the directionality of their gestures that accompany spatial prepositions like *in* and *out* (Özyürek, 2002).

People also adapt how they plan and describe routes according to whether they do it for themselves or for a partner unfamiliar with the environment (Hölscher et al, 2011). For an unfamiliar partner, people use more words and details, navigate along fewer, larger and more prominent streets and refer to more landmarks. Similarly, they adapt the level of detail they incorporate in describing landmarks, depending on whether their partner is familiar or unfamiliar with them (Isaacs & Clark, 1987).

Such adaptation in perspective choices is not limited to production, but extends to the interpretation of spatial descriptions as well. Attributional cues about the partner affected how people interpreted the perspective of ambiguous spatial descriptions (e.g., "give me the folder on the left", when partners occupied different viewpoints) in an online task, as reflected by the temporal and trajectory characteristics of their responses (Duran et al., 2011). Believing that their partner didn't know where they were seated (and could not consider their perspective) led people to more partner-centered responding, whereas believing that their partner was real (vs. simulated) led to more egocentric responding. Similarly, beliefs about whether the partner was an adult vs. a child influenced how participants planned their moves in a "tacit communication game", in which their intentions had to be conveyed exclusively through graphical means: they spent more time signaling the location of critical information to their partner when they believed they were interacting with a child (Newman-Norlund et al., 2009).

Across these studies, findings are broadly consistent with the view that partners share responsibility for mutual understanding and adapt their behavior in trying to minimize the collective effort of themselves and their partner; this has been termed as *the principle of least collaborative effort* (Clark, 1996; Clark & Wilkes-Gibbs, 1986). In situations where speakers address a partner who is imaginary or believed to be a child, or in situations where feedback is constrained, they expend considerable effort to adopt their partner's perspective or to convey spatial information to their partner. On the other hand, in circumstances where they interact with a real (or assumed to be real) partner, or a partner who can contribute contingently to the interaction (e.g., because they can see them), they may not invest as much effort in adopting the partner's perspective and instead rely on the partner to request clarifications, as needed. Thus, the attributions that people make about their partner and their ability to contribute to the interaction are critical in determining their perspective choices in collaborative spatial tasks.

Nonetheless, few studies have examined directly how partner-specific information can affect the memory representations that people recruit or generate in collaborative spatial tasks. In one study by Shelton and McNamara (2004), speakers were more accurate to make judgments about relations between array objects from the perspective that their partner had occupied earlier, when they had described the array to them. However, in this study partners could not coordinate freely during the description: speakers were instructed to describe arrays from the partner's perspective, and addressees did not know where speakers were relative to the array and could not provide any spoken feedback. This led to speakers using mostly partner-centered descriptions regardless of the degree of misalignment from their partner, and perhaps not surprisingly led to using the partner's viewpoint as an organizing direction in memory. Without these task constraints, it's unclear whether people would spontaneously incorporate their partner's viewpoint in memory.

3 The Partner Affects Memory Representations and Description Strategies

We recently adapted Shelton and McNamara's (2004) study to examine factors that affect whether people *spontaneously* represent their partner's viewpoint in spatial memory and to identify the description strategies that they *spontaneously* adopt when communicating spatial information. To do so, we dissociated the learning of the arrays from their description, and we did not constrain the interaction of partners when they reconstructed arrays.

In 18 pairs in our study, one participant, the Director, learned a table-top array of seven objects and later described it from memory to another participant, the Matcher, who reconstructed it by following the Director's descriptions. Across three blocks, Directors learned arrays under different conditions, which varied in terms of what Directors knew about their Matcher's viewpoint (i.e., the salience of the partner's viewpoint). In the first block, Directors didn't know that they would have to describe the array to a Matcher (No Intent condition). In the subsequent two blocks, Directors either knew that they would have to describe the array to a Matcher without knowing the Matcher's viewpoint (Intent condition), or knew that they would have to describe the array to the Matcher and also knew the Matcher's viewpoint, as the Matcher was co-present in the room during learning (Co-presence condition). The order of these two latter conditions was counterbalanced across pairs of participants, as was the degree of misalignment between partners during the description phase, which was 90°, 135°, or 180° across the three blocks.

The Directors' memory of arrays was assessed *prior* to descriptions through two tasks. The first involved *judgments of relative direction* (JRDs), which asked Directors to imagine a specific location and orientation, and to point, using a joystick, to another object from that imagined perspective (e.g., *Imagine being at the vase, facing the orange. Point to the button.*). The 48 JRD trials included eight imagined headings (0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°) and their order was randomized. Performance was assessed in terms of participants' orientation latency (the time from the offset of the instruction to adopt an imagined perspective to pressing the joystick button to indicate that they adopted that perspective) and their response latency (the time from the offset of the instruction to point to the target object to pressing the button to log their response after having deflected the joystick). Performance on JRD trials allowed us to determine the preferred direction participants used to organize the spatial relations in memory (e.g., Kelly et al., 2007). In a second task, Directors reconstructed the array by indicating the position of each object on a grid circle representing their table. This allowed us to assess their memory for relative positioning of objects and for systematic biases (e.g., Friedman & Kohler, 2003). The Matchers' memory of the arrays they reconstructed was assessed through the same tasks after the description phase, allowing us to examine the extent to which their representations were organized similarly to the Directors' and the extent to which these representations depended on the perspective of their Directors' descriptions.

In Galati et al. (2011), we focused on the Directors' performance in the memory tests and on the spatial perspectives they adopted in their linguistic descriptions. We found that, in the absence of advance information about the Matcher's viewpoint (in the No Intent and Intent conditions), Directors encoded arrays egocentrically, being faster to imagine orienting to and to respond from perspectives aligned with their own. On the other hand, when the Matcher's viewpoint was known in advance (in the Co-presence condition) it showed distinctive processing, at least when Matchers were known to be misaligned by 90° or 135°: Directors took longer to imagine orienting to headings aligned with these known viewpoints of their Matchers. We proposed that this was because, when orienting to headings aligned with their Matcher, Directors recalled their experience at learning and linked the Matcher's viewpoint to their representation of the array, incurring a processing cost. The Directors' array drawings also provided converging evidence for having represented their Matcher's viewpoint in memory: in the Co-presence condition, when Directors knew their Matcher's viewpoint in advance, their drawings showed a reliable rotational bias towards the Matcher and, to some extent, affected how distorted the relative positions of array objects were.

Directors also adapted the types of spatial descriptions they used according to the conditions at learning. However, advance knowledge of the Matcher's viewpoint did not determine on its own the perspective of their spatial expressions—that is, Directors didn't necessarily use more partner-centered expressions advancing the Co-presence condition. Instead, Directors' descriptions suggested strategic choices: when perspective-taking was relatively easy for both partners (at the small offset of 90º), they used Matcher-centered expressions more frequently, whereas when coordination was more difficult and perspective-taking was more computationally demanding for them (at the oblique 135°), they opted for their own perspective, often upon explicitly agreeing with their Matchers to do so. In fact, such explicit agreement between partners happened most often in the Co-presence condition, when partners had known in advance they'd be offset by 135°. Thus, knowing in advance each other's viewpoint enabled partners to mutually recognize when the communicative situation would be more demanding for each of them and to adapt their strategies accordingly to facilitate coordination.

Converging evidence regarding the partners' mutual commitment to maximize efficiency in communication came from other global strategies, like the Directors' overall perspective preference (which required over 70% of all person-centered expressions of a given block to be Director-centered or Matcher-centered). When Directors preferred their own perspective overall, they did so more frequently when offset by 135° from their Matcher, whereas when preferred their Matcher's perspective overall, they did so more frequently when offset by 90°. The Directors' initial description choices were also congruent with these strategies, as reflected by the alignment of the first two objects of their descriptions (i.e., whether these objects were aligned with the Director, the Matcher, or neither partner). When Directors knew about the subsequent description at learning, the initial perspective of their descriptions was more likely to be aligned with their own viewpoint when offset by 135° with their Matcher, and more likely to be aligned with their Matcher's viewpoint when offset by 90° with their Matcher.

Together, these findings suggest that partners shared responsibility for mutual understanding and adapted to the communicative situation flexibly; the burden of perspective-taking wasn't exclusively on the Director. Consistent with the principle of least collaborative effort (Clark & Wilkes-Gibbs, 1986), when partners recognized that one of them was likely to find the interaction difficult (e.g., when the Director described an array from a 135° offset), the other invested greater cognitive effort to ensure mutual understanding (e.g., the Matcher opted to interpret descriptions from the Director's perspective) to minimize their collective effort.

In the next two sections, we focus on behavioral adjustments beyond those of Directors in order to assess, first, how the conditions under which Directors learned arrays affected the efficiency of partners' coordination during the description, and then how the Matchers' resulting memory representation were affected by these conditions and by their Directors' description strategies.

4 Partners Consider Each Other's Cognitive Demands When Coordinating in Spatial Tasks

Partners in a joint activity monitor and coordinate their behavior by *grounding,* or exchanging ongoing evidence about what they do or do not understand (e.g., Clark & Brennan, 1991; Clark, 1996; Brennan, 2004). As an index of the partners' collaborative effort, we considered the number of conversational turns pairs took to reconstruct a given array. Uninterrupted stretches of speech by a Director or Matcher were counted as turns. Decreases in the number of turns suggest facilitation in grounding, whether due to a reduced cognitive cost of perspective-taking or due to successful coordination strategies (e.g., Clark & Wilkes-Gibbs, 1986).

In our study, practice on its own did not reliably influence the partners' collaborative effort as reflected by their number of conversational turns: although pairs took overall fewer turns across the three blocks, this was only a numerical trend. As Figure 1 shows, the number of turns patterned differently across the different levels of misalignment between partners according to what the Directors knew about their Matchers at learning. During the first block, the No Intent condition, when Directors had learned arrays without knowing about the upcoming description, partners took numerically (though not reliably) fewer turns to reconstruct arrays at the smallest offset of 90° relative to the other offsets. In the Intent condition, when Directors had learned arrays while knowing about the description but not their Matchers' viewpoint, pairs took fewer turns numerically when they were counter-aligned relative to the other offsets. And in the Co-presence condition, when the Directors had learned arrays while knowing in advance the Matcher's subsequent viewpoint, they took the fewest turns numerically when they knew the Matcher would be offset by 135°.

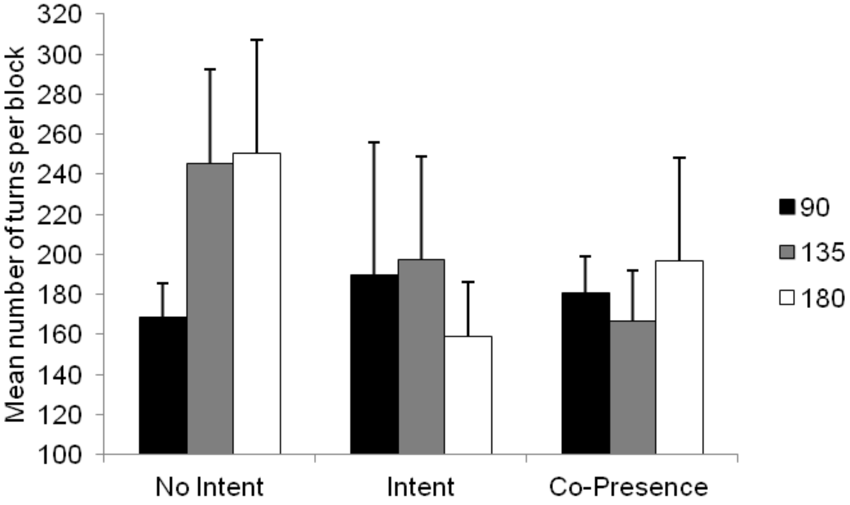


Fig. 1. Mean number of conversational turns across the three different levels of misalignment between partners, for each condition of partner salience

It may seem counterintuitive that partners tended to be more efficient when misaligned by the oblique and presumably computationally demanding offset of 135°, but their description strategies help contextualize this pattern. As we have found in Galati et al. (2011), when Directors knew they would be offset by 135°, they were more likely to use Director-centered expressions in their descriptions, having frequently agreed explicitly to do so with their Matchers, usually on their Matchers' own initiative. Indeed, the Directors' use of more egocentric expressions predicted the effort they expended when collaborating, as reflected by a reliable correlation with the number of turns: the greater the proportion of Director-centered expressions in the Co-Presence condition, the fewer turns partners needed to reconstruct the array.

Previously, we have claimed that when partners knew each other's viewpoint in advance, they were better able to mutually recognize when coordinating would be difficult and to agree on a strategy that would alleviate the demands on the partner with the greatest responsibility for mutual understanding in the task. Here, with turns as a proxy of partners' collaborative effort, we can corroborate that the partners' selected strategy (of reconstructing arrays with descriptions from the Director's perspective) was apt and successful in making their coordination more efficient.

5 Speakers' Descriptions Shape Their Partners' Memory Representations

Given the strategies that partners deployed during spatial descriptions, we wanted to determine whether the Matchers' memory representations were affected accordingly. Descriptions that differ in perspective, in terms of whether they involve a survey, bird's eye perspective or whether they guide the reader or listener along a route, have been shown to lead to comparable performance in spatial tasks, presumably because in interpreting them people construct equivalent mental models for the environment (Taylor & Tversky, 1992). However, it's unclear whether descriptions differing in person-centered viewpoint (i.e., whether they are egocentric vs. partner-centered) result in equivalent representations. Even though, based on such descriptions, people may construct spatial mental models that are equivalent in maintaining the spatial relations between objects, the preferred direction around which these spatial relations are organized may differ (see McNamara, 2003; Mou et al., 2004b, for a discussion of such allocentric representations having a preferred direction). This question hasn't been addressed with spontaneously produced descriptions.

In Section 3 we reviewed our findings from Galati et al (2011), where we examined how the availability of the partner's misaligned viewpoint influenced speakers' spatial memories and descriptions. We established that the perceptual information available during the description influenced the perspective of Directors' descriptions, with Directors using more Matcher-centered expressions when adopting their Matcher's viewpoint was relatively easy (when misaligned by a small offset). Additionally, we established that when partners knew each other's viewpoint in advance (in the Co-presence condition) they were better able to mutually recognize when perspective-taking would be most difficult for the Director (at the oblique offset of 135°) and agree on appropriate description strategies (describing arrays from the Director's perspective). Given these findings, in this section, we address whether the distribution of perspectives in the Directors' descriptions did in fact have an impact on how their Matchers organized their memory representations.

As we described in Section 3, the Matchers' memories were examined in the same way as the Directors', through JRDs and array drawings, after the description phase. To assess the preferred direction of Matchers' memory for the reconstructed arrays, in the JRD task, we examined Matchers' performance from headings aligned with their Director, from headings aligned with their own, and from all the remaining headings combined. Our goals were twofold: (1) first, to examine whether the Matchers' performance would be affected by the conditions during the description (the misalignment between partners and what Directors had known about the description phase in advance), and (2) insofar as these conditions affected the Directors' descriptions, as we had established in Galati et al. (2011), to examine whether Matchers' performance from either person-centered perspective in the JRD task correlated with the perspective of the Directors' descriptions.

Matchers' orientation and response latencies were affected somewhat differently by the conditions during the description. In terms of orientation latencies, the Director's perspective showed facilitation overall, as Figure 2 illustrates: Matchers were faster to orient to headings aligned with their Director than with themselves or all other headings. Notably, this facilitation of the Director's perspective was reliable in the No Intent and Intent conditions, but not in the Co-Presence condition where both partners had known each other's viewpoints in advance. The misalignment between partners during the description did not affect orientation times reliably, although, as Figure 2 suggests, when Matchers reconstructed arrays while offset by 135° from their partners, they were overall slower to adopt imagined headings than while offset by 90° or 180°.

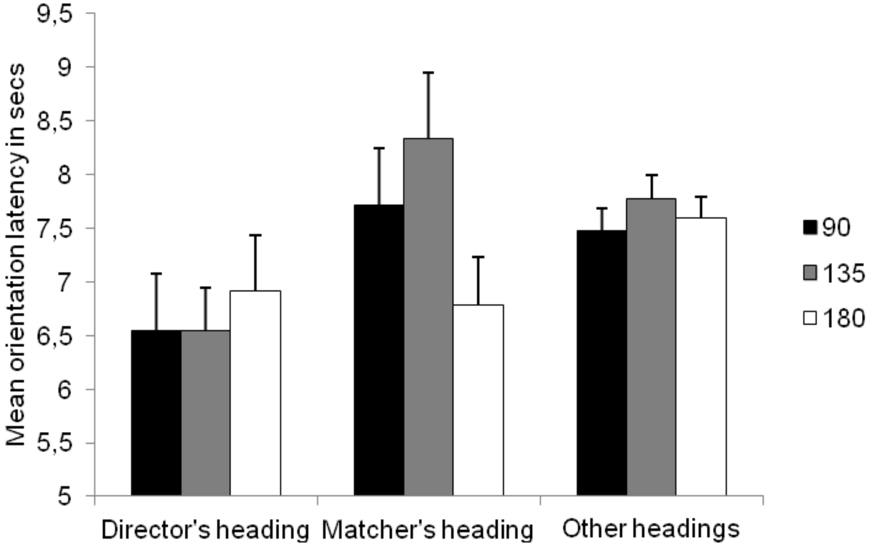


Fig. 2. The Matchers' mean orientation latencies from headings aligned with Directors, from headings aligned with themselves, and from other headings, across the three levels of misalignment between partners

On the other hand, in terms of response latencies, it was the Matcher's perspective that showed facilitation, as shown in Figure 3. Matchers were significantly faster to respond from headings aligned with their own than the Director or other headings. This pattern that held (reliably or marginally so), regardless of what the Director had known in advance about the description. As with orientation latencies, the misalignment between partners did not affect response latencies reliably, although, again, when Matchers reconstructed arrays while offset by 135° from Directors they were slower to respond than while offset by 90° or 180°.

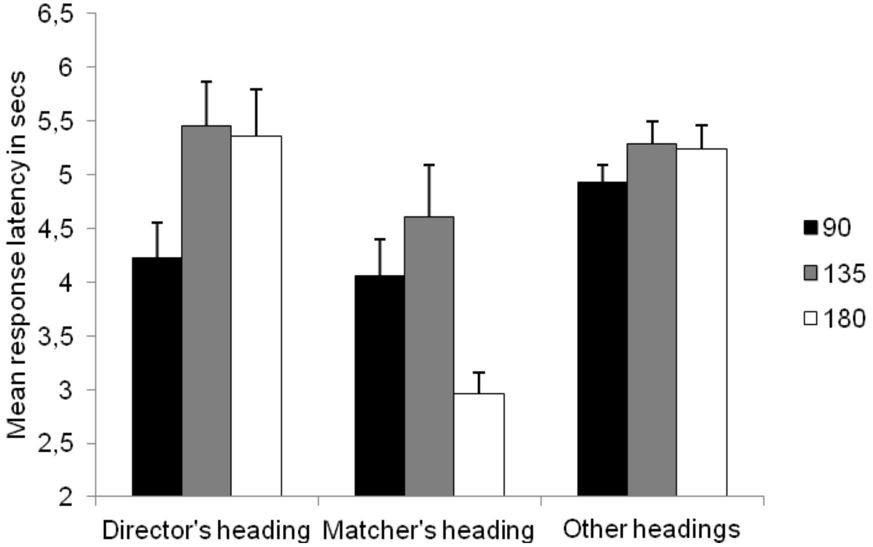


Fig. 3. The Matchers' mean response latencies from headings aligned with Directors, from headings aligned with themselves, and from other headings, across the three levels of misalignment between partners

Both latency measures, however, were affected similarly by the Directors' descriptions: the person-centered perspective of the Directors' descriptions was associated with facilitation in terms of both orienting to and responding from that heading. For orientation latencies, this was the case when partners had been offset by 90º: the greater the proportion of Matcher-centered expressions in the Directors' descriptions, the faster Matchers were to orient to headings aligned with their own, as suggested by a significant negative correlation. Considering that Directors were more likely to use Matcher-centered descriptions when offset by 90º than at greater offsets (Galati et al., 2011), we propose that the Directors' descriptions reinforced the Matcher's viewpoint as an organizing direction, and thus facilitated orienting to it. This negative correlation with Matcher-centered expressions also held for both latency measures in the Co-Presence condition, when Directors described arrays while knowing their Matcher's viewpoint in advance: the greater the proportion of Matcher-centered expressions in the Directors' descriptions, the faster Matchers were to orient to and respond from headings aligned with their own. Conversely, the greater the proportion of Director-centered expressions, the slower Matchers were to orient to and respond from headings aligned with their own. Additionally, when partners had been offset by 135º, as Matcher-centered expressions increased, Matchers were faster to respond from headings aligned with their own. That is, Matchers benefited especially from Matcher-centered expressions at the 135º offset, from which perspective-taking was demanding and from which Directors used primarily egocentric expressions in their descriptions (Galati et al., 2011).

6 Summary and Conclusions

The findings we report here contribute to a more nuanced understanding of partners' coordination in spatial tasks and of the memory representations that support and emerge from this coordination. Our study allowed us to examine whether certain circumstances (namely, the misalignment between partners' viewpoints and speakers' advance knowledge of it) affect whether speakers spontaneously incorporate their partner's viewpoint in memory and the description strategies they select to coordinate with their partners (Section 3). Additionally, our study allowed us to determine how these circumstances affect the efficiency of communication between partners (Section 4) and the memory representations that partners construct on the basis of speakers' descriptions (Section 5).

The first main conclusion emerging from our work is that, in collaborative tasks, when encoding spatial information and when subsequently describing it, people consider the cognitive demands of perspective-taking on themselves and on their partner, and adapt their representations and description strategies accordingly. When the partner's viewpoint is known while encoding spatial information, it seems to be represented, such that in spatial judgments orienting to it is slowed and in drawings the spatial configuration is rotated towards it. Also, knowing the partner's viewpoint in advance seems to enable partners to mutually recognize when coordinating is difficult, and to explicitly agree on strategies that reduce the cognitive demands on the partner with the greatest responsibility for mutual understanding. Speakers readily adopt their partner's perspective when perspective-taking is relatively easy, as when misaligned from their partner by a small offset. On the other hand, they opt for their own perspective, with their partner's consent or even initiative, when perspective-taking is difficult, especially when this is known while encoding the spatial information.

This adaption is consistent with the view that the attributions people make about their partner's ability to contribute to mutual understanding shapes behavior and leads to strategies that maximize the efficiency of communication (e.g., Duran et al., 2011; Brennan, 2004; Clark & Wilkes-Gibbs, 1986). Moreover, it underscores that the principles that govern coordination during spatial perspective-taking are not unlike those governing non-spatial perspective-taking (e.g., concerning their partner's conceptual perspective, their knowledge, or agenda; Schober, 1998). Partner-specific adjustments during both spatial and non-spatial perspective-taking appear to emerge from cognitive constraints acting on memory representations for shared experiences (see also Horton & Gerrig, 2005; Metzing & Brennan, 2003): if information about the partner is readily available or easily computed, it can be represented in memory and affect perspective-taking behavior; otherwise it won't.

Secondly, our findings suggest that the description strategies that partners select upon recognizing that coordination would be difficult are appropriate and successful in reducing their collective effort (thus maximizing the efficiency of communication). Partners who had known in advance that they would be misaligned by an oblique and computationally demanding offset were more efficient than partners who hadn't known, as reflected by the turns they took to reconstruct arrays. In fact, partners were numerically more efficient when they knew they would be misaligned by this oblique offset than by the other orthogonal offsets. The partners' efficiency is consistent with their explicit agreement to use the perspective of the partner for whom the task was most difficult in this perspective-taking situation. Thus, when partners were better able to realize that coordinating would be difficult, they selected appropriate and successful strategies. Currently we are examining other aspects of partners' coordination, beyond the effort they expended when coordinating. Specifically, by assessing the degree of distortion and rotational biases in how Matchers reconstructed the arrays on their tables (based on digital photographs of their reconstructions), we are investigating whether the strategies that partners deployed during the description affected not only their efficiency but also their accuracy in the task.

Additionally, our findings demonstrate that speakers' descriptions affect their conversational partners' resulting memory representations. The perspective of speakers' descriptions predicted the perspective that was facilitated when the partner subsequently made spatial judgments. This was especially so when partners had known each other's viewpoint in advance. The more speakers used partner-centered expressions in their descriptions, the more facilitation the partners showed for their own perspective; and conversely, the more speakers used egocentric expressions, the more facilitation the partners showed for the speakers' perspective.

Finally, our findings suggest that the partner who is reconstructing arrays based on another's spontaneous descriptions may represent both of their viewpoints in memory. Our Matchers' orientation latencies showed facilitation of their Directors' perspective, whereas their response latencies showed facilitation of their own perspective. It's not clear why the two latency measures were affected differently. Perhaps Directors served as a salient cue that helped Matchers to quickly adopt their imagined heading, thus facilitating orientation latencies. This cue may have been less relevant when identifying the location of a target after having adopted an imagined heading, in which case only headings aligned with Matchers' own showed facilitation in terms of response latencies. Nonetheless, Matchers appear to have represented both perspectives. This may be because Matchers, unlike their Directors, knew both of their viewpoints while they were reconstructing and learning arrays, and also because, despite any overarching preference, Directors' descriptions in all pairs included both Director-centered and Matcher-centered expressions.

Here, a comparison of the two collaborating partners' memory performance is pertinent. Previous studies have demonstrated that spatial information acquired through language results in memory representations that are functionally equivalent to those acquired through different sensory modalities (e.g., Avraamides & Kelly, 2010). Our findings don't address the issue of functional equivalence directly, as Directors and Matchers learned arrays under circumstances that differed not only in whether spatial information was acquired from vision vs. from language, but in other ways as well. For instance, Matchers learned arrays during the course of reconstructing them, on the basis of spontaneous descriptions, and while always knowing their partner's viewpoint. Directors, on the other hand, learned arrays through vision under non-interactive circumstances that were controlled, including with respect to whether or not they knew their partner's viewpoint. Despite these differences, both Directors' and Matchers' memory performance highlights that people construct their memory representations using the partner-specific information that is available, whether this information is available through the perceptual environment of the communicative situation (including the partner's position and orientation) or through descriptions emphasizing a particular viewpoint.

Together, our findings highlight some of the complex ways in which people adapt their memory representations and behavior when communicating spatial information. Future research can further clarify how coordination in spatial tasks, and the spatial representations supporting this coordination, are influenced both by partner-specific information and by egocentric preferences for organizing spatial information. But so far, it's evident that partners do consider each other's cognitive demands on the task when encoding and communicating spatial information. They are able to represent the partner's perspective when available—whether perceptually or through language—and they behave contingently: one partner's viewpoint influences the other's memory and description strategies, and in turn that partner's description strategies influences the other's memory.

Acknowledgments. Support was provided by research grant OSSMA-206912 from the European Research Council to M.N.A.

References

1. Avraamides, M.N., Kelly, J.W.: Multiple systems of spatial memory: evidence from described scenes. J. Exp. Psychol. Learn. Mem. Cogn. 36, 635–645 (2010)
2. Brennan, S.E.: How conversation is shaped by visual and spoken evidence. In: Trueswell, J., Tanenhaus, M. (eds.) Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-Action Traditions, pp. 95–129. MIT Press, Cambridge (2004)
3. Carlson-Radvansky, L.A., Irwin, D.E.: Frames of reference in vision and language: Where is above? Cogn. 46, 223–244 (1993)
4. Carlson-Radvansky, L.A., Logan, G.D.: The influence of reference frame selection on spatial template construction. J. Mem. Lang. 37, 411–437 (1997)
5. Clark, H.H.: Using language. Cambridge University Press, Cambridge (1996)
6. Clark, H.H., Brennan, S.E.: Grounding in communication. In: Resnick, L., Levine, J., Teasley, S. (eds.) Perspectives on Socially Shared Cognition, pp. 127–149. APA Books, Washington, DC (1991)
7. Clark, H.H., Krych, M.: Speaking while monitoring addressees for understanding. J. Mem. Lang. 50, 62–81 (2004)
8. Clark, H.H., Wilkes-Gibbs, D.: Referring as a collaborative process. Cogn. 22, 1–39 (1986)
9. Duran, N.D., Dale, R., Kreuz, R.J.: Listeners invest in an assumed other's perspective despite cognitive cost. Cogn. 121, 22–40 (2011)
10. Friedman, A., Kohler, B.: Bidimensional regression: Assessing the configural similarity and accuracy of cognitive maps and other two-dimensional data sets. Psychol. Methods 8, 468–491 (2003)

11. Galati, A., Michael, C., Greenauer, N., Mello, C., Avraamides, M.N.: How the conversational partner affects spatial memory and spatial descriptions. Talk given at the 17th European Society for Cognitive Psychology (ESCOP), Donostia-San Sebastian, Spain (2011)
12. Hölscher, C., Tenbrink, T., Wiener, J.M.: Would you follow your own route description? Cognitive strategies in urban route planning. Cogn. 121, 228–247 (2011)
13. Horton, W.S., Gerrig, R.J.: Conversational common ground and memory processes in language production. Discourse Process. 40, 1–35 (2005)
14. Isaacs, E.A., Clark, H.H.: References in conversations between experts and novices. J. Exp. Psychol. Gen. 116, 26–37 (1987)
15. Kelly, J.W., Avraamides, M.N., Loomis, J.M.: Sensorimotor alignment effects in learning and novel environments. J. Exp. Psychol. Learn. Mem. Cogn. 33, 1092–1107 (2007)
16. McNamara, T.P.: How Are the Locations of Objects in the Environment Represented in Memory? In: Freksa, C., Brauer, W., Habel, C., Wender, K.F. (eds.) Spatial Cognition III. LNCS (LNAI), vol. 2685, pp. 174–191. Springer, Heidelberg (2003)
17. Metzing, C., Brennan, S.E.: When conceptual pacts are broken: Partner-specific effects in the comprehension of referring expressions. J. Mem. Lang. 49, 201–213 (2003)
18. Mou, W., Zhang, K., McNamara, T.P.: Frames of Reference in Spatial Memories Acquired from Language. J. Exp. Psychol. Learn. Mem. Cogn. 30, 171–180 (2004)
19. Mou, W., McNamara, T.P., Valiquette, C.M., Rump, B.: Allocentric and egocentric updating of spatial memories. J. Exp. Psychol. Learn. Mem. Cogn. 30, 142–157 (2004)
20. Newman-Norlund, S.E., Noordzij, M.L., Newman-Norlund, R.D., Volman, I.A.C., de Ruiter, J.P., Hagoort, P., Toni, I.: Recipient design in tacit communication. Cogn. 111, 46–54 (2009)
21. Özyürek, A.: Do speakers design their co-speech gestures for their addressees? The effects of addressee location on representational gestures. J. Mem. Lang. 46, 688–704 (2002)
22. Schober, M.F.: Spatial perspective-taking in conversation. Cogn. 47, 1–24 (1993)
23. Schober, M.F.: Speakers, addressees, and frames of reference: Whose effort is minimized in conversations about location? Discourse Process. 20, 219–247 (1995)
24. Schober, M.F.: Different kinds of conversational perspective-taking. In: Fussell, S.R., Kreuz, R.J. (eds.) Social and Cognitive Psychological approaches to Interpersonal Communication, pp. 145–174. Lawrence Erlbaum, Mahwah (1998)
25. Schober, M.F.: Spatial dialogue between partners with mismatched abilities. In: Conventry, K.R., Tenbrink, T., Bateman, J.A. (eds.) Spatial Language and Dialogue, pp. 23–39. Oxford University Press, Oxford (2009)
26. Shelton, A.L., McNamara, T.P.: Spatial memory and perspective taking. Mem. Cogn. 32, 416–426 (2004)
27. Taylor, H.A., Tversky, B.: Descriptions and depictions of environments. Mem. Cogn. 20, 483–496 (1992)

Using Spatial Analogy to Facilitate Graph Learning

Linsey A. Smith and Dedre Gentner

Department of Psychology, Northwestern University
2029 Sheridan Road
Evanston, IL 60208
linsey@u.northwestern.edu,
gentner@northwestern.edu

Abstract. Graphical depictions of complex interactions pose a challenge to spatial reasoning. In this research, we tested whether analogical processes can be harnessed to help students learn to solve complex graphical reasoning problems. Specifically, we asked whether a brief training experience using spatial analogies could help students learn about *stock-and-flow* graphs. The basic idea of our intervention was to juxtapose contrastive graphs and encourage students to compare them. In two studies, we test the following predictions derived from structural alignment theories of analogy: (1) comparing contrastive graphs during training will lead to better performance in a graph-understanding task than will studying the same exemplars sequentially; and (2) comparing high-similarity pairs will lead to better performance than will comparing low similarity pairs. The results support both of these predictions, indicating that even a brief analogical comparison task can confer relational insight. Further, these results corroborate prior evidence that a structural alignment process underlies analogical comparison.

Keywords: Analogy, Analogical Comparison, Structural Alignment, Spatial Learning, Graphical Reasoning.

1 Introduction

Comparison of exemplars is a powerful learning process that has been shown to improve learning in a variety of domains. Indeed, according to Gentner [1], “The simple, ubiquitous act of comparing two things is often highly informative to human learners… Comparison is a general learning process that can promote deep relational learning and the development of theory-level explanations” (pp. 247, 251). Analogical comparison has been shown to aid learning across a broad range of topics, ranging from preschoolers learning new words [2] through elementary school children learning estimation methods [3] to business school students learning contract negotiation skills [4]. Within the spatial domain, there is evidence that spatial analogies can help learners to extract and use common spatial structure between two exemplars. For example, preschoolers who are given a challenging mapping task from one model room to another perform better if they first compare two models than if they interact with the same two models one-at-a-time [5].

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 196–209, 2012.

In this research we asked whether a brief analogical training experience, in which students were encouraged to make comparisons and identify contrasts, could help them learn important relational principles involved in complex graph integration problems. The outline of this paper is as follows. First, we lay out a theoretical framework for this work. We use the structure-mapping theory of analogy, which proposes that analogical comparison involves a process of structural alignment [6-7]. We then review research that illustrates how structural alignment is helpful for learning. Next we propose graph learning as a particularly fruitful domain in which to explore structural alignment as a learning tool, and introduce the specific kind of graphs that we investigated. We then present our experiments and review the results. We consider theoretical and applied implications of our findings, and close with a discussion of study limitations and future directions.

1.1 Analogical Comparison Fosters Learning

Comparison is powerful learning process [3], [7-8]. According to Structure-Mapping Theory (SMT) [6], [9-10], this is because comparison entails a structural alignment process that promotes a focus on common relational structure. This allows learners to move beyond superficial, possibly idiosyncratic features of particular examples [2], [11-12].

Under Structure-Mapping Theory [6], [9-10], carrying out a comparison involves aligning two structured representations so that matching objects and relations are placed into correspondence with one another (structural alignment). Once aligned, inferences can then be projected from one representation to another[1]. A key point of SMT is that common relations are more likely to be highlighted during comparison than are common object properties. This is because the structural alignment process favors matches that are connected to other matching information. For example, adults asked to match elements between two pictures are more likely to choose correspondences based on common relational role (rather than matching similar objects) if they have previously compared the two pictures [13].

Structural alignment paves the way for at least three distinct kinds of learning. First, as noted above, structural alignment highlights shared relational structure [4], [8], [13]. This can give rise to a new relational abstraction, which can then be transferred and applied to new situations [4], [8]. Second, rather paradoxically, highlighting commonalities also facilitates noticing differences that are connected to the shared structure, known as alignable differences [14-17]. A third consequence of structural alignment is that inferences may be brought from one situation to the other.

1.2 Analogical Comparison in Spatial Learning

While analogy (structural alignment) is a domain-general process, *spatial analogy* is a fundamental and pervasive kind of analogy. In spatial analogy, one or both analogs contain spatial relations. For example, one can use a cross-domain spatial comparison

[1] Many current models of analogical comparison have adapted these basic assumptions of SMT (for reviews, see Gentner & Forbus, 2011; Kokinov & French, 2003).

to describe the layered structure of Earth by likening it to the layers of a peach. Several studies show that within-domain, concrete spatial comparisons facilitate spatial learning [5], [11], [43]. For example, young children are successful at learning a non-obvious spatial concept when asked to compare two spatial structures, one of which exemplifies the concept and the other which does not [43]. As another example, children learn novel spatial relations better when they compare two exemplars that depict the relation than when they see the exemplars separately [11]. Most studies on spatial comparison have focused on concepts and examples that are almost entirely spatial. An open question is whether providing a spatial comparison can facilitate learning spatial representations with a strong conceptual component, such as graphs and diagrams, where the spatial representation serves to illustrate concepts that are not themselves spatial. There is reason to think that spatial analogy can encourage conceptual learning; in natural language, space is frequently analogized to abstract domains (e.g., *She was* in between *jobs*), indicating that spatial analogy can serve as a springboard for abstract, conceptual knowledge [45]. In this work, we begin to address the question of whether spatial comparison can simultaneously confer both spatial and conceptual relational insight. The current studies focus on learning about graphs, a particularly challenging type of spatial representation.

1.3 Graphs: A Complex Relational Task

Successful graph comprehension requires highly sophisticated spatial and conceptual reasoning. Graphs simultaneously convey spatial relations (one line above another) and conceptual relations (A exceeds B) [18]. It is widely accepted that graph comprehension entails at least three major, intertwined component processes [18-20]. First, viewers must encode the visual array and identify the important visuospatial relationships (e.g., a straight line slanting upward). Second, viewers must identify the underlying conceptual relations that those visuospatial relations represent (e.g., an increasing linear relationship between x and y). Finally, viewers must relate those relations to the variables depicted (e.g. a constant increase in carbon dioxide emissions over time). In sum, when one looks at a graph they must be able to simultaneously identify both the spatial and underlying conceptual relations depicted (see [21] for a related claim about diagrammatic representations more generally). Because of this relational complexity, it is not surprising that students of all ages have difficulties understanding graphs [18], [22-30].

Our question was whether analogical comparison—a process that promotes relational learning—would be a useful tool for learning the challenging spatial task of integrating complex graphical representations. In the experiments presented here, we focused on reasoning about stock-and-flow (SF) graphs. Conceptually, a stock is some entity amount that is accumulated over time by inflows and/or depleted by outflows. Stocks can only be changed via these flows. The amount of stock in a system is determined by the relationship between inflow and outflow: when inflow exceeds outflow, the stock will increase; when outflow exceeds inflow, the stock will decrease; and when inflow equals outflow, the stock will stabilize.

Stocks and flows are pervasive across domains—for example, they capture the dynamics of water in a bathtub (Figure 1), cash flow of a bank account, and CO2 levels in the atmosphere. These stock and flow relations are often depicted graphically, as in Figure 2. SF graph problems, even simple ones, are unintuitive and difficult, even for highly educated people with substantial training in science, technology, engineering, and mathematics (STEM) [23], [29-33].

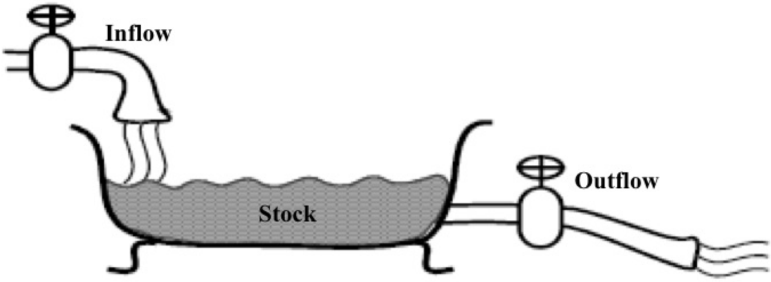


Fig. 1. Stocks and Flows in a bathtub. The amount of water in the tub is the stock. Water entering the tub through the faucet is the inflow. Water leaving the tub through the drain is the outflow.

1.4 The Current Experiments

In this set of studies, we tested whether presenting spatial analogies between graphical systems can help students learn to reason about stock-and-flow graphs like those depicted in Figure 2. The basic idea of our intervention was to juxtapose contrastive graphs and encourage students to compare them. This intervention was based on two principles of comparison processing derived from structure-mapping theory: (1) abstraction: analogical comparison reveals common structure [2-3], [8], [13]; and (2) contrast: analogical comparison highlights alignable differences—differences along a common dimension or predicate that plays the same role in the common structure [15-16].

These principles, taken together, predict that if learners align two analogous but contrasting examples, the common structure will become more salient and any alignable differences will become more noticeable [16]. This prediction has been borne out in studies of relational mapping and transfer in adults [4], [34] and children [5], [35-37], [43], in both conceptual and spatial domains. For example, Gentner et al. [43] found evidence that comparison can help children learn a non-obvious spatial concept, namely that *triangles confer stability in construction.* Specifically, when children were shown two toy buildings, a stable one that contained a triangle and a wobbly one that did not, children could use the alignment between them to identify the distinctive part (the triangle) as important for stability.

A third principle that is particularly relevant for research on learning is that alignment is easier and less error-prone for novice learners (both children and adults) when the items being compared are highly similar in their surface features as well as in their relational structure, i.e., the items are *literally similar* [38-41], [43].

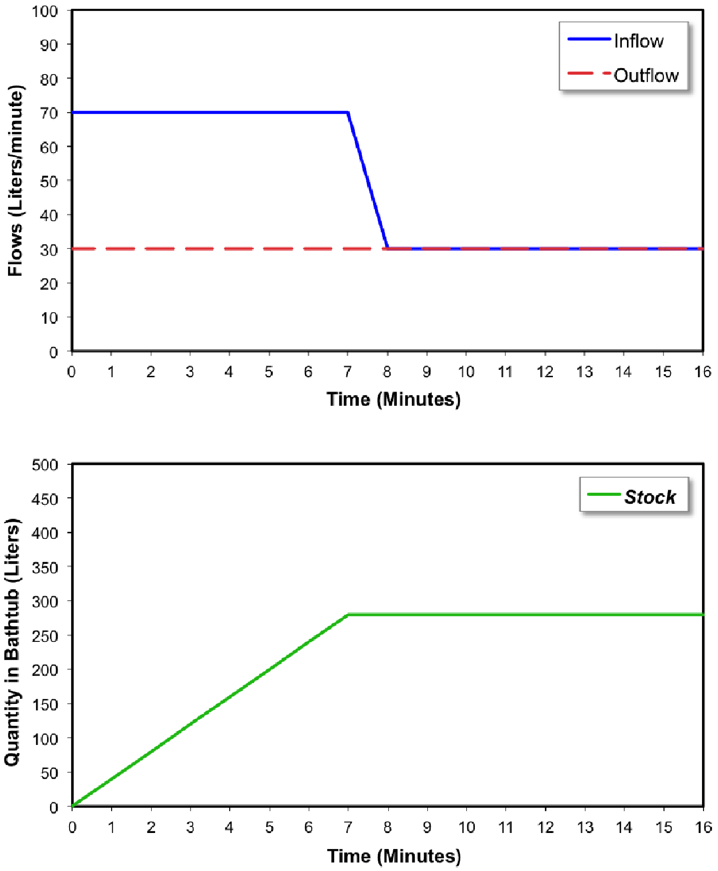


Fig. 2. A typical set of stock and flow graphs. The top graph depicts the changing rates of inflow from the faucet (solid line) and outflow through the drain (dotted line) over time. The line in the bottom graph shows the resulting change in the stock amount, i.e., bathtub water. Notice that as long as inflow exceeds outflow, the stock continues to rise. In contrast, when inflow equals outflow, the stock stabilizes.

Researchers suggest that the literal similarity advantage exists because object similarities support the required relational alignment. For example, in the part-learning study just described, young children (3-year-olds) were far better at aligning the creatures and noticing the contrasting parts when the pairs were highly similar (making them easy to align). Similar results have been obtained with adults. For example, Markman and Gentner [16] found that people list more relational similarities and alignable differences for literally similar scenes than for analogous scenes that contained fewer object- or surface matches. Even in online sentence processing, literal-similarity matches are processed faster than purely relational matches [42].

The studies consisted of a self-paced graph training task, followed by a set of graphical integration problems involving stocks and flows, which are described below. In the first study, we examined whether comparing examples leads to better performance on the graphical integration task than studying the same examples sequentially. In the second study, we varied the similarity of the pairs being compared during training, the details of which we will discuss later.

2 Experiment 1

2.1 Method

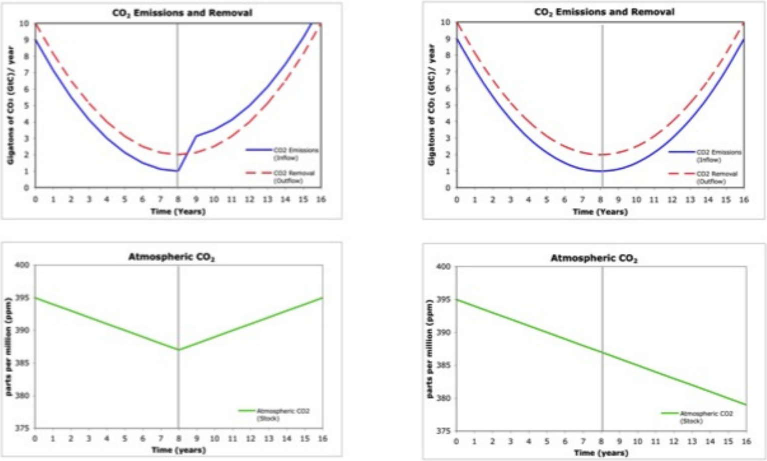
Participants. 32 undergraduate students from Northwestern University took part in the study individually or in groups of two. Participants completed the task in 15-25 minutes and for their time they received credit towards a course requirement.

Materials and Procedure. The experimenter gave one task booklet to the participant and upon completion they returned the booklet to the experimenter. The booklet contained a graph-training task followed by a graphical integration test. To make the task more concrete, all graphs were described in the context of CO_2 levels, where the stock was the amount of CO_2 in the atmosphere, inflow was the rate of CO_2 emissions, and outflow was the rate of CO_2 removal from the atmosphere (e.g., as it is taken up by plants).

Graph-Training Task. During the training phase, participants saw three examples of stock and flow graphs, similar to the graphs in Figure 3. To facilitate structural alignment, each example looked exactly the same up to the midpoint of the x-axis (time = 8). After the midpoint the examples differed in which of the three basic relationships between inflow, outflow, and stock they depicted: when inflow exceeded outflow, the stock was increasing; when outflow exceeded inflow, the stock was decreasing; and when inflow was equivalent to outflow, the stock was stable[2]. Thus, each of the examples only differed in one key relation between the three variables. Participants were randomly assigned to the *Sequential* or the *Comparison* training condition. In the Sequential condition, participants saw the three examples on separate pages. After seeing each example, participants were asked to explain the graphs by describing "What is going on in the TOP graph" and also "What is going on in the BOTTOM graph" (emphasis in the original instructions). The order in which the examples were shown was counterbalanced across participants. In the Comparison condition, participants saw two examples side-by-side and were asked to describe both similarities

[2] There are several more complex relations involving changes in net flow and the shape of the stock graph, but systematically varying those would compound the number of examples to be used. Thus in these studies we only focus on the three most basic relationships between stock and flows.

and differences between the two sets of graphs by listing “What is **similar** about the TOP (BOTTOM) graphs” and “What is **different** about the TOP (BOTTOM) graphs” (emphasis in the original instructions). Participants in the Comparison group only saw two stock-and-flow graph examples at one time; in order to make sure they saw all three examples, we gave them two separate comparisons to make. Thus, the Comparison group saw one of the examples twice (in two different comparison sets). The repeated example and the position of the example on the page (left or right) were counterbalanced across participants.

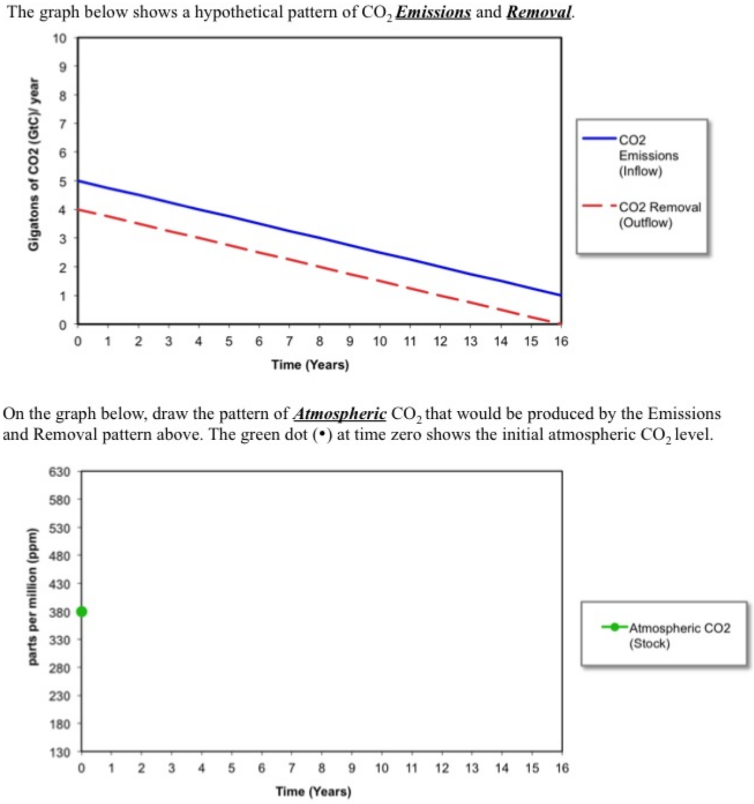


Example 1

Example 2

Fig. 3. Sample Comparison Examples. The inflow/outflow (top) graphs are the same until the midpoint, when the inflow (solid line) trajectory changes. Likewise, the stock (bottom) graphs are the same up until the midpoint, when the stock trajectory changes, corresponding to the change in the inflow/outflow graph. In the training task, participants were directed to compare and contrast the top two graphs, and then compare and contrast the bottom two graphs.

Graphical Integration Task. We adapted the *graphical integration* task from Booth Sweeney and Sterman [23]. In their original study, highly educated graduate students were presented with a picture of a bathtub and graphs showing the inflow and outflow of water, then asked to draw the trajectory of the stock of water in the tub. We used similar problems, although they were introduced in the context of CO_2 levels in the atmosphere (Figure 3). Participants solved seven graphical integration problems.


The graph below shows a hypothetical pattern of CO$_2$ ***Emissions*** and ***Removal***.

On the graph below, draw the pattern of ***Atmospheric*** CO$_2$ that would be produced by the Emissions and Removal pattern above. The green dot (•) at time zero shows the initial atmospheric CO$_2$ level.

Fig. 4. Sample Graphical Integration problem. Participants were given a graph that depicted inflows and outflows to the stock over time. They had to draw the resultant stock in the bottom graph.

2.2 Measures

Problem Score. For each graphical integration problem, participants received either 0 or 1 point. Participants received one point if their response maintained the three basic relations between stock and flow. For example, if the inflow was greater than outflow from t=0-8, then the participant needed to draw a stock that was continually increasing from t=0-8. Quantitative inaccuracies were not penalized. Participants could achieve a maximum score of 7 across the seven problems. Two raters blind to condition scored each problem. There was high interrater agreement, (96%, κ = 0.88); all disagreements were resolved through discussion.

2.3 Results and Discussion

Our prediction was that participants who were given the opportunity to compare contrastive graphs would perform better on the graphical integration problems. This prediction was borne out in the data. Participants who compared examples performed better (M=4.75, SE=0.49) on the graphical integration test than participants who studied the examples separately (M=3.32, SE=0.78), t(30)=2.18, p<.05, d=0.77. Why do we see this performance advantage for the Comparison group? We suggest that the act of comparing the graphs enabled people to both (1) identify the relations common to both graphs and (2) notice relational contrasts between them. That is, when learners were given the opportunity to align two analogous but contrasting examples, the common structure became more salient and the alignable differences between the graphs were more noticeable [16]. These two phenomena are exemplified in the similarity/difference listings from two of the Comparison participants:

- "From t=0-8 inflow exceeds outflow." (Similarity)
- "From t=8-16 inflow still exceeds outflow in [the top left graph], but inflow is less than outflow in [the top right graph]." (Difference)

- "Both CO_2 contents increase from time 0 to 9" (Similarity)
- "In [the bottom left graph]; total stock CO_2 goes down after 8 yrs. vs [the bottom right] graph where the stock CO_2 value continues to increase." (Difference)

Our results are consistent with the claim that the structural alignment process both highlights common relational structure and accentuates alignable differences. Furthermore, these data suggest that spatial analogy can facilitate learning about spatial representations with a strong conceptual component. In experiment two, we wanted to test a further prediction of structural alignment models of analogical comparison—namely, comparing examples that share greater overall similarity (i.e., surface and structural similarity) will be more beneficial for learning than comparing examples where there is less surface similarity.

3 Experiment 2

Prior work demonstrates that structural alignment is easier for learners when the items being compared are highly similar in their surface features as well as in their relational structure [38], [43]. The claim is that, in cases of high similarity, surface similarity works in the service of relational similarity, and thus effectively guides learners to the correct alignment. Maximizing the likelihood that a learner achieves a successful structural alignment increases the likelihood that they will notice important relational commonalities and differences. Thus, a greater likelihood of successful alignment should translate into a greater likelihood of successful relational learning. Several studies have demonstrated a learning advantage for high similarity comparisons. In the toy building task mentioned above, children learned better when the two compared buildings shared high surface similarity, in contrast to low surface similarity [43]. Most studies that report a high similarity advantage in learning by comparison have focused on children's learning [5], [11], [43]; our question is whether we will see a similar advantage for high similarity with adults in a complex arena such as graph integration. In this study, we varied the similarity of examples that participants compared during training. Participants either compared graphs that shared both relational (structural) similarity and surface similarity—i.e., they had high overall similarity—or they compared graphs that shared structural similarity but were perceptually dissimilar—they had low overall similarity.

3.1 Method

Participants. 62 undergraduate students from Northwestern University took part in the study individually or in groups of two. Participants completed the task in 15-25 minutes and for their time they received credit towards a course requirement.

Materials and Procedure. The procedure was as in Experiment 1—an experimenter handed a booklet to the participant. Upon completion the participant gave the booklet back to the experimenter. The booklet contained a graph-training task followed by a graphical integration test.

High Alignment vs. Low Alignment Training. All participants compared examples during training, what differed was the overall similarity between the examples. One group of participants compared example graphs that shared both structural similarity and perceptual similarity. Specifically, the compared graphs contained the same relations between variables. For example, in Figure 3 both of the top graphs show outflow above (exceeding) inflow from t=0-8. In addition, the trajectories or shapes of the lines in the graphs were similar; in Figure 3, for example, the outflow line is parabolic in both graphs. These graphs were considered highly alignable because they shared both relational and surface similarity. For the sake of clarity, we call this the *Same Shape* training condition. Another group of participants compared graphs that maintained relational similarity, but were less perceptually similar. Thus, the same relations between inflow, outflow and stock were present (e.g., outflow exceeds inflow), but the shapes of the variable lines were different (e.g., the inflow was a parabolic function in one graph and an exponential function in the other). These graphs were considered less alignable because surface similarity could not facilitate alignment to the same degree. We call this the *Different Shape* condition. Participants were asked to list the similarities and differences for each comparison set, as in Experiment 1.

Graphical Integration Test. The graphical integration test was as in Experiment 1.

Measures

Problem Score. We scored each graphical integration response as in Experiment 1. For each graphical integration problem, participants received either 0 or 1 point, for a maximum of 7 points across seven problems.

3.2 Results

As predicted, participants who compared Same Shape examples performed better on the graphical integration problems (M=4.35, SE=0.40) than those who compared Different Shape examples (M=3.32, SE=0.47), t(60)=1.67, $p<0.05$, d=0.42, one-tailed. Overall, these results are consistent with our prediction that performance is related to the ease of alignment, with students who were exposed to High Alignability (Same Shape) training performing better than those that were exposed to Low Alignability (Different Shape) training. Comparing highly similar graphs enabled people to more easily identify the important relational commonalities and differences between the graphs, as exemplified in one participant's similarity/difference listings:

- "For the first 8 years, the CO_2 removal (outflow) is greater than CO_2 emission inflow" (Similarity)
- "After 8 years, [the left graph] has a greater inflow than outflow while [the right graph] has same amount of inflow and outflow" (Difference)

In contrast, comparing less similar graphs made it more difficult for people to focus on the relevant relational commonalities and contrasts. Below is a representative similarity/difference listing for participants in the Different Shape condition. These participants tended to describe superficial characteristics of the graphs rather than relational aspects.

- "They both measure inflows and outflows of CO_2; they have the same key, and the same axis measurements; same colors; same titles" (Similarity)
- "[The left graph] is smooth; [the right graph] is straight until sharp junction" (Difference)

In sum, we found that pairs that were easier to spatially align (because they were perceptually similar) were more helpful in training, and led to better performance on the graphical integration test, than pairs that were more difficult to align. These results are consistent with the claim that, in early learning, comparing examples that are readily alignable—such as pairs that share overall similarity—is especially beneficial [5], [37], [43].

4 Discussion

These experiments provide initial evidence that the principles of structure-mapping can be used effectively to promote students' learning in a domain with a high degree of relational complexity. Specifically, spatial alignment (spatial analogy) of examples facilitated the sophisticated spatial and conceptual reasoning required for the task. Participants who compared examples of stock-and-flow graphs during training were able to transfer their understanding to graphical integration problems. Our results also support the claim that ease of spatial alignment contributes to graph learning. Participants who saw perceptually similar graphs were better able to align them and notice the key relational commonalities and differences between the variables on the graphs—e.g., that inflow exceeds outflow. This advantage for ease of spatial alignment is consistent with prior findings on spatial learning [5], [43].

In future work, we aim to further explore variability in the surface and structural similarity between examples. It would also be useful to identify other aspects of graphical examples that may make them easier or harder to align. Another issue that should be explored is how to better facilitate learning via comparison. In our studies, overall performance across conditions was not at ceiling—participants have room to grow in their learning. In addition to exploring the issue of optimal variation in examples, it would also be useful to develop ways to guide the comparison process more effectively. In the above studies, the comparison task was fairly open-ended—people were only asked to describe similarities and differences between the graphs. Prior work has shown that greater scaffolding during the comparison process leads to better learning [44]; it seems likely that constructing a more guided comparison task would be advantageous for helping students hone in on the multitudinous and complex relations embedded within graphs.

Overall, our findings offer evidence that spatial analogical alignment can be used effectively for graph learning. In our study, detailed predictions from structure-mapping theory and research were found to be applicable for promoting students' graphical learning and reasoning.

Acknowledgements. This research was supported by NSF grant SBE-0541957, the Spatial Intelligence and Learning Center (SILC). We thank Garrett Honke and the Cognition and Language Lab for help with this research.

References

1. Gentner, D.: The development of relational category knowledge. In: Gershkoff-Stowe, L., Rakison, D.H. (eds.) Building Object Categories in Developmental Time, pp. 245–275. Erlbaum, Hillsdale (2005)
2. Namy, L.L., Gentner, D.: Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. J. Experimental Psychology: General 131(1), 5–15 (2002)
3. Star, J.R., Rittle-Johnson, B.: It pays to compare: An experimental study on computational estimation. J. Experimental Child Psychology 101, 408–426 (2009)
4. Gentner, D., Loewenstein, J., Thompson, L.: Learning and transfer: A general role for analogical encoding. J. of Educational Psychology 95(2), 393–405 (2003)
5. Loewenstein, J., Gentner, D.: Spatial mapping in preschoolers: Close comparisons facilitate far mappings. J. of Cognition and Development 2(2), 189–219 (2001)
6. Gentner, D.: Structure-mapping: A theoretical framework for analogy. Cognitive Science 7(2), 155–170 (1983)
7. Gentner, D., Markman, A.B.: Structure mapping in analogy and similarity. American Psychologist 52, 45–56 (1997)
8. Gick, M.L., Holyoak, K.J.: Schema induction and analogical transfer. Cognitive Psychology 15(1), 1–38 (1983)

9. Gentner, D.: Why we're so smart. In: Gentner, D., Goldin-Meadow, S. (eds.) Language in Mind: Advances in the Study of Language and Thought, pp. 195–235. MIT Press, Cambridge (2003)
10. Gentner, D.: Bootstrapping children's learning: Analogical processes and symbol systems. Cognitive Science 34(5), 752–775 (2010)
11. Christie, S., Gentner, D.: Where hypotheses come from: Learning new relations by structural alignment. J. Cognition and Development 11(3), 356–373 (2010)
12. Gentner, D., Medina, J.: Similarity and the development of rules. Cognition 65, 263–297 (1998)
13. Markman, A.B., Gentner, D.: Structural alignment during similarity comparisons. Cognitive Psychology 25, 431–467 (1993)
14. Gentner, D., Gunn, V.: Structural alignment facilitates the noticing of differences. Memory and Cognition 29(4), 565–577 (2001)
15. Gentner, D., Markman, A.B.: Structural alignment in comparison: No difference without similarity. Psychological Science 5(3), 152–158 (1994)
16. Markman, A.B., Gentner, D.: Splitting the differences: A structural alignment view of similarity. J. Memory and Language 32, 517–535 (1993)
17. Gentner, D., Sagi, E.: Does "different" imply a difference? A comparison of two tasks. In: Sun, R., Miyake, N. (eds.) Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society, pp. 261–266. Erlbaum, Mahwah (2006)
18. Carpenter, P., Shah, P.: A model of the perceptual and conceptual processes in graph comprehension. J. Experimental Psychology: Applied 4, 75–100 (1998)
19. Tversky, B.: Spatial schemas in depictions. In: Gattis, M. (ed.) Spatial Schemas and Abstract Thought, pp. 79–111. MIT Press, Cambridge (2001)
20. Pinker, S.: A theory of graph comprehension. In: Freedle, R. (ed.) Artificial Intelligence and the Future of Testing, pp. 73–126. Erlbaum, Hillsdale (1990)
21. Gattis, M.: Mapping relational structure in spatial reasoning. Cognitive Science 28(4), 589–610 (2004)
22. Bell, A., Janvier, C.: The interpretation of graphs representing situations. For the Learning of Mathematics 2(1), 34–42 (1981)
23. Booth Sweeney, L., Sterman, J.D.: Bathtub Dynamics: Initial Results of a Systems Thinking Inventory. System Dynamics Review 16, 249–294 (2001)
24. Culbertson, H.M., Powers, R.D.: A study of graph comprehension difficulties. AV Communication Review 7, 97–110 (1959)
25. Gattis, M., Holyoak, K.J.: Mapping conceptual to spatial relations in visual reasoning. J. Experimental Psychology: Learning, Memory, and Cognition 22, 231–239 (1996)
26. Kozhevnikov, M., Hegarty, M., Mayer, R.E.: Revising the visualizer/verbalizer dimension: Evidence for two types of visualizers. Cognition and Instruction 20, 47–77 (2002)
27. Maichle, U.: Cognitive processes in understanding line graphs. In: Schnotz, W., Kulhavy, R.W. (eds.) Comprehension of Graphics, pp. 207–226. Elsevier Science, New York (1994)
28. Shah, P., Carpenter, P.A.: Conceptual limitations in comprehending line graphs. J. Experimental Psychology: General 124, 43–61 (1995)
29. Sterman, J.D., Booth Sweeney, L.: Cloudy Skies: Assessing Public Understanding of Global Warming. System Dynamics Review 18, 207–240 (2002)
30. Sterman, J.D., Booth Sweeney, L.: Understanding Public Complacency About Climate Change: Adults' Mental Models of Climate Change Violate Conservation of Matter. Climatic Change 80, 213–238 (2007)

31. Cronin, M.A., Gonzalez, C., Sterman, J.D.: Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. Organizational Behavior and Human Decision Processes 108, 116–130 (2009)
32. Cronin, M., Gonzalez, C.: Understanding the building blocks of system dynamics. System Dynamics Review 23(1), 1–17 (2007)
33. Pala, Ö., Vennix, J.A.M.: Effect of system dynamics education on systems thinking inventory task performance. System Dynamics Review 21(2), 147–172 (2005)
34. Catrambone, R., Holyoak, K.J.: Overcoming contextual limitations on problem-solving transfer. J. Experimental Psychology: Learning, Memory, and Cognition 15(6), 1147–1156 (1989)
35. Gentner, D., Namy, L.: Comparison in the development of categories. Cognitive Development 14, 487–513 (1999)
36. Mutafchieva, M., Kokinov, B.: Does the family analogy help young children to do relational mapping? In: Proceedings of the European Conference on Cognitive Science, pp. 407–412. Erlbaum, Hillsdale (2007)
37. Gentner, D., Loewenstein, J., Hung, B.: Comparison facilitates children's learning of names for parts. J. Cognition and Development 8, 285–307 (2007)
38. Gentner, D., Ratterman, M.J., Forbus, K.D.: The roles of similarity in transfer: Separating retrievability from inferential soundness. Cognitive Psychology 25, 524–575 (1993)
39. Gentner, D., Toupin, C.: Systematicity and surface similarity in the development of analogy. Cognitive Science 10, 277–300 (1986)
40. Paik, J.H., Mix, K.S.: Preschooler's use of surface similarity in object comparisons: Taking context into account. J. Experimental Child Psychology 95(3), 194–214 (2006)
41. Richland, L.E., Morrison, R.G., Holyoak, K.J.: Children's Development of Analogical Reasoning: Insights from Scene Analogy Problems. J. Experimental Child Psychology 94, 246–273 (2006)
42. Gentner, D., Kurtz, K.: Relations, objects, and the composition of analogies. Cognitive Science 30, 609–642 (2006)
43. Gentner, D., Levine, S., Dhillon, S., Poltermann, A.: Using structural alignment to facilitate learning of spatial concepts in an informal setting. In: Kokinov, B., Holyoak, K.J., Gentner, D. (eds.) Proceedings of the Second International Conference on Analogy. NBU Press, Sofia (2009)
44. Kurtz, K.J., Miao, C., Gentner, D.: Learning by analogical bootstrapping. J. Learning Sciences 10(4), 417–446 (2001)
45. Lakoff, G., Johnson, M.: Metaphors We Live By. University of Chicago Press, Chicago (1980)

Activity Effects on Path Integration Tasks for Children in Different Environments

Eva Neidhardt[1] and Michael Popp[2]

[1] Universität Koblenz-Landau, Campus Koblenz, Germany
neidhardt@uni-koblenz.de
[2] Universität der Bundeswehr, München, Germany
michael.popp@unibw.de

Abstract. In each of the three presented studies kindergarten children and school children walked a path of about one kilometer in a macro environment At up to six locations subjects stopped and were asked to point into the direction of the path origin with their outstretched arm and finger and later with a mechanical pointer or, in the case of the virtual environment, with a laser pointer. Pointing accuracy was taken as a measure for path integration. Kindergarten children from small German towns and from a primary school in Namibia as well as school children from Munich were tested. The Munich school children were also assessed in a virtual reality condition. Results indicate that children's activity reports influence pointing accuracy. Implications for gender differences and ideas on affordances of children's future real environments are discussed.

Keywords: Spatial activity, children, path integration, pointing, virtual reality.

1 Introduction

Experiments with animals convincingly demonstrated a mechanism called „path integration" (e.g. Maurer & Séguinot, 1995), a processing of travel information, which enables shortcut finding after walking a new path with several segments. Path integration can be defined as the ability to demonstrate knowledge about bearing and distance of the direct connection between two not directly related points. In humans path integration can be measured by differences between distance estimates and true distance or by angular deviations between pointing direction and true direction e.g. walking a path with several turns and at the end pointing to the origin (starting point). Path integration in animals has been established by observation of shortcut-behaviour. Etienne, Maurer and Saucy (1988) proved the importance of visual and motor information systems for path integration in rodents.

However, in one of the best-known theories of spatial representation in humans (e.g. Siegel & White, 1975), it is assumed that survey knowledge has to be installed before human persons are able to find shortcuts: Landmarks serve as anchor points for route knowledge. The routes have to be integrated into a network of landmarks and

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 210–219, 2012.

routes, combined with certain features such as bearings and distances. According to Siegel and White (1975) survey knowledge is necessary primarily to shortcut finding or pointing ability. Now it is well accepted that the involved processes are more sophisticated.

1.1 Cognitive Processes Involved in Path Integration

Dead reckoning is a strategy where actual information is continuously integrated to stay informed about the position relative to a reference point, e.g. home. Piloting is mainly based on external acoustic or visual signals. Dead reckoning uses internal and external information about velocity and direction. Both mechanisms seem to work complementarily for spatial orientation in macro spatial environments (Etienne, 1992). Most daily tasks involve coordination of self-to-object distances and direction (Rieser & Pick, 2007). Distance estimation is known to depend on visual and acoustic ("optical and auditory flow cues", see Rieser & Pick, 2007) as well as body sense information (e.g. Kearns, Warren, Duchon, & Tarr, 2002; Popp, Platzer, Eichner & Schade, 2004). Locomotion is crucial in this process of updating spatial position information (Rieser & Pick, 2007).

The ability to know the bearing to the origin of a walked path, hence to be able to point to this target which cannot be seen at the moment of pointing, is only one component emerging from the spatial information updating process. As integration processes with respect to walking along several path segments are assumed, this will be called *path integration*. Dead reckoning is hence an important component of path integration. Dead reckoning suffers from drifting errors, resulting from inexact spatial updating. These drifting errors are probably small with short paths, and they probably accumulate to relevant deviation angles in pointing, when the path gets longer. Hence, additional external information is needed to improve the spatial information system and to render information about the own position more reliable (Newcombe & Huttenlocher, 2000). A *cognitive map* as a mental representation, i.e. personal spatial knowledge resulting from experiencing spatial activities, may provide further information to improve actual position information (Rieser & Pick, 2007). Even pre-school children obviously use both kinds of information, path integration (i.e. spatial updating in a kind of dead reckoning process) and cognitive maps (i.e. personal spatial knowledge). Both, drifting with longer paths and better pointing in familiar surroundings can be demonstrated experimentally (Neidhardt, 2002). Distance of the walked path and familiarity with the surrounding are clearly situational factors that affect pointing accuracy.

1.2 Research Focus

The focus here is on dispositions, competencies emerging from spatial activity experiences in a more abstract sense. Several authors have claimed that orientation abilities have to be learnt as well as other cognitive abilities (e.g. Rieser & Pick, 2007). For the special path integration competence measured by pointing to the invisible origin of a path just walked it was shown that preschoolers who have places outside their homes where they go on their own, unaccompanied, show better pointing in real as well as in virtual environments (e.g. Neidhardt & Popp, 2010). School children with more real world spatial activity experience also have better results when pointing to invisible objects at locations just passed by while walking (Neidhardt & Schmitz, 2001).

Here the thesis will be proved more systematically, integrating aspects of cultural comparisons. Cultural differences implying different home ranges (i.e. areas where children can move without being surveyed) have been claimed to be responsible for a range of spatial abilities, for example for memory of location (Ecuyer & Robert, 2004). They are also discussed as source of gender related differences in many aspects of spatial cognition.

For all studies presented below, the hypothesis to be tested is whether children with more spatial experience show better pointing accuracy in path integration tasks than children with less spatial experience.

2 General Methods

For kindergarten children the decisive factor will be the answer to the question if there is any place outside home where they go on their own, without parents or other guiding people. For school children this question is no longer valid as separation predictor between high and low spatial experience groups: Children in Germany normally still walk to school unaccompanied. Exceptions are children in very large cities: in the Munich school children sample there were only few who reported to get to school on their own without parents or older children going with them. The other school children were asked if they spend a lot of their free time "roaming around".

In all studies presented below the environments differed in many aspects, yet we carefully chose paths with similar characteristics: They all were of about one kilometer length, there were six pointing locations from where the origin of the path (target where the children had to point to) could not be seen nor any adjacent potential landmark, and the correct pointing direction from each of the pointing locations deviated more than 30° from the path so that just pointing back could not be successful.

3 Study 1

In the first experiment 33 children (four to six years old) from a kindergarten in Wetzlar, a small rural German town, were led along two paths of about one kilometer each. The two groups compared differed in the time they normally spend outdoor in the forest. As in general, the hypothesis for this study is that spatial experience predicts pointing accuracy.

3.1 Method

Subjects: 20 boys and 13 girls from a kindergarten in the small rural town Wetzlar participated in this experiment. About half of them (13 boys and 3 girls, mean age M = 5.30 yrs, SD = 1.05 yrs) were members of a forest project team: For at least six months they had not entered the kindergarten building. They had chosen to spend all their kindergarten time outdoors in the forest. A little site caravan there served as shelter for this group. This group will be called "the forest group". The other half of the children (7 boys and 10 girls, mean age M = 4.94 yrs (SD = 1.04 yrs) had never been part of the forest project team. They will be named "the indoor group" – which, strictly speaking, is not accurate as every German kindergarten group enjoys outdoor activities.

Procedure: The children were tested individually. First they answered to some interview questions about demographic data and way finding experience („Are there places you are going to alone, without your parents, for example to see your friends?"). This short interview served as warming-up procedure between child and experimenter. After the interview the child was asked to accompany the experimenter on a path leading either from the kindergarten (town path) or from the site caravan in the forest (forest path) to an endpoint about half a mile away. The town path lay in an area of small houses with little gardens, while the forest path lead to a little river. On their way, children were asked to point to the kindergarten or to the site caravan (target), respectively from six different locations. The target or any direct hint to the target could not be seen from any of the pointing locations. Pointing was first done directly without any additional instrument, and was measured via compass. In a second assessment children had to use a pointer to show the correct bearing. This procedure was chosen because earlier tests had proven that children's pointing performance gets worse when they only use the pointer and the experimenter's bearing assessment is better when looking at the pointer compared to reading the compass. Absolute deviation between the correct bearing according to GPS information and children's pointing was taken as pointing accuracy.

3.2 Results

First we compared the forest group's and the indoor group's pointing results. Contrary to our expectations in the ANOVAs with group (forest vs. indoor group) as independent factor and pointing error (in degrees) as dependent factor there were no significant differences for the path in the forest and only a small effect ($F(1,32)= 4.3$, $p<.05$, $\eta^2=.12$) in the town.

In a second analysis we put all children together and formed new groups: the "alone group" with children who answered "yes" to the question if there were places outside home they went without company (n=11) and the "accompanied group" with children who said that there were no places outdoor which they visited on their own (n=22).

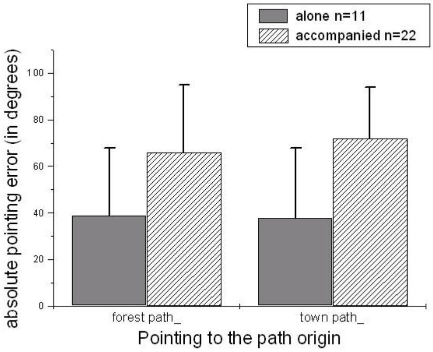


Fig. 1. Pointing accuracy of children going on their own somewhere outdoor ("alone group") compared to children who only go in company ("accompanied group")

As can be seen from figure 1 the "alone group" performs clearly better than the "accompanied group": In analyses of variance with individual travel as independent factor (walking alone vs. always accompanied) this difference is significant for the forest path (F(1,32)= 5.9, p<.05, η^2=.16) and very clear for the town path (F(1,32)= 13.8, p<.001, η^2=.32). Taken together and controlling for adherence to the forest group or the indoor group in a 2 (walking alone vs. always accompanied) x 2 (forest path vs. town path) analysis of variance with group adherence as covariate, and path (forest vs. town) as repeated measure the effect is still significant (F(1,30)= 12.5, p<.005, η^2=.29). The interaction effect for path (forest vs. town) x group adherence fails to reach significance (F(1,30) = 3.2, p<.10, η^2=.10), all other effects are far from reaching significance (p>.10).

3.3 Discussion

This experiment shows that "outdoor activity" in itself seems not to be the important factor but rather the question if the children are allowed to take responsibility for their path finding. Still, evidence for this hypothesis cannot be convincing from this study alone as the groups are rather small.

In the last years 190 kindergarten children in seven studies were conducted in about the same way (Neidhardt, 2004): Children walked paths in familiar surroundings starting at the kindergarten. At several locations they were asked to point to the kindergarten door. Path integration was assessed as described above as absolute deviation from correct pointing. The meta analysis from these data shows very clearly that "alone group" children perform better than "accompanied group" children (fixed combined effect of the seven studies: $_{\text{Hedges g}}$(7,72,118) = .53, p<.001). Our interpretation to this finding is that children's experiences in taking responsability for outdoor wayfinding improve spatial orientation competencies relevant for path integration.

4 Study 2

Only recently our "forest children" group was compared to a group of kindergarten children in Namibia, Africa. Those children usually go everywhere on their own. Therefore their pointing accuracy was assumed to be even better than the accuracy of German "alone group" children. Other potential factors may work in favour of this hypothesis as less building density or against this hypothesis as the fact that children's homes are farer away from their kindergarten in the African compared to the German sample, yet we estimated these factors not to be as influential as children's self-initiated outdoor experience.

4.1 Method

Subjects: In the Namibia study 14 boys and 16 girls from Mphe Thuto Primary School in Tjsaka participated. Mean age was 5.6 yrs (SD = 0.4 yrs). All children were "alone group" children: All stated that they would go on their own everywhere they wanted in their free time.

Procedure: The procedure was the same as for the German children. The path was chosen so that the Namibian children could not see the origin of the path, Mphe Thuto Primary School, from any of the six pointing locations. The study was conducted in Namibia by Sarah Monzel.

4.2 Results

As can be seen in figure 2 the Namibian children clearly outperformed the German children

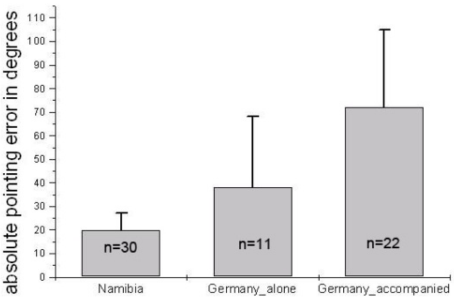


Fig. 2. Pointing accuracy of kindergarten children in Namibia compared to German children who go alone (middle) or do not go alone to places outdoor

The difference is significant even if only the German "alone group" is taken into account (ANOVA with country (German vs. African children) as independent variable and with absolute difference to correct bearing when pointing to the kindergarten door as dependent variable: F(1,40)=9.8, p<.005, η^2=.20). The German sample is small, however, the other German samples do not perform better as every other comparison also shows Namibian children's superiority in this task.

4.3 Discussion

From earlier studies (e.g. Neidhardt & Popp, 2010) it is known that 20° is measuring variation in this task. Even if African children can point more accurately this cannot be measured with this kind of task. This means that within measuring accuracy the kindergarten children from Mphe Thuto Primary School in Tjsaka show perfect path integration. Evidently, our interpretation that the huge home range in the Namibian sample may be responsible for their good results cannot be more than a further hint as there are many other differences between German and Namibian children than just wayfinding experience.

5 Study 3

Munich is one of the big German cities with lots of traffic. Here, children in first and second grade are brought to school by their parents. Even in second grade about half of the pupils are not allowed to visit their friends on their own. In this study second grade children were brought to an unfamiliar surrounding within the Universität der Bundeswehr Campus. As before it was supposed that children who go outdoor on their own ("alone group") show better pointing accuracy in the path integration task than the group of children who are always accompanied ("accompanied group"). In this study the main question was if there was a difference between real and virtual environments in the spatial activity effect.

5.1 Method

Subjects: 39 second graders from two classes of an inner city school in Munich (19 girls, 20 boys) with a mean age of 7.6 yrs (SD=0.6 yrs) participated in this study.

Procedure: The experiment took place on the campus of the Universität der Bundeswehr in Neubiberg in the south east of Munich. The area consists of approximately 100 buildings, nested between dense vegetation with small paths and without any road names and other orientation aids. The children were brought by bus to the university campus for two days together with their teachers. School lessons were given in one of the campus rooms. Children individually left their class and were brought to the starting point of the real walking path or to the vision dome. The VR "NeuViberg"is a nearly perfect copy of the reality. It is shown to the subjects in form of an immersive 180° projection in a 5m VisionDome device. (s. figure 3).

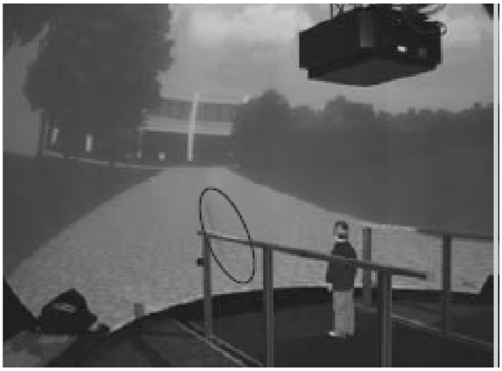


Fig. 3. Vision-dome (virtual reality condition)

All children were tested first in the real world condition and then in the virtual environment.

Real world condition: As starting "point" a large stone had been chosen. Children were first asked to point to the stone with the pointer to get used to it. As in the studies before these children walked a path on the university campus of about one kilometer length, but this time it was certain that they did not previously know the path as the campus is not open to public. At six locations along the path, the children were asked to point to the starting point (stone) which was invisible from the pointing locations.

Virtual environment condition: Each child was positioned in the center of the Vision-dome, looking towards the starting point. The experimenter demonstrated the appearance of the rotating scene and trained the use of the fixed pointer to indicate directions (see above) using the position of the piece of rock as the starting point. After that, each child walked virtually to the six pointing locations. At each pointing location the scene started to rotate and the child was asked to say "stop", whenever the fixed pointer in the center of the rotating scene pointed to the direction of the rock at the starting place. That direction was noted.

5.2 Results

When relating the results of the Munich second graders in an unfamiliar real world surrounding (figure 4, left columns) to those of small town kindergarten children in a familiar environment (figure 1) it is obvious that the pointing errors are about the same size but that standard deviations are larger in the unfamiliar surrounding. Pointing errors obviously get much larger in the virtual environment.

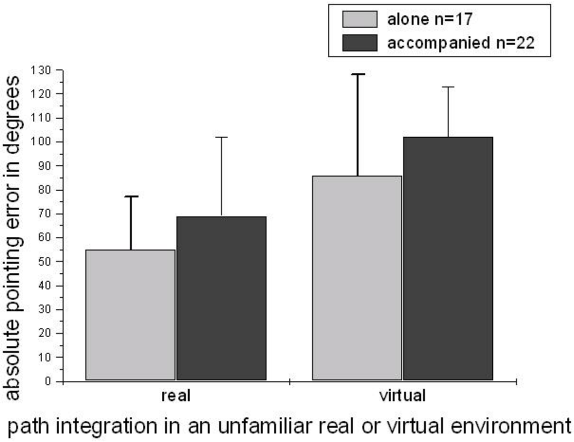


Fig. 4. Munich second graders' pointing accuracy at the Universität der Bundeswehr campus

The differences between children who said that they were allowed to visit their friends on their own (fig. 4, grey columns, "alone group") and those children who were denied this possibility (fig. 4, black columns, "accompanied group") were only marginally significant in both the real world and the virtual reality environment, due to huge standard deviations. Still it can be seen that there is no interaction effect (η^2<.01). The activity factor is working in both conditions, in the real world as well as in the virtual environment condition. Taken together the effect is statistically significant (2 factor (alone vs. accompanied) ANOVA: F1,36=6.5, p<.05, η^2=.15).

5.3 Discussion

Even though the activity effect is small it can still be demonstrated despite the difficulty of the path integration task in a very complex unfamiliar real or virtual surrounding.

6 General Discussion

In all three studies it was shown that children who report to be allowed to go outdoors on their own and who evidently make use of this liberty show better pointing results. Spatial orientation, measured here with a very special task, seems to be better in those "alone group" children. Choosing their own path seems to help children develop this competence. A very convincing argument is the finding that children who are not restricted in their freedom of movement outdoor in Namibia, show almost perfect pointing accuracy. Children's environment in Germany's big cities prevents many school children even in first and second grade from exploring their environment on their own. In a Munich sample of second graders as in a lot of kindergarten samples self directed outdoor path finding seems to make a difference in spatial orientation.

Between studies children varied with respect to cultural environment: In study 1 children from Wetzlar participated, a small German town, in study 3 children from

Munich, one of Germany's largest cities, in study 2 children from Wetzlar and from Tjsaka in Namibia. They all go to school or kindergarten, yet there are probably huge differences in many aspects of their daily life. Our studies focused on one aspect only: children's self-guided, i.e. unaccompanied activities outdoors. This factor was estimated by asking the children and it quasi-experimentally varied between groups. Ignoring all other possible sources this factor showed significant effects for children's bearing estimates in all three studies.

From many studies we also know that in Germany girls are more often in an "accompanied group". Sex related differences are not focused here. Restricted home range does not only hinder spatial development in girls: German towns and cities seem to give priority to traffic fluency, thereby discouraging parents to allow children to explore their "natural" surrounding on their own. If we want our children to develop normal spatial competencies city planners should reflect about priorities. Children need safe areas near their homes.

References

Ecuyer Dab, I., Robert, M.: Spatial Ability and Home-Range Size: Examining the Relationship in Western Men and Women (Homo sapiens). Journal of Comparative Psychology 118, 217–231 (2004)

Etienne, A.S.: Navigation of a small mammal by dead reckoning and local cues. Current Directions in Psychological Science 1, 48–52 (1992)

Etienne, A.S., Maurer, R., Saucy, F.: Limitations in the assessment of path dependent information. Behaviour 106, 81–111 (1988)

Kearns, M.J., Warren, W.H., Duchon, A.P., Tarr, M.: Path integration from optic flow and body senses in a homing task. Perception 31, 349–374 (2002)

Maurer, R., Séguinot, V.: What is modelling for? A critical review of the models of path integration. Journal of Theoretical Biology 175, 457–475 (1995)

Neidhardt, E.: Orientierung bei Vorschulkindern: Zwei Feldexperimente zur Pfadintegration. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie 34(4), 185–193 (2002)

Neidhardt, E., Popp, M.: Spatial tests, familiarity with the surroundings and spatial activity experience–how do they contribute to spatial orientation in real macro environments? Journal of Individual Differences 31, 59–63 (2010)

Neidhardt, E., Schmitz, S.: Entwicklung von Strategien und Kompetenzen in der räumlichen Orientierung und in der Raumkognition: Einflüsse von Geschlecht, Alter, Erfahrung und Motivation. Psychologie in Erziehung und Unterricht 48(4), 262–279 (2001)

Newcombe, N.S., Huttenlocher, J.: Making space: The development of spatial representation and reasoning. The MIT Press, Cambridge (2000)

Popp, M.M., Platzer, E., Eichner, M., Schade, M.: Walking With and Without Perception of Distance in Large Scale Urban Areas in Reality and in Virtual Reality. Presence 13(1), 61–76 (2004)

Rieser, J.J., Pick, H.L.: Using locomotion to update spatial orientation: What changes with learning and development? In: Plumert, J.M., Spencer, J.M. (eds.) The Emerging Spatial Mind (2007)

Siegel, A.W., White, S.H.: The development of spatial representations of large-scale environments. In: Reese, H.W. (ed.) Advances in Child Development and Behavior, vol. 10, pp. 10–55. Academic Press, New York (1975)

Influence of Rotational Axis and Gender-Stereotypical Nature of Rotation Stimuli on the Mental-Rotation Performance of Male and Female Fifth Graders

Sarah Neuburger[1], Vera Heuser[1], Petra Jansen[2], and Claudia Quaiser-Pohl[1]

[1] University of Koblenz-Landau, Institute of Psychology,
Universitaetsstrasse 1, 56070 Koblenz, Germany
{neuburger,vheuser,quaiser}@uni-koblenz.de
[2] University of Regensburg, Institute of Sports Science,
Universitaetsstrasse 31, 93053 Regensburg, Germany
Petra.Jansen@psk.uni-regensburg.de

Abstract. The male advantage in the mental rotation of two- or three-dimensional objects in mind is well documented across various age groups. The current study examined the influence of task characteristics on this gender difference by comparing the mental rotation performance of 148 fifth-grade boys and girls in three stimulus conditions (male-stereotyped objects, female-stereotyped objects, Shepard and Metzler's cube figures) and two rotational-axis conditions (rotations in picture plane only vs. rotations in depth). In line with the hypotheses, boys slightly outperformed girls in the in-depth condition, but not in the picture-plane condition. Unexpectedly, however, boys tended to outperform girls in the female-objects task. Overall, results suggest that rotational axis is more influential in determining the gender difference than the stereotyped nature of the stimuli. Findings are discussed with regard to the influence of working memory on mental rotation.

Keywords: mental rotation, gender differences, stimulus material, rotational axis.

1 Introduction

1.1 Theoretical Background

In mental rotation, which is a subcomponent of spatial abilities and refers to the rotation of two- or three-dimensional objects in mind, male subjects are usually found to outperform female subjects [1]. The male performance advantage has been demonstrated in various age groups, including pre-adolescent children [2-6] and infants [7-8]. With regard to the gender difference in age groups younger than ten years, results are however mixed [e.g. 4 vs. 5-6]; therefore, the age in which the male advantage emerges is still a controversial topic. The causes of the gender effect in

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 220–229, 2012.

mental rotation are most appropriately conceptualized within a psychobiosocial framework [9-10], because both biological factors, e.g. specific genes [11-12] and sex hormone levels [13], and socio-cultural processes, e.g. stereotypes [14], gender role identity [15-16] and patrilineal versus matrilineal society structures [17] have been found to influence the spatial test performance of male and female subjects. Some findings suggest the gender difference in mental rotation to depend on task characteristics [18]. The largest gender effect of about one standard deviation has been found in the "Mental Rotations Test" [MRT, 19-20]. This might be due to several features of the MRT, among them stimulus material and rotational axis. The MRT requires the mental rotation of cube-figures drawings with foreshadowed depth dimension, originally designed by Shepard and Metzler [21], in all three Cartesian axes, and both stimulus characteristics and dimensionality/axis of the rotation might contribute to the gender difference.

With regard to stimulus features, gender stereotypes as well as the degree to which stimulus-similar objects are part of the everyday environment of male versus female subjects might contribute to the gender effect. Objects similar to the cube figures of the MRT, like blocks, dominos, cube puzzles, and LEGO material, are more frequently part of boys' environment [22-23]. Because of the gender difference in stimulus familiarity, boys are more likely to process such stimuli holistically [24], which might in turn support efficient mental rotation. Furthermore, cube figures might activate gender stereotypes of male superiority because they remind of male-stereotyped objects and thus lead to stereotype threat effects [14, 25].

Usually, the cube figures of Shepard and Metzler [21] are used in in-depth rotation tasks, i.e. target stimuli have to be rotated not only around the picture-plane z-axis, but also around the depth-plane y-axis. Neuburger et al. [6], who assessed the mental-rotation performance of elementary-school children, used picture-plane rotation of the cube figures and found a considerably smaller male advantage in fourth graders than Titze et al. [5], who used in-depth rotations. Thus, in addition or alternatively to stimulus features, rotational axis provides an explanation for the large gender effect that is found in the MRT compared to other mental-rotation tasks.

However, the greater size of the gender effect in the MRT might also simply be explained by task difficulty, which in turn is probably influenced by both stimulus features and rotational axes [26]. In order to test this assumption, it seems reasonable to compare male and female subjects' performance levels across several mental-rotation tasks, thus examining if the size of the gender effect simply varies as a function of task difficulty or if there are certain task characteristics (e.g. stimulus features or rotational axis) that differentially affect male and female subjects' performance beyond task difficulty.

Since the gender effect in the MRT is reliably found from the age of ten years onwards and since male and female subjects' mental-rotation strategy has been found to differ already in fifth graders [27], the age group of fifth graders was chosen for the current study, which examined the effect of gender-stereotypical stimuli and rotational axis on the gender effect in mental-rotation performance.

1.2 Design and Hypotheses

The study had a 2x2x3-mixed design with the between-subjects variables gender and rotational axis and the within-subjects variable stimulus material. Subjects solved three mental-rotation tasks, each with a different stimulus type (male-stereotyped objects, female-stereotyped objects, cube figures). About half of the subjects solved picture-plane mental-rotation tasks, in which the target stimuli had to be rotated only around the Cartesian z-axis, and the other half of the subjects solved in-depth rotation tasks, in which target stimuli were additionally rotated around the Cartesian y-axis. Rotational axis was varied between subjects for practical reasons (participating schools allowed a test duration of one lesson only); furthermore, substantial order effects would have been likely to occur in a within-subject design in which children would have solved two rotation tests with the same stimulus material, because stimulus familiarity has been found to induce substantial training effects and changes in strategy use [24, 28], and these order effects would have complicated the design and data interpretation. Stimulus material was varied within subjects so that correlations between the three tasks could be computed as a preliminary validity indicator of the newly designed male- and female-objects tasks. It was expected that the male advantage would be larger in the in-depth rotation tasks than in the picture-plane tasks, and also larger with the male-stereotyped objects and the cube figures than with the female-stereotyped objects.

2 Method

2.1 Participants

By contacting local schools, 148 fifth graders, who were between 9.50 and 11.92 years old (M = 10.71 years; SD = 0.37) were recruited to participate in the study. As tests were administered in classes, each class was randomly assigned to one of the two conditions. In the picture-plane condition, 60 subjects (41 boys/19 girls) were tested, and in the in-depth condition, 88 subjects (56 boys/32 girls) were tested. There were no significant age differences between the two conditions or between boys and girls (all $p > .10$). Parents as well as the participants themselves gave their written, informed consent.

2.2 Material

Task format was the same as in the "Mental Rotations Test" of Vandenberg and Kuse [19]. The paper-pencil tasks consisted of twelve test items with one target on the left side and four comparison stimuli on the right (see Fig. 1). Two of the four comparisons were rotated versions of the target and had to be crossed out by the participants. The male and female stereotyped objects were constructed with 3ds Max 2012 (http://www.autodesk.de). The male objects in the test items were: car, digger, soccer goal, hammer, cannon, model airplane, corsair, revolver, saw, screw wrench, toy soldier, and tractor. The two male practice items were: truck and screw driver. The female test items were: buggy, ballet slipper, ironing board, iron, hairbrush, handbag, necklace, dress, pot, doll, hair ribbon. The two female practice items were: mirror and teapot. The cube-figures task was taken from Titze, Jansen, and Heil [29]. In addition to the mental-rotation tasks, a questionnaire of perceived gender-stereotyped nature of the stimuli was administered, in which the male and female objects and one cube figure were displayed in a random order and rated by the subjects on a scale from 1 ("typical for boys") to 5 ("typical for girls").

2.3 Procedure

Before administering the first mental-rotation task, the concept of "mentally rotating objects" was introduced by rotating a real, familiar object (a pair of scissors) in front of the class. In the next step, the mental-rotation task was explained on an overhead projector, and 2-3 practice items were solved. In each class, the cube-figures task was administered after the two other tasks; in half of the classes, the male-objects task was solved before the female-objects task, and in the other half of the classes, the female-objects task was solved before the male-objects task. For each task, subjects had 3 minutes to correctly solve as many items as possible. After finishing the three mental-rotation tasks, the questionnaire assessing the perceived gender-stereotyped nature of the stimuli was administered.

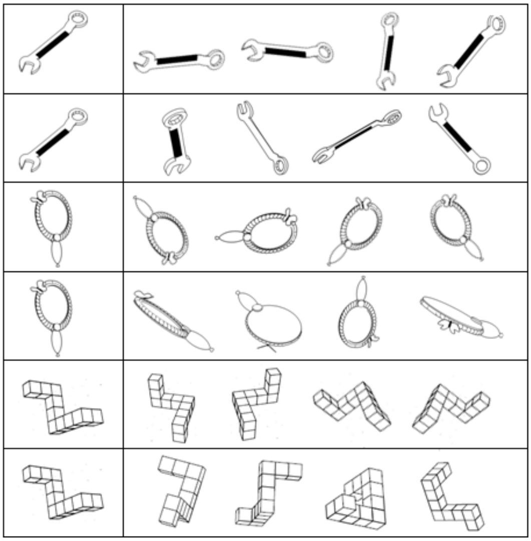


Fig. 1. Example items from the mental-rotation tasks (male objects/picture-plane, male objects/in-depth, female objects/picture-plane, female objects/in-depth, cube figures/ picture plane, and cube figures/in-depth)

3 Results

Analysis of the gender-stereotyped nature questionnaire showed that the mean ratings of the male objects (M = 1.67; SD = 0.44) significantly diverged from 3 ("neither typical for boys nor for girls") towards the male pole of the scale, t (147) = 36.67; p < .001, while the mean ratings of the female objects (M = 4.29; SD = 0.33) significantly diverged from 3 towards the female pole of the scale, t (147) = 47.81; p < .001.

Although the mean ratings of the cube figure were close to the gender neutral value (M = 2.69; SD = 0.72), they also differed significantly from 3 towards the male pole, t (146) = 5.17; p < .001. Thus, the gender-stereotyped nature of the stimulus material was confirmed.

In both rotational axis conditions, the three mental-rotation tasks correlated significantly: In the picture-plane condition, the highest correlation was found between the male-objects and the female-objects task (r = .71; p < .001), and the female-objects task correlated more strongly with the cube-figures task (r = .60; p < .001) than the male-objects task (r = .38; p < .01); in the in-depth rotation, the female- and male-objects tasks correlated even more strongly with the cube-figures task (female objects: r = .51; p < .001, male objects: r = .43; p < .001) than with each other (r = .34; p < .001).

A 2 (gender) x 2 (rotational axis) x 3 (stimulus material)-ANOVA with number of correctly solved items as dependent variable revealed the following main effects: First, there was a significant main effect of rotational axis, F (1,144) = 72.67; p < .001; ŋ² = .34: performance was significantly higher in the picture-plane condition than in the in-depth condition. Second, there was a significant main effect of stimulus material, F (1,288) = 66.74; p < .001; ŋ² = .32: performance in the cube-figures task was lower than in the male-objects task (p < .001; d = 0.57) and in the female-objects task (p < .001; d = 0.62); however, performance did not differ between the male-objects and the female-objects task (p > .10; d = 0.04). No main effect of gender was found, F (1,144) < .001; p > .10; ŋ² < .001. In addition to the main effects, the following interactions were found: First, there was a significant interaction of gender and rotational axis, F (1,144) = 4.82; p < .05; ŋ² = .03; simple effect analyses (see Fig. 2) showed a nonsignificant higher performance of girls in the picture-plane condition, F (1,58) = 1.52; p = .22; ŋ² = .025, and a marginally significant higher performance of boys in the in-depth condition, F (1,86) = 3.89; p = .052; ŋ² = .043.

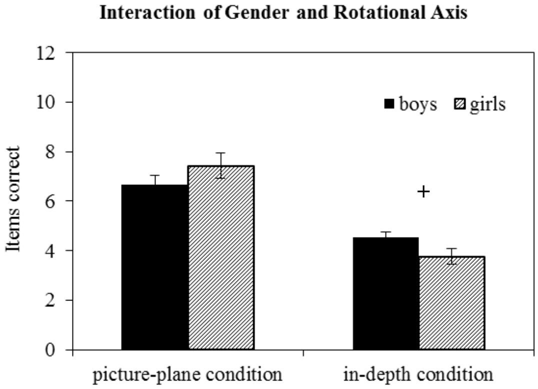


Fig. 2. Mental-rotation performance (means) as a function of gender and rotational axis. Error bars represent standard errors of the mean; + = marginally sign. difference, p < .10.

Second, there was a significant interaction of gender and stimulus material, F (2,144) = 5.30; p < .01; ŋ² = .04; however, simple effect analyses did not show any significant gender difference: on the descriptive level, boys slightly outperformed girls in the female-objects task, F (1,146) = 2.43; p = .12; ŋ² = .016, and in the cube figures task, F (1,146) = 1.25; p = .27; ŋ² = .008, while girls very slightly outperformed boys in the male objects task, F (1,146) = 0.57; p = .45; ŋ² = .004. As the absence of a gender difference in the cube-figures task is in contrast to the usually reported large gender differences in this task, boys' and girls' scores on the cube-figures task were compared between the picture-plane and the in-depth condition. In line with the hypotheses, a significant male advantage was found in the in-depth condition (d = 0.56; p < .05), while a nonsignificant female advantage was found in the picture-plane condition (d = 0.32; p > .10).

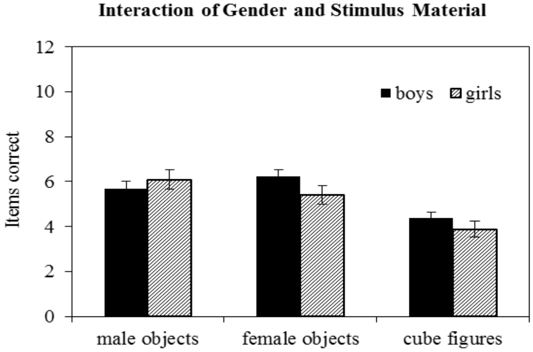


Fig. 3. Mental-rotation performance (means) as a function of gender and stimulus material. Error bars represent standard errors of the mean.

The overall ANOVA also revealed a significant interaction between rotational axis and stimulus material, F (2,288) = 60.84; p < .001; ŋ² = .30; simple effect analyses showed a significantly higher performance in the picture-plane condition for the male-objects task, F (1,146) = 127.45; p < .001; ŋ² = .466; and for the female-objects task, F (1,146) = 88.81; p < .001; ŋ² = .378, but not for the cube-figures task, F (1,146) = 0.40; p = .84; ŋ² < .001. The three-way interaction of gender, rotational axis, and stimulus material did not reach significance, F (2,288) = 1.08; p > .10; ŋ² = .01.

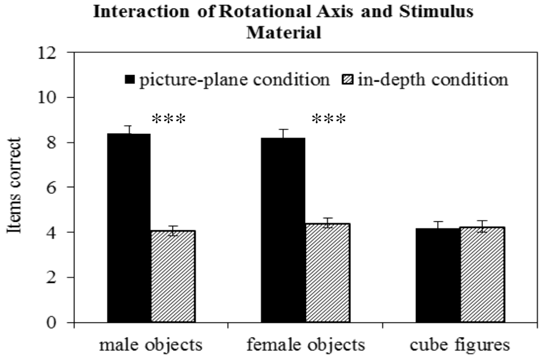


Fig. 4. Mental-rotation performance (means) as a function of rotational axis and stimulus material. Error bars represent standard errors of the mean; *** = sign. difference, p < .001.

4 Discussion

Results suggest the gender effect in mental rotation to depend on rotational axis. The male advantage emerged only with in-depth rotations, i.e. when target stimuli had to be rotated in the Cartesian y-axis, but not in picture-plane rotations, i.e. when target stimuli had to be rotated only in the Cartesian z-axis/ line of sight. This effect might be explained by gender-specific differences in spatial experience with regard to the rotation of foreshortened objects in three-dimensional space, e.g. in computer games [30-31]. Since, overall, performance was lower in the in-depth condition, the larger gender effect in this condition might also be due to the higher task difficulty, which would be in line with previous studies showing that sufficiently difficult two-dimensional rotation tasks also produce a large male advantage in adults [32]. However, in the present study, performance level in the cube-figures task did not differ between the picture-plane and the in-depth condition. Nevertheless, the male advantage disappeared in the picture-plane cube-figures task. Thus, the effect of the rotational axis on the gender difference in the cube-figures task cannot be explained by a higher task difficulty in the in-depth condition.

In the current study, the gender effect was also slightly influenced by stimulus material. However, the direction of the interaction between gender and stimulus material was not in line with the hypothesized effects of stimulus stereotypicality; therefore, and in face of the nonsignificant simple effect results, the effects of stimulus material cannot be interpreted as straightforward as the effects of the rotational axis. The unexpected direction of the stimulus by gender-interaction might have been due to specific spatial features of the female and male stereotyped objects in combination with gender-specific differences in mental-rotation strategy [27], which might have made the male objects easier for girls, and the female objects easier for boys. Furthermore, this effect might have been caused by attentional differences, e.g. due to an increased effort in case of own-gender inconsistent stimuli which led to a deeper level of processing and thus greater rotation accuracy. Another explanation might be that there have been differences in subjects' certainty criterion – maybe the well known own-gender consistent stimuli reduced the carefulness and accuracy with which subjects solved the task. Both mechanisms would support the crucial role of working memory as a main component of mental rotation [33-34]: solving a mental-rotation task requires maintaining a spatial image (storage component) and its simultaneous transformation (processing component). Previous studies demonstrated a substantial male advantage in the speed of information processing in visuospatial working memory [e.g. 34]. Spatial working memory has been found to mediate the effects of gender on mental-rotation performance [35]; furthermore, working memory capacity is strongly affected by stereotype threat, i.e. the situational activation of negative stereotypes [36-37]. In the female-objects condition of the present study, it was intended to reduce stereotype threat for girls and thus to promote their task-related information processing in working memory by replacing the MRT cube figures by female-stereotyped stimuli. However, results suggest that these stimuli did not promote girls' performance. Independent of stimulus material, the gender effect disappeared in the picture-plane rotations, probably because of the reduced working memory load in these two-dimensional tasks.

The present study is limited by the fact that differences between the complexity and familiarity of the three stimulus types were not controlled. The concrete, familiar stimuli of the male-objects and female-objects task are more easily represented and more likely processed as a unitary mental image than the abstract cube figures. This difference probably explains the higher performance of both boys and girls in the two objects tasks. However, performance in the male-objects and the female-objects task did not differ; therefore, these two tasks appear to be comparable with regard to the average complexity and familiarity of the included stimuli. Interestingly, in the in-depth condition, performance in the two objects tasks dropped to the cube-figures level, which might have been due to the more difficult distractor stimuli in the two objects tasks compared to the cube-figures task, in which some distractors differ from the target with regard to feature differences and are thus more easily detected. Since the three tasks were not parallelized concerning stimulus familiarity, stimulus complexity, and distractor stimuli, this study should be considered as a pilot study for further investigations. Further studies should examine the impact of specific stimuli and task features on different subprocesses of male and female subjects' mental-rotation, which might contribute to a better understanding of the causal mechanisms underlying this gender difference and the cognitive process of mental rotation itself.

References

1. Voyer, D., Voyer, S., Bryden, M.P.: Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. Psychological Bulletin 117, 250–270 (1995)
2. Linn, M.C., Petersen, A.C.: Emergence and characterization of sex difference in spatial ability: a meta-analysis. Child Development 56, 1479–1498 (1985)
3. Johnson, E.S., Meade, A.C.: Developmental patterns of spatial ability: An early sex difference. Child Development 58, 725–740 (1987)
4. Levine, S.C., Huttenlocher, J., Taylor, A., Langrock, A.: Early sex differences in spatial skill. Developmental Psychology 35, 940–949 (1999)
5. Titze, C., Jansen, P., Heil, M.: Mental rotation performance and the effect of gender in fourth graders and adults. European Journal of Developmental Psychology 7, 432–444 (2010)
6. Neuburger, S., Jansen, P., Heil, M., Quaiser-Pohl, C.: Gender differences in pre-adolescents' mental-rotation performance: Do they depend on grade and stimulus type? Personality and Individual Differences 50, 1238–1242 (2011)
7. Moore, D.S., Johnson, S.P.: Mental rotation in human infants: a sex difference. Psychological Science 19, 1063–1066 (2008)
8. Quinn, P.C., Liben, L.S.: A sex difference in mental rotation in young infants. Psychological Science 19, 1067–1070 (2008)
9. Halpern, D.F., Wai, J., Saw, A.: A psychobiosocial model: Why females are sometimes greater than and sometimes less than males in math achievement. In: Gallagher, A.M., Kaufman, J.C. (eds.) Gender Differences in Mathematics: An Integratice Psychological Approach, pp. 48–72. Cambridge University Press, New York (2005)

10. Hausmann, M., Schoofs, D., Rosenthal, H.E.S., Jordan, K.: Interactive effects of sex hormones and gender stereotypes on cognitive sex differences – a psychobiosocial approach. Psychoneuroendocrinology 34, 389–401 (2009)
11. Bock, R.D., Kolakowski, D.: Further evidence of sex-linked major-gene influence on human spatial visualizing ability. American Journal of Human Genetics 25, 1–14 (1973)
12. Pezaris, E., Casey, M.B.: Girls who use "masculine" problem-solving strategies on a spatial task: Proposed genetic and environmental factors. Brain and Cognition 17, 1–22 (1991)
13. Hausmann, M., Slabbekorn, D., Van Goosen, S.H.M., Cohen-Kettenis, P.T., Güntürkün, O.: Sex hormones affect spatial abilities during the menstrual cycle. Behavioral Neuroscience 114, 1245–1250 (2000)
14. Moè, A., Pazzaglia, F.: Following the instructions! Effects of gender beliefs in mental rotation. Learning and Individual Differences 16, 369–377 (2006)
15. McGlone, M.S., Aronson, J.: Stereotype threat, identity salience, and spatial reasoning. Journal of Applied Developmental Psychology 27, 486–493 (2006)
16. Ortner, T.M., Sieverding, M.: Where are the gender differences? Male priming boosts spatial skills in women. Sex Roles 59, 274–281 (2008)
17. Hoffman, M., Gneezy, U., List, J.A.: Nurture affects gender differences in spatial abilities. Proceedings of the National Academy of Science 25 (2011), doi:10.1073/pnas.1015182108
18. Jansen-Osmann, P., Heil, M.: Suitable stimuli to obtain (no) gender differences in the speed of cognitive processes involved in mental rotation. Brain and Cognition 64, 217–227 (2007)
19. Vandenberg, S.G., Kuse, A.R.: Mental rotations. A group test of three-dimensional spatial visualization. Perceptual and Motor Skills 47, 599–604 (1978)
20. Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., Richardson, C.: A redrawn Vandenberg and Kuse mental rotations test: Different versions and factors that affect performance. Brain and Cognition 28, 39–58 (1995)
21. Shepard, R.N., Metzler, J.: Mental rotation of three-dimensional objects. Science 171, 701–703 (1971)
22. Connor, J.M., Serbin, L.A.: Behaviorally based masculine- and feminine-activity-preference scales for preschoolers: Correlates with other classroom behaviors and cognitive tests. Child Development 48, 1411–1416 (1977)
23. Etaugh, C.: Introduction. The influence of environmental factors on sex differences in children's play. In: Liss, M.B. (ed.) Social and Cognitive Skills: Sex Roles and Children's Play, pp. 1–19. Academic Press, New York (1983)
24. Bethell-Fox, C.E., Shepard, R.N.: Mental Rotation: Effects of Stimulus Complexity and Familiarity. Journal of Experimental Psychology: Human Perception and Performance 14, 12–23 (1988)
25. Moè, A.: Are males always better than females in mental rotation? Exploring a gender belief explanation. Learning and Individual Differences 19, 21–27 (2009)
26. Birenbaum, M., Kelly, A.E., Levi-Keren, M.: Stimulus features and sex differences in mental rotation test performance. Intelligence 19, 51–64 (1994)
27. Geiser, C., Lehmann, W., Corth, M., Eid, M.: Quantitative and qualitative change in children's mental rotation performance. Learning and Individual Differences 18, 419–429 (2007)
28. Sims, V.K., Mayer, R.E.: Domain specificity of spatial expertise: The case of video game players. Applied Cognitive Psychology 16, 97–115 (2002)
29. Titze, C., Jansen, P., Heil, M.: Mental rotation performance in fourth graders: No effect of gender beliefs (yet)? Learning and Individual Differences 20, 459–463 (2010)

30. Terlecki, M.S., Newcombe, N.S.: How important is the digital divide? The relation of computer and videogame usage to gender differences in mental rotation ability. Sex Role 53, 433–441 (2005)
31. Quaiser-Pohl, C., Geiser, C., Lehmann, W.: The relationship between computer-game preference, gender, and mental-rotation ability. Personality and Individual Differences 40, 609–619 (2006)
32. Collins, D.W., Kimura, D.: A large sex difference on a two-dimensional mental rotation task. Behavioral Neuroscience 111, 845–849 (1997)
33. Hyun, J.-S., Luck, S.J.: Visual working memory as substrate for mental rotation. Psychonomic Bulletin & Review 14, 154–158 (2007)
34. Loring-Meier, S., Halpern, D.F.: Sex differences in visuospatial working memory: components of cognitive processing. Psychonomic Bulletin & Review 6, 464–471 (1999)
35. Kaufman, S.B.: Sex differences in mental rotation and spatial visualization ability: Can they be accounted for by differences in working memory capacity? Intelligence 35, 211–223 (2007)
36. Schmader, T., Johns, M., Forbes, C.: An integrated process model of stereotype threat effects on performance. Psychological Review 115, 336–356 (2008)
37. Wraga, M., Helt, M., Jacobs, E., Sullivan, K.: Neural basis of stereotype-induced shifts in women's mental rotation performance. Social Cognitive and Affective Neuroscience 2, 12–19 (2006)

Towards a Revision of the Typology of Motion Verbs

Sander Lestrade and Nina Reshöft

University of Bremen

Abstract. Path and manner of movement are generally taken to be the core distinguishing features in descriptions of motion events. It is proposed here that this two-dimensional characterization of the lexical semantics of motion verbs needs to be reconsidered. We introduce a new research method, the cross-linguistic dictionary-lookup analysis, by means of which additional meaning dimensions can be identified for motion verbs in Dutch, English and German.

1 Introduction

This paper studies the way in which spatial and non-spatial semantic dimensions may be conflated in motion verbs across languages. According to Talmy [11, 12, 13, 14], a motion event can be described in terms of an object, the *Figure*, which moves with respect to a reference object, the *Ground*. The *Path* describes the course followed by the Figure with respect to the Ground. In addition, a motion event may have a *Manner* or a *Cause*, which Talmy [12, p. 61] analyzes "as constituting a distinct external event."

Talmy claims that these ingredients are expressed separately, and classifies languages into two broad types [12, p. 57]. In this binary typology, which is based on the locus of path, *satellite-framed languages* typically encode path in so-called *satellites* (e.g. particles), using the main verb to specify Manner or Cause. English is of this type, as illustrated in (1).

(1)	*The bottle*	*floated*	*out of*	*the cave*
	Figure	Motion+Manner	Path	Ground

By contrast, *verb-framed languages* such as Spanish typically encode the Path of the movement in the main verb and optionally specify Manner in satellites, for example coverbs or adverbials:

(2)	*La botella*	*salió*	*de la cueva*	*(flotando)*
	Figure	Motion+Path	Ground	Manner
	'The bottle floated out of the cave'			

Although Talmy [12, p. 57] points out that the distribution of labor may not always be this clear and additional semantic ingredients may be found in single elements of motion expressions, subsequent cross-linguistic research on lexicalization patterns of motion events generally has made use of this simplistic dichotomy only. This (implicitly) assumes that Manner and Path are the only two

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 230–241, 2012.

major dimensions that characterize cross-linguistic differences in the (verbal) expression of motion, because of which the two categories have become rather broadly understood notions that gloss over many potentially interesting meaning aspects. Indeed, according to Slobin ([10, fn. 5]; cf. also [7, 2]), manner "covers an ill-defined set of dimensions that modulate motion, including motor pattern, rate, rhythm, posture, affect, and evaluative factors."

Only recently, it has been proposed to look at other dimensions and combinatorial possibilities too. For example, a third type of *equipollently-framed languages* is sometimes added in which "[p]ath and manner are expressed by equivalent grammatical forms" ([10, p. 25]; cf. also [15, 3]). In these languages, Manner and Path are either both expressed by different verbs in a serial verb construction, simultaneously expressed in a single verb, or both expressed by verbal prefixes, as for example in Spanish *callejear* 'walking around streets' [2, p. 121]. Also, whereas motion verbs like *swim* and *fly* have mostly been analyzed as manner verbs, Frawley [4, p. 174-175] notes that they probably do not so much encode the manner of motion but rather the *location* at or *medium* in which the motion event takes place (water and air, respectively).

This paper too subscribes to a more in-depth analysis of motion verbs and tries to find out which additional dimensions can be identified in motion verbs in a more systematic way, using a *cross-linguistic dictionary-lookup method* (Cf. [6] for a more elaborate introduction of this method).

2 The Cross-Linguistic Dictionary-Lookup Methodology

To identify the spatial and non-spatial semantic dimensions of motion verbs, we annotated the dictionary definitions of a large set of motion verbs for English, Dutch and German. This selection of languages was simply motivated by the fact that the authors were most familiar with them. (But since all three happen to be traditionally analyzed as satellite-framed, Spanish is currently being added for contrast.) First, we simply went through dictionaries of the different languages (viz. *The American Heritage Dictionary of the English Language* and the *Deutsches Wörterbuch*, both available online at www.thefreedictionary.com, and the *van Dale Groot woordenboek der Nederlandse taal*, electronic version 1.2), and copied all potential motion verbs with their (relevant) definitions. If distinct relevant definitions of a motion verb were given, the entry was doubled in our list. For example, *to veer* occurs twice in our list, once as 'to turn aside from a course, direction, or purpose; swerve' and once as 'to shift clockwise in direction'.

Next, we narrowed down our list of motion verbs by selecting only those verbs which inherently, that is, by dictionary definition, describe *translational motion*, i.e., motion in which "an object's base location shifts from one point to another in space" [14, p. 35]. Specifically, we ignore change-of-posture verbs such as *bow* and *kneel*. Also, we included only the 'to move from place to place' use of *float*, excluding 'to remain suspended within or on the surface of a fluid without sinking'. Although the latter can be used in motion expressions, describing for example a boat floating toward a waterfall, the motion in this event, at least according to our procedure, is not due to the lexical semantics of the verb, but to (other elements of) the construction instead.

We further selected for verbs for which (i) the Figure of the motion event is the syntactic subject of the verb (thus including some transitive verbs and uses, such as *enter*), and (ii) parts, in case of composite verbs, have no or a different independent meaning (i.e. the meaning is non-compositional; e.g. we excluded Dutch *wegrijden* (lit.: 'away'-'drive') 'drive away', but included German *zurücksetzen* 'drive back', which is used for (people moving in) vehicles only and can be used intransitively, whereas *setzen* 'put' is more general and can be used transitively only). By the first criterion, we leave for future work a further distinction that can be made between caused motion (largely expressed by transitive verbs) and locomotion (expressed by intransitives; contrast Dutch *vellen* 'to make fall' and *vallen* 'to fall').

As a third step, we classified and annotated the definitions along the following dimensions, which will be discussed in more detail below:

- Path: specification of the internal organization of the course of motion (form, length, orientation) without reference to a Ground
- Ground: specification of the relation between Figure and (some region with respect to) the reference object
- Figure: type of Figure that the motion verb selects for or, in case of complex Figures, specification of the organization of its parts (e.g. a horse for *gallop*, or 'in drops, in a stream' in *swarm*)
- Manner: way in which specific parts of the body move or are positioned (for animates) or way in which the Figure behaves during the motion event
- Means: additional means used in the motion event (such as vehicles or instruments)
- Speed: velocity with which motion takes place
- CausalStructure: purpose, motivation, cause of motion, result
- Time: temporal specification of motion event, such as pattern, duration, frequency (e.g. *(ir)regularly, suddenly, for a short time*)
- Sound: sound caused by motion event
- Context: specification of context in which the motion event takes place (e.g. 'usually to music' in one of the definitions of *dance*)

As we are still developing our coding scheme and adding more languages to our data set, these dimensions and their definitions should be considered provisional. Nevertheless, we believe the results we present here are reliable, as evidenced by our interannotator-agreement score of K=.86, which was determined on the basis of a random subset of 100 English verb definitions annotated by both authors.[1] Also, we have purposefully restricted our claims in this paper to the current state of affairs.

[1] In total 276 categories were used for the 100 verbs in this test. The kappa coefficient measures pairwise interannotator agreement correcting for change agreement. Generally, scores above .8 are considered reliable (cf. [1]).

We followed the general corpus procedure by determining the annotation on the meaning in the context of the definition only. Hence, *directly* in English *cut* ‘to go directly and often hastily’ is about the form of the course of motion (Path), not about some immediate performance (Time) it might express in (most) other cases. The same holds for definitions that include a descripiton of Speed. For example, depending on the context *quick(ly)* should be interpreted as either Speed or Manner. In *patter*, ‘to move with quick, light, softly audible steps’, it seems to decribe Manner, whereas in *buzz* ‘to move quickly and busily’, it is about the velocity with which the movement takes place (Speed).

Sometimes, the distinction between two dimensions is hard to determine. For example, in the definition of *rebound*, ‘to spring or bounce back after hitting or colliding with something’, we have annotated *after hitting or colliding with something* as Context, although one may argue that it could be CausalStructure too. Similarly, in *parade*, ‘to march in a public procession’ one could argue for both Context and Ground. Probably, this situation can never be solved completely, but we keep trying to refine our definitions in order to reduce such ambiguities as much as possible.

As a criterion for the identification of the different motion dimension we used their independent manipulability. If two dimensions are independently modifiable, they should be kept separately. Although not fully developed yet, we believe that our classification will thus eventually lead to a better structured semantic organization of motion verbs and events.

Note that, as a result of this procedure, what has come to be analyzed as Path in the framework of Talmy is partly subsumed under Ground in our analysis, reserving Path for an arguably more intuitive use. *Path*, in our analysis, exclusively refers to the lexically specified properties of the course of motion, for example its form (‘zigzag’) or inherent orientation (that is, using previous points of the motion path only, e.g. ‘downward’). Instead, *Ground* refers to the specification of the relation between the Figure and (a location with respect to) a reference object. Generally, the role or type of this reference object is further specified too. For example, in *enter* ‘to come or go into’ (the inside of) the GroundRole is the Goal of the motion event, in *leave* ‘to go out of or away from’ the Ground functions as a Source, and in *swim* ‘to move through water by means of the limbs, fins, or tail’ the GroundRole is a Medium and the GroundType is specified as ‘water’. We will also analyze as a Ground those specifications of the orientation of the Path that use a reference object. For example, *toward a dance partner* in English *balance* ‘to move toward and then away from a dance partner’ is annotated as a Ground even though it specifies orientation too. The crucial difference between both types of orientations (i.e., those annotated as Paths and those annotated as Goals) is that the former type does not specify the development of the relation between Figure and Ground but rather uses the preceding position of the Figure itself. Instead, orientations toward Grounds by definition make use of a reference object. We will leave for future research a further analysis of these different kinds of Ground specification in terms of its role (Goal, Source, . . .) or type (water, air, . . .). Our main point for now is that Path and Ground could be considered independent dimensions, as the Path dimension can be specified without the use of Grounds, an option that is very frequently chosen indeed, as Section 3 will show.[2] For example, it does not matter whether one goes into a house (Goal) via a downward, long, or circular path (Path).

The dimensions of Manner, Speed and Time may seem difficult to keep apart at first, but here too, the intuition should be clear. Indepently from how one moves his limbs, e.g. whether one is crawling or hopping (Manner), the translational movement can take place quickly or slowly (Speed) and, again independent from that, be regular or sudden (Time). Since Speed is about the velocity of the translational motion and Time is about temporal organization, we classify *whirling motion* in *reel* 'to go round and round in a whirling motion' as Time, not as Speed.

The following examples may further illustrate the annotation procedure (note that we, at least for now, ignore passives in our annotation):

(3) *surge*: 'to roll or be tossed about on waves, as a boat'

a. TypeOf=roll
b. Path=about
c. Ground=on waves
d. Figure=as a boat

(4) *patter*: 'to move with quick, light, softly audible steps'

a. TypeOf=move
b. Manner=with quick, light steps
c. Sound=with softly audible steps

TypeOfs, such as *roll* in (3) are the motion verbs that are used and further modified in the definition of the motion verb. We will say more about them below.

In (4), *quick steps* could both be argued to belong to Manner (modifying the way in which the body parts move) and Time (being about the temporal pattern in terms of frequency of repetition), since the two are related. In such cases, we tried to determine which of the two is primary and which is dependent, or at least less important, annotating for the former only. In the same example, *light steps* clearly is about the way in which the feet are moved and hence easily assigned to the Manner dimension. Also the Sound classification of *softly audible* should not be problematic.

The difference between Path and Ground is further illustrated by the following examples:

(5) *plunge*: 'to move forward and downward violently'

a. TypeOf: move
b. Path: forward and downward
c. Manner: violently

[2] This also holds the other way around, which cannot be shown empirically as a result of our choice not to discuss the role of the Ground here. However, we would argue that one can, for example, *enter* ('to come or go into') a Ground via many different Paths and that verbs thus may specify a Ground without specifying the Path of motion.

Forward and *downward* in (5) concern the internal organization of the motion path. They are defined with respect to preceding points of the path, without referring to a Ground. The same holds for *upward* in (6). In this case, however, the Ground is specified too (having a Source role).

(6) *leap*: 'to spring or bound upward from or as if from the ground; jump'
  a. TypeOf: spring; bound
  b. Path: upward
  c. Ground: (as if) from the ground
  d. Synonym: jump

Similarly, *pace* in (7) has both a Path and a Ground specification. The Path has a *back and forth* form or orientation and is located *across* some reference object.

(7) *pace*: 'to walk or stride back and forth across'
  a. TypeOf: walk; stride
  b. Path: back and forth
  c. Ground: across

Motion verbs may have several (relevant) dictionary entries, which we have numbered in our data set distinctly. As illustrated for Dutch *joggen* 'to jog' in (8), some of these entries are characterized as being specific variants of some other motion verb.

(8) Dutch *joggen*: 'Hardlopen als ontspanning, louter ten behoeve van de lichamelijke conditie' (*jog* 'Run for recreation, for physical condition only')
  a. TypeOf=hardlopen (run)
  b. CausalStructure=als ontspanning, louter ten behoeve van de lichamelijke conditie (for recreation, for physical condition only)

In fact, a single entry may contain several TypeOfs, as shown in some of the above examples already. When multiple TypeOfs are used within one definition, the entry was multiplied accordingly. We then automatically enriched each copy with semantic information from one TypeOfs only.

However, *TypeOfs* are often themselves further specified for various dimensions and therefore may have their own TypeOf, as shown in (9).

(9) Dutch *hardlopen*: 'Snel en lang achtereen lopen (als oefening of als vertoning)' (*run* 'Walk quickly and for a long period (as an exercise or display)')
  a. TypeOf=lopen (walk)
  b. Speed=snel (quickly)
  c. Rhythm=lang achtereen (for a long period)
  d. CausalStructure=(als oefening of als vertoning)((as an exercise or display))

Thus, *joggen* in (8) is in fact indirectly lexically specified for the dimensions of *hardlopen* in (9), which on its turn is indirectly specified for whatever is said about its TypeOf *lopen* ‘walk’ (9-a). When TypeOfs themselves have multiple entries (and hence possible referents) in our dataset, we have manually specified the relevant entry that is used in the definition. In this case, the TypeOf that *hardlopen* makes use of seems to be the sixth entry of *lopen*, viz. (*zich op de benen snel voortbewegen* ‘move quickly on the legs’).

A small number of motion verbs, such as English *go* and *move*, Dutch *bewegen* and *gaan* and German *sich bewegen*, is very frequently used as TypeOfs. In our analysis, we have considered them as motion primitives and did not annotate them to prevent their dimensions from becoming overrepresented through the feeding procedure.

Before going to the results, probably a brief comparison between our procedure and the more commonly used method of introspection is in place. The advantage of our method with respect to introspection is that the latter is much less objective. If one wanted to show the importance of some dimension, say the type of Ground, it is fairly easy to come up with a list of verbs in which this dimension indeed can be identified or even be argued to be of defining importance. For example, what seems to be of importance for prototypical manner verbs such as *swim* and *float* in fact is not so much the way in which some agent moves, the Manner, but rather the substance in which the motion event is conceptualized, i.e. the GroundType: Dogs and humans swim in radically different manners, but still both are said to *swim* when they move through water. Unfortunately, introspection is a rather self-fulfilling method as one is likely to be much more sensitive to data that confirm the research hypothesis. Instead, verb definitions in dictionaries are established independently from our research goals and hence offer a much more objective data set.

In such a comparison, it is important to evaluate our method by the research question it addresses, as putative shortcomings are dependent on one’s research goals. As every corpus linguist will know, actual word uses as found in a corpus may differ from the ones given in a dictionary. However, our goal is *not* to give a semantic characterization of individual motion verbs nor to achieve full coverage of the complete set of motion verbs of the languages. Instead, our aim is to sketch the range of dimensions that typically occur in the class of motion verbs of a specific language. For this, a corpus study is probably infeasible.

Nevertheless, also for specific verbs our method may in fact be quite rewarding. Consider the uses of otherwise problematic *cluster concepts* such as *climb*:

(10) a. Bill climbed (up) the mountain.
b. Bill climbed down the mountain.
c. The snake climbed (up) the tree
d. ?*The snake climbed down the street.

According to Jackendoff [5, p. 353, following Fillmore 1982], *climbing* covers independent conceptual conditions: first, an individual is traveling upward, and second, the individual is moving with characteristic effortful grasping motions. In his examples, (10-a) violates neither, (10-b) and (10-c) each violate one of them, and (10-d) violates both. Using introspection (at least when done for all motion verbs), one may easily miss out on either of these dimensions, or think they are both necessary. Instead, as shown in (11), our dictionary entry of *climb* "correctly" gives the following definitions, thereby covering Jackendoffs intuitions:

(11) *climb*

a. 'To move oneself upward'
b. 'To rise slowly, steadily, or effortfully; ascend'
c. 'To move in a specified direction by using the hands and feet'

3 Results

In total, we analyzed more than 1000 motion verb definitions for our three languages: 316 for Dutch, 484 for English, and 326 for German. These definitions correspond to 197 unique verbs in Dutch, 303 in English, and 225 in German. Probably, numbers for German and Dutch are lower because of the compositional nature of many motion verbs in these languages (which were then excluded, as explained in the previous section).

In Table 1, the proportion of verbs that is specified for each meaning dimension is given (with proportions before feeding TypeOf information between parentheses). Note that proportions are not additive, as a single verb may be specified for a number of dimensions.

Table 1. Proportions of verbs specified for specific dimension after feeding (proportions before feeding in parentheses)

dimension	Dutch	German	English
Ground	0.51 (0.43)	0.60 (0.47)	0.37 (0.30)
Path	0.58 (0.36)	0.51 (0.24)	0.44 (0.32)
Manner	0.41 (0.37)	0.41 (0.20)	0.52 (0.44)
Speed	0.18 (0.14)	0.32 (0.15)	0.17 (0.13)
Figure	0.17 (0.14)	0.19 (0.10)	0.05 (0.05)
Context	0.17 (0.10)	0.07 (0.06)	0.08 (0.07)
CausalStructure	0.16 (0.16)	0.10 (0.08)	0.11 (0.07)
Time	0.11 (0.10)	0.17 (0.11)	0.18 (0.13)
Means	0.07 (0.05)	0.11 (0.07)	0.06 (0.06)
Sound	0.04 (0.04)	0.04 (0.03)	0.06 (0.05)

Before feeding TypeOf information, the main dimensions that can be identified via our dictionary-lookup method are, indeed, Ground, Manner and Path, which together could be said to cover Talmy's Manner and Path (cf. Section 4 below). Of all possible correlations between dimensions within languages, we only find negative correlations exceeding the arbitrarily set threshold of .2 between Manner and Ground in Dutch (R = -.34; Pearson's χ^2 = 15.07, p=.0001), between

Path and Ground in German (R= -.025, $\chi^2 = 18.7882$, df = 1, p-value = 1.461e-05), and between Manner and Ground in English (R=-.34, $\chi^2 = 54.8457$, df = 1, p-value = 1.304e-13; all significant after Bonferroni correction; for these and all other calculations and manipulations, R was used [8]). These negative correlations nicely suggest that lexical specification of Manner, Path, and Ground exclude each other (but see below).

After feeding TypeOf information and multiplying entries for their number of TypeOfs as explained in the previous section, we get 372 definitions for Dutch, 606 for English, and 367 entries for German. Not surprisingly, after feeding TypeOf information the proportions of verbs that are specified for some specific dimension also increase. In table 2 it is shown how many dimensions are specified per verb. Zero-dimensional verbs are either motion primitives such as *move* (see previous section) or verbs that are defined by means of synonyms only (the information of which we did not feed in), e.g. *journey.* Some examples of one-dimensional verbs in English are *creep, cycle, wobble,* and *wheel.*

Table 2. Number of verbs with certain number of specified dimensions

	number of dimensions							
Language	0	1	2	3	4	5	6	7
Dutch	27	60	124	91	40	22	6	1
English	15	206	227	96	37	19	2	1
German	31	69	88	82	67	27	6	0

Recall that before feeding TypeOf information, we only found negative correlations between Manner and Ground for English and Dutch and between Path and Ground for German. If we again look into correlations between dimensions after including TypeOf information, we find a very different pattern as shown in Table 3.[3] Whereas Manner and Ground still exclude each other in English, we instead find a number of positive correlations in Dutch and German. The negative correlation between Manner and Figure thas is found in German is not significant after Bonferroni correction. The same holds for the correlation between Means and Figure in Dutch and Means and Path in German; all other correlations are significant.

4 Discussion

Preliminary as they are, the numbers in Table 1 already show that the simple two-way typology of motion verbs does not hold. Our three Germanic languages are all traditionally classified as satellite-framed, hence (what we analyze as) Ground and Path information should not occur in the dictionary definitions as this information is generally subsumed under the notion of Ground in the simple view. Given Talmy's characterization of satellite-framed languages, which are

[3] In the tables below, correlations below .2 are not reported and replaced by zero.

Table 3. Correlations between dimensions specified in motion verbs exceeding the arbitray threshold of R=.2. Correlations that are non-significant after Bonferroni correction are marked with a “*”.

ENGLISH	Ma	P	Me	F	G	Sp	So	CS	T	Co
Manner	1									
Path	0	1								
Means	0	0	1							
Figure	0	0	0	1						
Ground	-0.2	0	0	0	1					
Speed	0	0	0	0	0	1				
Sound	0	0	0	0	0	0	1			
CausalStructure	0	0	0	0	0	0	0	1		
Time	0	0	0	0	0	0	0	0	1	
Context	0	0	0	0	0	0	0	0	0	1
DUTCH	Ma	P	Me	F	G	Sp	So	CS	T	Co
Manner	1									
Path	0.22	1								
Means	0	0	1							
Figure	0	0	0.21*	1						
Ground	0	0	0	0	1					
Speed	0	0	0	0	0	1				
Sound	0	0	0	0	0	0	1			
CausalStructure	0	0	0	0	0	0	0	1		
Time	0	0	0	0	0	0	0	0	1	
Context	0	0	0	0	0.30	0	0	0	0	1
GERMAN	Ma	P	Me	F	G	Sp	So	CS	T	Co
Manner	1									
Path	0.27	1								
Means	0	0.21*	1							
Figure	-0.27*	0	0	1						
Ground	0	0	0	0	1					
Speed	0.46	0.34	0	0	0	1				
Sound	0	0	0	0	0	0	1			
CausalStructure	0	0	0	0	0	0	0	1		
Time	0	0	0	0.34	0	0	0	0	1	
Context	0	0	0	0	0	0	0	0	0	1

said to prefer manner verbs over path verbs, our results provide counterevidence for Dutch, English and German. Instead, Ground, Manner and Path seem to be equally important dimensions. This counterevidence becomes even stronger when our Ground and Path dimensions are taken together and contrasted with Manner for a more direct comparison with the standard view.

Especially after feeding, our results show that motion verbs are much more complex than the traditional dichotomy suggests. However, at the current stage it cannot be said in which ways these dimensions relate to typological claims about Path and Manner encoding. Several issues need to be considered in order to make stronger claims about methodological profits, interactions between dimensions, and cross-linguistic differences in the semantics of motion verbs. First, the validity of cross-linguistic dictionary lookup method is an important aspect in this regard. Problems like definitional practices as well as the lexicographical treatment of polysemy should be considered. Second, a proper interpretation of the correlations (within and between languages) needs to await a more careful inclusion of TypeOf information, which, at present, is not completely without errors. Another important question is whether additional dimensions are needed. Some of the dimensions we identified might be broken further down into subdimensions, which, as further analysis need to show, may have to be considered dimensions of their own. For a cross-linguistic comparison, it is also important to look at semantic dimensions of motion verbs in "verb-framed" languages. At present, we are adding information from Spanish indeed, which is the main example of a verb-framed language.

5 Conclusion and Future Research

Looking at the dictionary definitions of a large sample of motion verbs in three languages, we have shown that the traditional simple view of motion verbs as expressing either Manner or Path does not hold. Our findings suggest that motion verbs are far more complex and cannot be analyzed in terms of Path and Manner only.

References

1. Carletta, J.: Assessing agreement on classification tasks: the Kappa statistic. Computational Linguistics 22, 249–254 (1996)
2. Cifuentes Férez, P.: Human locomotion verbs in English and Spanish. International Journal of English Studies 7(1), 117–136 (2007)
3. Croft, W., Barddal, J., Hollmann, W., Sotirova, V., Taoka, C.: Revising Talmy's typological classification of complex events. In: Boas, H. (ed.) Contrastive Construction Grammar, pp. 201–235. John Benjamins, Amsterdam (2008)
4. Frawley, W.: Linguistic Semantics. Lawrence Erlbaum, Hillsdale (1992)
5. Jackendoff, R.: Foundations of language. Brain, meaning, grammar, evolution. Oxford University Press, Oxford (2002)
6. Lestrade, S., Reshöft, N.: Extending the typological toolbox with the dictionary-look-up method: A case study of the cross-linguistic encoding of motion events (submitted)
7. Pourcel, S.: Motion in language and cognition. In: Soares da Silva, A., Torres, A., Gonçalves, M. (eds.) Linguagem, cultura e cognio: estudos de linguistica cognitiva, vol. 2, pp. 75–91. Almedina, Coimbra (2004)
8. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2009), http://www.R-project.org
9. Slobin, D.: Verbalized events: a dynamic approach to linguistic relativity and determinism. In: Niemeier, S., Dirven, R. (eds.) Evidence for Linguistic Relativity, pp. 107–138. John Benjamins, Amsterdam (2000)

10. Slobin, D.: The many ways to search for a frog: linguistic typology and the expression of motion events. In: Strömqvist, S., Verhoeven, L. (eds.) Relating Events in Narrative: Typological Perspectives, pp. 219–257. Lawrence Erlbaum Associates, Mahwah (2004)
11. Talmy, L.: Semantic structures in English and Atsugewi Doctoral dissertation. University of California, Berkeley
12. Talmy, L.: Lexicalization patterns: semantic structure in lexical forms. In: Shopen, T. (ed.) Language Typology & Syntactic Description, vol. 3, pp. 57–149. Cambridge University Press, Cambridge (1985)
13. Talmy, L.: Path to realization: a typology of event conflation. In: Proceedings of the Seventeenth Annual Meeting of the Berkeley Linguistics Society, pp. 480–520 (1991)
14. Talmy, L.: Toward a Cognitive Semantics, vol. 1. MIT Press, Cambridge (2000)
15. Zlatev, J., Yangklang, P.: A third way to travel: The place of Thai and serial verb languages in motion event typology. In: Strömqvist, S., Verhoeven, L. (eds.) Relating Events in Narrative: Typological and Contextual Perspectives. Lawrence Erlbaum Associates, Mahwah (2003)

Assessing Similarities of Qualitative Spatio-temporal Relations[*]

Alexander Klippel[1], Jinlong Yang[1], Jan Oliver Wallgrün[1], Frank Dylla[2], and Rui Li[1]

[1] Department of Geography, GeoVISTA Center, The Pennsylvania State University, PA, USA
{klippel,jinlong,wallgrun,rui.li}@psu.edu
[2] SFB/TR 8 Spatial Cognition, Universität Bremen, Germany
dylla@sfbtr8.uni-bremen.de

Abstract. In this article we analyze behavioral data to advance knowledge on how to assess similarities of events and spatial relations characterized by qualitative spatial calculi. We have collected a large amount of behavioral data evaluating topological relations specified in the Region Connection Calculus and Intersection Models. Several suggestions have been made in the literature on how to use associated conceptual neighborhood graphs to assess the similarities between events and static spatial relations specified within these frameworks. However, to the best of our knowledge, there are few (to none) approaches that use behavioral data to formally assess similarities. This article is contributing to this endeavor of using behavioral data as a basis for similarities (and associated weights) by (a) discussing a number of approaches that allow for transforming behavioral data into numeric values; (b) applying these approaches to nine data sets we collected in the last couple of years on conceptualizing spatio-temporal information using RCC/IM as a baseline; and (c) discussing potential weighting schemes but also revealing essential avenues for future research.

1 Introduction and Background

Every calculus with jointly exhaustive and pairwise disjoint (JEPD) relations (such as RCC and IM) has a conceptual neighborhood graph (Cohn & Renz, 2008).

To navigate through daily life, humans use their ability to conceptualize spatio-temporal information, which ultimately leads to a system of categories. Likewise, the disciplines of the spatial sciences focus on conceptualization and categorization as a means to structure spatio-temporal information. Although challenged by several researchers, similarity is one of the most important and most commonly used tools to aid in the process of conceptualization and categorization in both artificial and natural cognitive systems (Bruns & Egenhofer, 1996; Goldstone & Son, 2005; Nedas & Egenhofer, 2008; Rissland, 2006; Schwering, 2008; Tversky, 1977). In the spatial sciences and in related branches of artificial intelligence, an approach has been developed that allows the

[*] This research is funded by the National Science Foundation (#0924534) and Deutsche Forschungsgemeinschaft (DFG) grant SFB/TR 8 Spatial Cognition.

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 242–261, 2012.

specification of similarity measures for spatio-temporal data: qualitative spatio-temporal representation and reasoning (QSTR). Calculi developed in the general area of QSTR allow for meaningful processing of spatio-temporal information because they focus on categorical (discrete) changes or salient discontinuities (Galton, 2000), which are thought to be relevant to an information processing system (both human and artificial). While qualitative calculi are naturally appealing and are, on a general level, widely acknowledged in the cognitive sciences, too[1], there is comparatively little behavioral assessment of the cognitive adequacy of these calculi. This is an astonishing fact given that these calculi are often intended to improve processes at the human-machine interface and are on several occasions claimed to be cognitively adequate (Clementini, Di Felice, & van Oosterom, 1993; Knauff, Rauh, & Renz, 1997; Knauff, 1999). However, in our opinion, the systematic behavioral evaluation of QSTR is an essential missing piece that will lead to refined and improved QSTR models and in return significantly increase their value and their usability in numerous applications (e.g., information retrieval, spatial query languages, formalizing the semantics of spatial language).

This paper will discuss a framework for defining cognitively adequate similarities/weights by detailing strategies to transform results from behavioral experiments on how humans conceptualize spatio-temporal information into both qualitative (category-based) and quantitative similarity measures. These measures are tailored towards formal theories in the spatial sciences and will be applicable to theories of spatio-temporal representation and reasoning. Hence, we will contribute to the formal basis of the semantics of spatio-temporal information.

To further motivate the general questions we are addressing in this paper, consider the spatial scenes in the right part of Figure 1. The scenes show the development of an oil spill in relation to an island. We focus on relations distinguished by prominent topological calculi commonly used in spatial information theory and in the cognitive sciences to identify potentially important aspects of spatio-temporal information. The individual icons reflect distinctions made by the Region Connection Calculus (RCC, Randell, Cui, & Cohn, 1992) as well as Egenhofer's Intersection Models (IM, Egenhofer & Franzosa, 1991). An important aspect for assessing the similarity of these scenes as well as modeling spatio-temporal information is that these relations can be organized to form a so-called conceptual neighborhood graph (CNG, Freksa, 1992, left part of Figure 1). Two relations, R_1 and R_2 are conceptual neighbors if it is possible for R_1 to hold over a tuple of objects at a certain point in time, and for R_2 to hold over the tuple at a later time, with no other (third) mutually exclusive relation holding in between (Cohn, 2008). A neighborhood graph has one node for each relation $R \in \boldsymbol{R}$, and an edge between two nodes if the corresponding relations are conceptual neighbors. The important aspect to keep in mind, which adds to the transformative nature of this paper, is that virtually every calculus with jointly exhaustive and pairwise disjoint (JEPD) relations (such as RCC and IM) has a conceptual neighborhood graph (Cohn & Renz, 2008), and that, hence, the methods proposed here are universally applicable amongst all such calculi.

[1] Lewin, 1936/1966; Piaget & Inhelder, 1948/56/67; Lu, Harter, & Graesser, 2009.

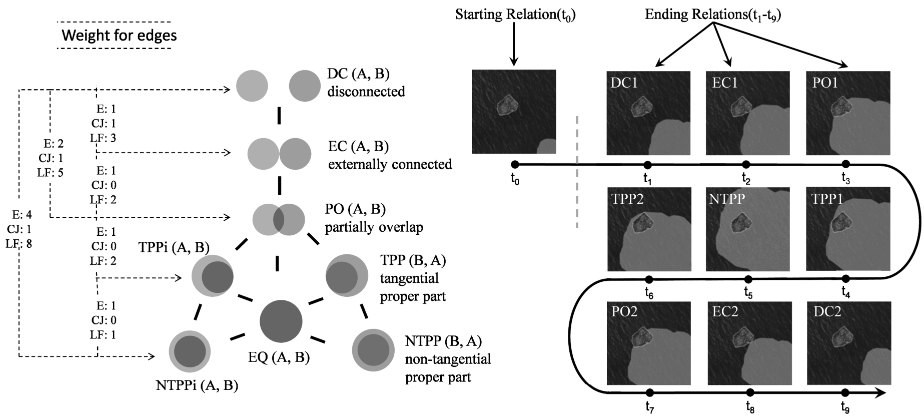


Fig. 1. The left side shows a conceptual neighborhood graph based on RCC-8 and IM. The dotted lines reflect the discussion in the text showing similarity assessments for three models exemplarily (E: equal weights; CJ: Camara and Jungert, 2007; LF: Li and Fonseca, 2006). On the right side the development of an oil spill is depicted in relation to an island. Each of the scenes could either be a transition or an ending relation.

One of the main characteristics of the relations displayed in Figure 1, but also relations from all other JEPD qualitative calculi, is that we can measure how similar the scenes (or their genesis) on the right side are by employing the conceptual neighborhood graph on the left side. Organizing spatial relations in this graph-like format has the advantage that graph theoretical measures can be applied to determine the similarity between these scenes which is essential for numerous information retrieval and formal semantic tasks (Bruns & Egenhofer, 1996; Papadias & Delis, 1997; Wallgrün, Wolter, & Richter, 2010). As both the spatial and the cognitive sciences focus on qualitative distinctions made, for example, by topology (Galton, 2000; Johnson, 1987; Klix, 1971), we have found—theoretically—a bridge between formal and cognitive spatial semantics. To demonstrate this aspect, we will focus on the similarity of four relations in Figure 1, DC (t1), EC (t2), PO (t3), and NTPP (t5): The simplest approach (Bruns & Egenhofer, 1996; Dylla & Wallgrün, 2007; Rada, Mili, Bicknell, & Blettner, 1989; Schwering, 2007), in a nutshell, assigns all edges in the CNG an equal weight of 1 and similarity/dissimilarity is established by counting the number of edges between two relations. Hence, the dissimilarity (weight) between DC and EC would be 1, the dissimilarity between DC and PO would be 2, and the dissimilarity between DC and NTPP would be 4.

This rather simplistic view has, of course, been challenged and several researchers have proposed (introspectively) alternative weighting schemes for CNGs. For example, Camara and Jungert (2007), in seeking to define a query language for dynamic processes, suggested a grouping of topological relations that are the basis of the CNG in Figure 1 into DC (disconnected) on the one hand and all other relations on the other. If we apply this strategy, dissimilarity between DC and all other relations would be 1, while the dissimilarity among all other relations would be basically 0. Another approach by Li and Fonseca (2006) takes into account that there may be different weights between different conceptual neighbors. They assign, for example, a weight of 3 to the edge between DC and EC, and a weight of 2 to the edge between EC and PO, a weight of 1 to the edge between TPP and NTPP. Hence the three dissimilarities from the example would be: DC-EC - 3, DC-PO - 5, and DC-NTPP - 8.

Many other approaches have been discussed. Consider the issue of the level of granularity which determines the number of basic relations that are assumed. The immediately relevant distinction here is between RCC-5 and RCC-8. While RCC-8 can be mapped onto eight relations distinguished by Egenhofer's intersection models (IM), RCC-5 cannot be directly mapped onto the coarser level of the IM (cf. Knauff et al., 1997; Renz, 2002) due to the ontologically different status of the boundaries. Hence, depending on the model we apply, the similarities will change, too (see also Clementini et al., 1993 for a different 5 relation solution, 4 plus 1 to be precise).

This simple example and associated literature demonstrate a number of important issues:

- Formal calculi using JEPD are omnipresent in research in the spatial sciences and are both theories as well as integral parts of spatial information systems. They are vital to various applications for spatial representation and reasoning, and are used frequently to establish similarities especially to aid human-computer interaction.
- There is some arbitrariness in designing and using these approaches guided by both formal constraints and requirements arising from a specific formalism (e.g., RCC versus IM) or by the introspection of a researcher. This lack of guidance as to which approach to use has been identified as a major obstacle in the usefulness of QSTR (Schultz, Amor, & Guesgen, 2011).
- There are few behavioral approaches that have evaluated QSTR[2]. However, to the best of our knowledge, except for our own work (Klippel, accepted), there is no behavioral research that addresses the possibility that similarities (as an expression of cognitive conceptualization processes) between qualitative spatial relations may change depending on the semantics of a specific domain[3]. For instance, one big question is: What happens to similarities of relations in the example in Figure 1, if we use different domains such as such as a lake and a house or a hurricane and a peninsula? Do we expect to be able to use the same similarities (weights) between relations?

We strongly believe that similarity measures should not be designed introspectively. As these measures are often intended to improve the interface of humans and computers/machines, it is essential to ground the assessment of similarities in behavioral research.

The remainder of this article is structured as follows: In Section 2, we will provide a very short overview of the behavioral data that we have collected over the last couple of years and that we will reanalyze here to discuss weights for CNGs; Section 3 discusses different methods that potentially allow for establishing weights; Section 4 provides a summary and lays out ideas for future research efforts.

[2] For an overview of research evaluating QSTR see Klippel, Li, Yang, Hardisty, & Xu, in press; Mark, 1999.

[3] Mark and Egenhofer 1995 have speculated that this might be the case.

- We will be using the following abbreviations: CNG: Conceptual neighborhood graph, CN: Conceptual neighbor; TEC: Topological equivalence class; OSM: Overall similarity matrix; QSTR: Qualitative Spatio-temporal Representation and Reasoning; RCC: region connections calculus; IM: Intersection models. We will also use RCC terminology for topological relations: DC: disconnected, EC: externally connected, PO: partial overlap, TPP: tangential proper part, NTPP: non-tangential proper part.

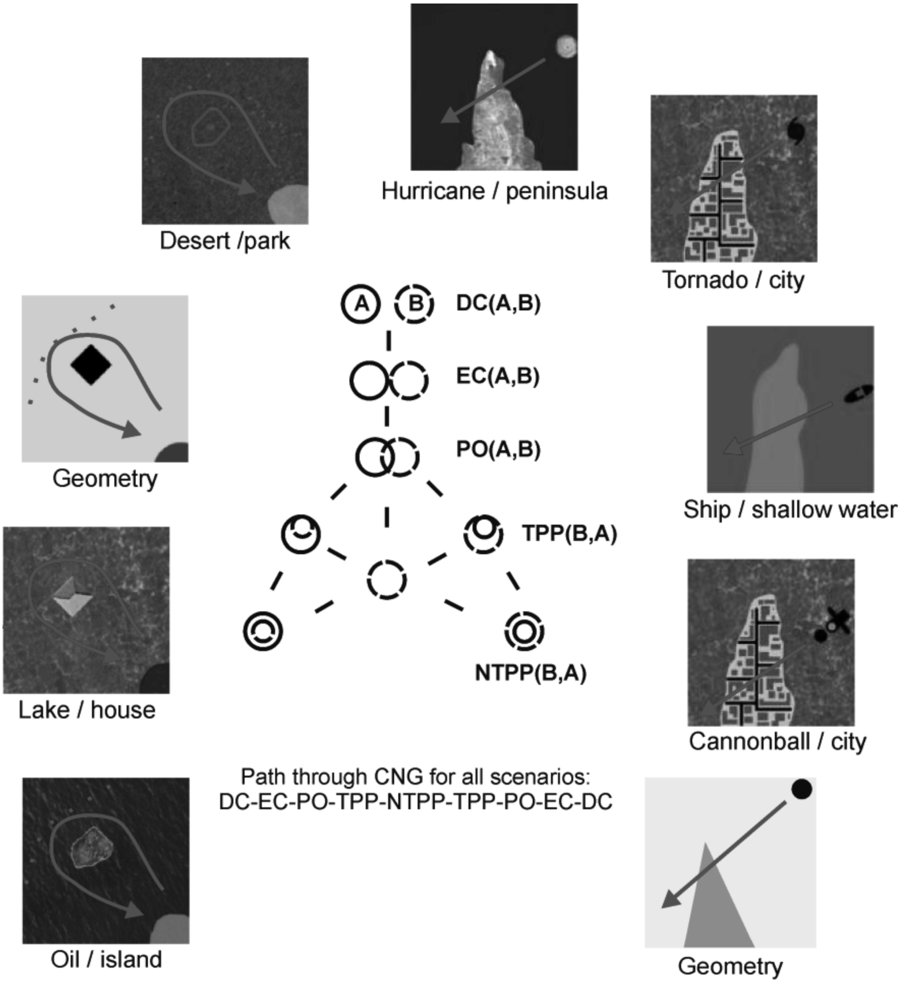


Fig. 2. Example of the nine scenarios that we reanalyze in this paper to derive weights for conceptual neighborhood graphs (translation: geometric figures, hurricane/peninsula, tornado/city, ship/shallow water, cannonball/city; scaling: geometric figures, oil spill/island, house/lake, desert/recreation park).

2 Data Collection

In this article, we reanalyze data we collected investigating cognitive conceptualizations of earth dynamics. We have designed nine experiments (for more details see Klippel, accepted and Yang, Klippel, & Li, submitted) using two different types of dynamics—movement patterns that can be considered translations and movement patterns characterized as scaling. Additionally, we used different semantic domains with different entities (translation: geometric figures, hurricane/peninsula, tornado/city, ship/shallow water, cannonball/city; scaling: geometric figures, oil spill/island, house/lake, desert/recreation park). Figure 2 gives a general idea of how the animated icons were designed for the different domains, while Figure 1 already showed static snapshots of the actual animations for the oil spill/island domain. The important aspect that makes all nine experiments comparable is that the stimuli used in each experiment are identical from a topological perspective (using either RCC-8 or IM). In each of the nine experiments, animations are designed such that nine different yet topologically equivalent movement patterns can be distinguished. The main distinguishing criterion is borrowed from cognitive theories on event conceptualization (Regier & Zheng, 2007), that is, patterns are separated based on the topological relation they can end in (see Figure 1). All movement patterns start in the DC relation and could end in one of the nine possible ending relations depicted in Figure 1. Within each topologically identical pattern we realized eight instances. This means that for each of the nine experiments (semantic domains), 72 animations were created: eight animations/instances each for nine topologically equivalent movement patterns.

In our experiments we employed a grouping paradigm, which is classically used to elicit conceptual knowledge. Participants (N = 20 in each of the nine experiments) have the task to sort the animated icons into groups with larger within-group than between-group similarities. The task can be characterized as *free classification* (Pothos, 2005) or *category construction* (Medin, Wattenmaker, & Hampson, 1987), meaning that participants created all their groups from scratch without any limitations regarding the number of groups or which icons should be placed together. All 72 animations have to be sorted into groups before the experiments were considered complete. The grouping behavior for each participant is recorded in a similarity matrix in binary form: two icons that are placed into the same group are coded as 1; two icons in different groups are coded as 0. All nine scenarios have 72 animated icons such that each matrix for each participant has 5194 cells of which 2556 are meaningful (others are redundant or encode the relation of an icon with itself). Summing over all individual similarity matrices within each domain nine overall similarity matrices (OSMs) are created. OSMs encode overall similarity assessment between icons in the following way: The highest possible similarity corresponds to N, the number of participants. For example, if all 20 participants placed a certain pair of animated icons together into a group, 20 individual '1's are added up. In contrast, if a pair of two animated icons is never placed into the same group, their similarity is recorded as '0'.

As all nine domains are topologically identical, we have obtained a large number of similarity ratings for pairs (conceptual neighbors) of topological relations. Each topological equivalence class had eight instances in all nine domains assessed by 20 participants in each experiment. Hence, we have a total of 8 times 9 times 20 (= 1440) similarity assessments for each topological relation combination (for CNs but also for all other possible combinations). To give the reader a first impression of how these similarities are distributed across the nine different domains and across TECs, Figure 3 is visualizing the raw similarities from the OSMs in so-called heat maps. We reduced the size but the overall patterns reveal that there are potentially interesting differences across domains.

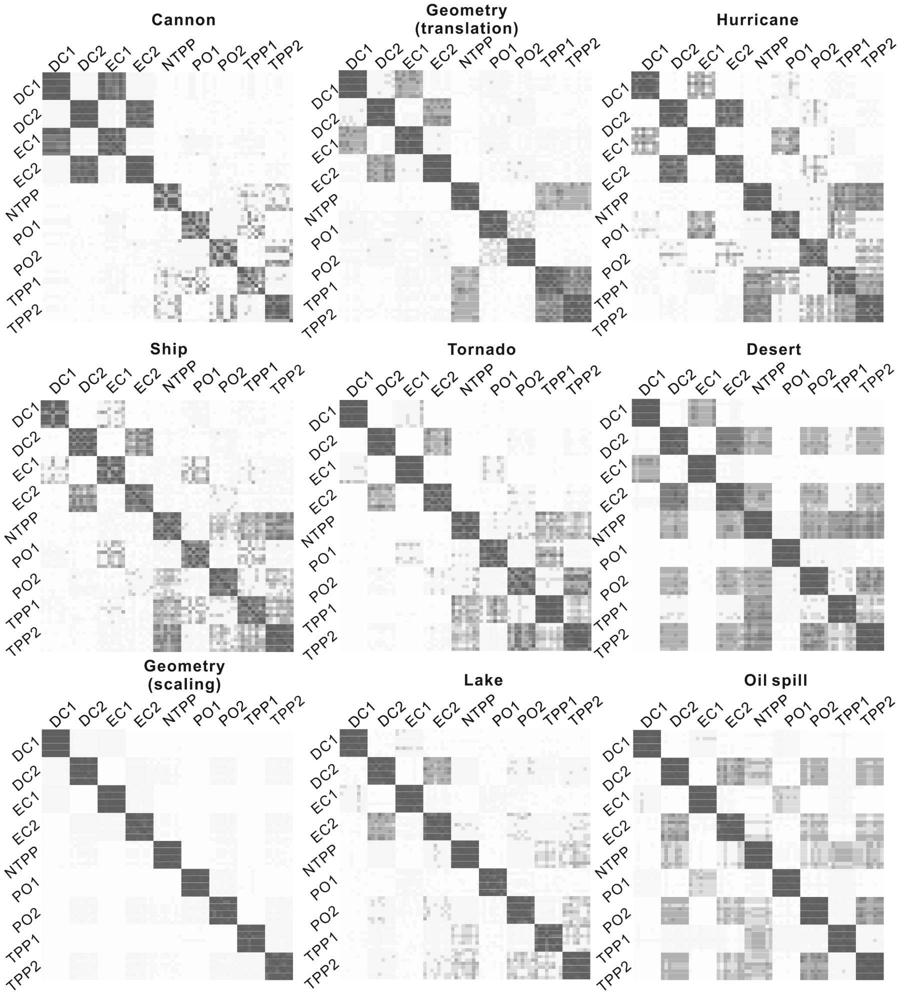


Fig. 3. Heat maps visualizing the raw similarities of all nine experiments/domains. Each heat map shows icons and TECs in the same order to allow direct comparison. Only labels for TECs are provided (not individual animations). Dark gray colors indicate high similarities, light gray to white colors indicate low or 'no' similarities (a color version of this figure can be found at min.us/m_sc2012figure3 for better readability).

3 Tailoring the Cognitive Adequacy of QSTR

In this section we are discussing how behavioral data (see Section 2) can be used to derive similarities/weights for conceptual neighborhood graphs (and potentially for pairs of topological relations that are not conceptual neighbors). We will be discussing several methods such as normalizing raw data, cluster analysis, and cluster validation techniques. In addition to using raw data, cluster analysis is chosen as it is the most common method to analyze grouping data and is thought to reveal natural groupings.

3.1 Raw Similarities

Raw similarities have been briefly introduced in Section 2. Additionally, Figure 3 provides an overview of how raw similarities are distributed within the OSMs of all nine experiments/domains. For the purpose of using raw similarities as a possibility for assessing weights of edges in conceptual neighborhood graphs (as well as, potentially, for relations that are formally not conceptual neighbors) several adjustments/standardizations have to be performed. While the behavioral data characteristics of our data allow for straightforward comparisons given that each experiment/domain had the same number of participants and the same number of icons per TEC, we will discuss normalization approaches for the purpose of creating a universal method for deriving weights on the basis of raw similarities. Raw similarities have the advantage that they can be employed not only for conceptual neighbored TECs but for all pairs of TECs.

The OSMs depicted in Figure 3 contain a large amount of redundant information. As we are concerned here with deriving weights for pairs of TECs—primarily for neighbored TECs in a CNG but the same approach can be applied to any pair of TECS—we can simply focus on the $k \times k$ sized submatrix ($k = 8$ in our case) consisting of all rows corresponding to the first TEC and all columns corresponding to the second TEC. Using $CN\_inst_{i,j}$ for the entries of this matrix, the raw similarity of the two CNs is simply computed as:

$$RSim_{CN} = \sum_{i=1}^{k} \sum_{j=1}^{k} CN\_inst_{i,j}$$

Once the raw similarities for each combination of TECs have been extracted, they have to be normalized to adjust for the specifics of the experimental setup, i.e., the number of instances in each TEC and the number of participants in each experiment. The obtained data can be normalized using row standardization taking into account all values such that individual values will be between 0 and 1. As a first step, we will look into conceptual neighbors only. Raw similarities of all CNs can be normalized to $NRSim_{CN_i}$ in the following way:

$$NRSim_{CN_i} = \frac{RSim_{CN_i}}{\sum RSim_{CN_j}}$$

3.2 Fusion Coefficients

Cluster analysis has the goal to identify natural groupings of entities (e.g., animated icons) and is frequently used in a number of disciplines (Everitt, 2001; Romesburg, 2004). There is not one specific algorithm but rather a family of cluster algorithms. In hierarchical cluster analysis, entities are stepwise assigned to groups based on similarities which are coded in similarity/proximity matrices (here: the nine OSMs). For most cluster analyses, the similarity matrix is recalculated after each clustering step, reflecting the existence of new groups. Cluster algorithms differ with respect to the way similarities are re-calculated after each grouping step (for an overview see Everitt, 2001; Romesburg, 2004). The first step in each clustering process is to combine those entities into groups that have the highest similarities. Similarities (or dissimilarities, respectively) can be considered distances and as such they are used to create so called dendrograms that reflect the clustering process (see Figure 4). Dendrograms provide an indication when two entities or groups of entities are fused (grouped) together and the distance at which they are fused is referred to as a *fusion coefficient*. All fusion coefficients are stored in a so-called *cophenetic matrix*. As similarities/dissimilarities are differently calculated by different clustering methods, there is a different cophenetic matrix for every clustering method.

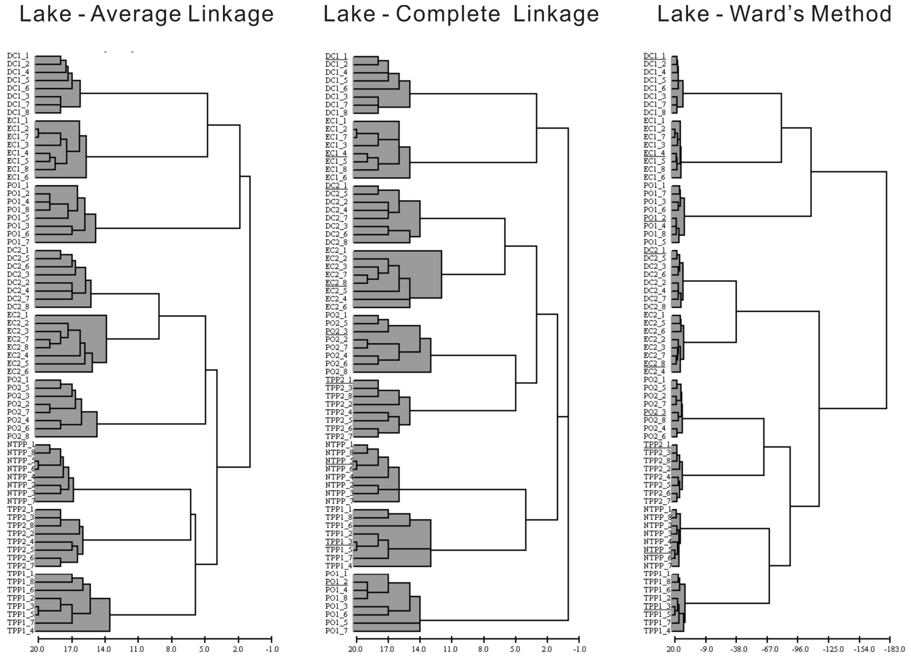


Fig. 4. Displayed are three dendrograms which reflect the clustering process for three different clustering methods (Ward's, average linkage, and complete linkage) for the same experiment/domain (lake/house). The dendrograms are visual representations of fusion coefficients indicating the value (distance) at which individual clusters are merged.

For the purpose of deriving similarities/weights for CNs on the basis of fusion coefficients, we have to briefly discuss two prerequisites. First, as different clustering methods will have different fusion coefficients (see Figure 4), we follow advice from the cluster validation literature (Ketchen & Hult, 2000; Kos & Psenicka, 2000; Milligan, 1996) and compare different methods, here: Ward's method, average linkage, and complete linkage. Second, while topology is overall a strong grouping criterion, there are situations in which TECs are indistinguishably merged (a potential indication of high similarity), individual instances of TECs might have ended up in a group with instances of a different TEC, or in the worst case, instances of certain TECs are spread across several groups.

Figure 5 shows the results for cases in which it is possible to read out one fusion coefficient as a measure of how similar two TECs are and use this as a weight for CNs. To make the data comparable, we normalized the fusion coefficients in the same way as the raw similarities (see Section 3.1).

3.3 Cluster Validation Techniques

The formal characterization based on topological equivalence classes allows for specifying a theoretical partition of the animated icons, P. This is an ideal scenario as we can employ cluster validation methods to assess whether the clustering structure Ccreated by participants matches the theoretical partition established through topological equivalence or the CNG (Halkidi, Batistakis, & Vazirgiannis, 2002b, Halkidi, Batistakis, & Vazirgiannis, 2002a). One way of comparing C and P is to calculate indices such as Rand Statistics, Jaccard Coefficient, and the Folkes and Mallows index.

These indices build on the following information: Let $C = \{C_1, \dots, C_m\}$ be the clustering structure that results from analyzing the grouping behavior of the participants recorded either in individual similarity matrices or the OSM. Let $P = \{P_1, \dots, P_n\}$ be the partition of the stimulus (animated icons) that is based on formal requirements (such as the differences between RCC-8 and RCC-5) or some introspective assumptions made by a researcher (e.g., Li & Fonseca, 2006). To be able to compare the formally derived partitioning P with the obtained results C, the following numbers are computed by comparing the containing clusters for each pair of animated icons (x_v, x_u):

- SS-a: the number of pairs of animated icons that belong to the same cluster in both, the clustering structure C and the partition P.
- SD-b: the number of pairs that belong to the same cluster in C but to different clusters in P.
- DS-c: the number of pairs that belong to the same cluster in P but to different clusters in C.
- DD-d: the number of pairs that belong to different clusters in both C and P.

The numbers for a, b, c, and d add up to the number of pairs of animated icons M. The Jaccard coefficient J, for example, is then calculated as $J = \frac{\mathrm{a}}{\mathrm{a+b+c}}$ and provides a similarity measure for comparing C and P.

To derive actual weights for the conceptual neighborhood graph based on these indices, we adapt this general approach and compute individual indices for two conceptually neighbored TECs R1 and R2 in the following way: We focus on only those icons that belong to either R1 or R2. We then consider the individual grouping of a participant and reduce it to just these icons. The resulting clustering is used for C and compared to a clustering P in which all icons from R1 and R2 are grouped into a single cluster. This means we compare the groupings of the participants to a grouping in which TECs R1 and R2 are completely combined. The values a, b, c, and d as well as the indices are then computed as described above and averaged over all participants. Using this approach, the Rand and Jaccard indices will always be the same because there is only one cluster in P and, hence, b *and* d are always zero. It also has to be noted that in the case that a is also zero (which means that the icons from R1 and R2 form two disjoint groups in C), we consider the Folkes and Mallows index to be zero (maximally dissimilar), while it is not defined in the original definition.

3.4 Comparing Similarity/Weighting Approaches

In the following, we will compare different strategies to derive similarities/weights for conceptual neighborhood graphs applying the methods discussed above to give an overview of potential weights through the perspective of similarities. The raw similarities of the nine different experiments/domains that we re-analyzed were already shown in Figure 3. Similarities are visualized as heat maps: dark gray colors indicate high similarities; light gray to white colors indicate low or no similarities. Columns and rows are organized by TECs with eight instances (animated icons) within each TEC. The order of columns and rows is kept constant (i.e., in alphabetical order) such that the heat maps are directly comparable. From the dark gray colors along the diagonals (top left to bottom right) we can infer that for most experiments/domains, the similarities within a TEC are very high. This indicates that topology is a strong grouping criterion. There are some exceptions that we will discuss in the following (e.g., the proper part relations in the cannon scenario). We also find that other TECs form strong conceptual groups, but that these similarities are susceptible to change across different scenarios.

In addition to the visualization of the raw similarities in Figure 3, Figure 5 shows the normalized weights derived by analyzing the behavioral data using the methods discussed in Sections 3.1 to 3.3.

We ran a correlation analysis over all index combinations for all nine scenarios and, as indicated by the graphs in the figures, found higher (partially near perfect) correlations between raw similarities and validation indices and slightly lower yet high correlations between fusion coefficients and raw similarities. For the time being we are only looking at fusion coefficients for CN TECs, not for individual icons as the goal of cluster analysis is to identify natural groupings.

We can make several observations:

- Fusion coefficients are not specified for all CNs. There are two reasons for this: First, although topology overall is a strong basic grouping criterion, there are some exceptions in which instances of a TEC are split and are not members of the same group. The consequence is that fusion coefficients cannot be specified between TECs. Fortunately, these cases are rare. Second, in the case that two TECs are merged together to an extent that they become indistinguishable, specifying a fusion coefficient does not make sense. In these cases, it would be most appropriate to use dissimilarities and define dissimilarities of indistinguishably merged TECs as being minimal (e.g., '0'). Specifying an exact value for similarity is more involved and for the time being, we left the value unspecified; it should be the highest possible similarity.

- The fusion coefficients deliver more pronounced graphs, compared to, for example, the raw similarities (hence the lower correlation coefficients). This is good news and bad news. On the one side, this is the intention of clustering methods, that is, strengthening within group similarities and pronouncing between group differences. On the other hand, individual clustering methods may introduce biases. While the details on how clustering methods create groups are known (Aldenderfer & Blashfield, 1984; Romesburg, 2004), it is not necessarily transparent how this plays out in a specific calculation (e.g., why they result in differences in one experiment/domain but not in another). We found that particularly complete linkage is behaving differently than other methods: (a) weights for CNs are more often not defined (see discussion above); (b) the behavior of the graphs is sometimes contrary to graphs of other methods (e.g., EC1-PO1 in the hurricane and ship scenario).
- One important observation that we will pick up in the outlook again is that there are substantial and significant (see also Klippel, accepted) differences between the similarities across the nine different scenarios. This is exemplified by the different shapes of the graphs across the different scenarios (see Figure 5). As the scenarios are topologically identical and differences such as metric information and speed have been minimized in the experimental setup, we have to conclude that weights between CNs are not independent of the semantics of a domain.
- One additional aspect that makes the assignment of weights difficult is that contextual factors play a role. In the case of the experimental data that we reanalyzed, all paths through the CNG were identical and symmetric: DC-EC-PO-TPP-NTPP-TPP-PO-EC-DC. As several domains show, the similarity between two relations, for example DC and EC, can change in dependence on whether these CN occur at the beginning of a movement pattern (DC1-EC1) or at the end of a movement pattern (DC2-EC2). This can be nicely seen by comparing the shape of the graph of the geometry-translation (GeoT) domains in Figures 5 with all other scenarios. GeoT has a near perfect symmetric shape (in contrast to other domains) with high similarities between non-overlapping CNs as well as high similarities of proper-part CNs. In contrast, similarities of CNs involving PO are relatively low. For a modeling context, this poses a challenge as contextual factors (whether a relation occurs at the beginning or at the end of a movement pattern) have to be taken into consideration.

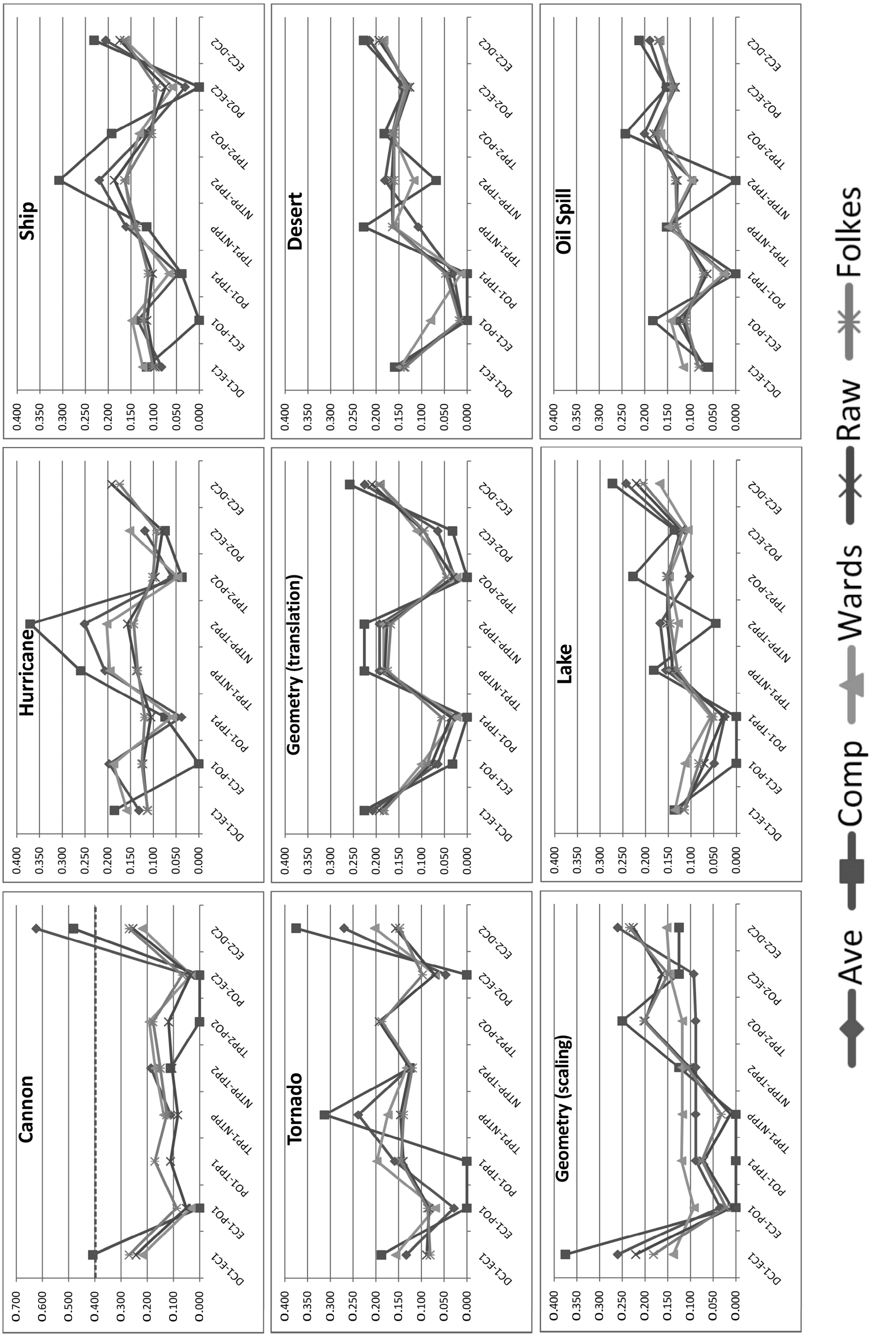


Fig. 5. Each graph visualizes the similarities / weights for CN for each of the nine scenarios using raw similarities and fusion coefficients (a color version of this figure is available at min.us/m_sc2012figure5 for better readability)

- The similarity values in Figure 5 can also be used to potentially answer the question whether an approach based on the region connection calculus is supported or whether IMs are favored. Obviously, and revealed through approaches on behaviorally evaluating QSTR approaches, TECs are forming conceptual groups. In other words, the number of JEPD relations offered by several calculi is higher than the number we would deem cognitively adequate (Clementini et al., 1993; Klippel & Li, 2009; Mark & Egenhofer, 1995). However, while coarser versions of both RCC-8 and IM exist (RCC-5 and coarse IM), these calculi do not match with respect to which relations are merged. In the case of RCC, the relations DC and EC are merged to form DR (discrete from); in the case of IM, the relations EC and PO are merged. Looking at the graphs in Figures 5, we find that some domains tend to support RCC-5, while others support coarse IM:
  - DC1-EC1 > EC1-PO1: Cannon, geometry translation, desert, geometry scaling, and lake.
  - DC1-EC1 =< EC1-PO1: Hurricane, ship, tornado, and oil.
- All scenarios, except for geometry scaling, show high similarities for CNs with proper part relations (TPP1-NTPP, NTPP-TPP2). This finding is consistent with the transition from both RCC-8 to RCC-5 and the fine to coarse transition in the IMs. This finding is also consistent with Li and Fonseca's (2006) assumption that TPP/NTPP relations are very similar to each other. However, in their model these two relations receive the highest similarity of all relations which clearly is not always supported by the data discussed here.

4 Conclusions and Outlook

4.1 Tangible Findings

All current approaches that either propose equal weights or some weighting scheme do so either introspectively or based on formal requirements. None of these approaches capture the "cognitive reality" that similarities between CNs change dependent on domain semantics. While it is still difficult to capture / derive weights directly from our data, it is clear that we need a deeper understanding of the processes at work to be able to guide weight assignments (see below for a theoretical discussion).

It is also clear that there is not simply a single formalism that will be able to capture similarity universally. We have seen a) that Clementini's (Clementini et al., 1993) proposal to use as few as five relations potentially is a step in the right direction as several TECs are very similar to each other; b) however, which TECs are considered as being more or less similar is dependent on the domain. In the analysis we showed that some scenarios follow RCC-5 while others may be better captured using the coarse version of IM.

One aspect important for using qualitative approaches to capture similarities of events is that similarities may be asymmetric (see discussion in Section 3). This aspect has been pointed out early on by Tversky (e.g., Tversky & Gati, 1978) and we do find aspects of asymmetry in nearly every domain we analyzed (e.g., whether a hurricane moves toward the coast, DC-EC-PO …, or away from the coast, PO-EC-DC). That means it matters whether two TECs are in AB or BA order. The similarity again is not something arbitrarily assigned but is an indication of an underlying (commonsense) process model that has guided participants in performing the grouping task.

4.2 A Reflection on the Methods Used

One aspect to keep in mind is that our experimental design follows a classic grouping paradigm, that is, two icons considered as being similar to each other are placed into the same group by participants. Also referred to as category construction or free classification, this method has gained widespread acceptance across a number of disciplines (Medin et al., 1987; Pothos, 2005; Roth et al., 2011), despite several limitations. The coding of similarities is binary, that is, either '0' or '1' with the implication that integers are used throughout individual similarity matrices as well as in the OSM. While the OSM reflects that individual icons may belong to more than one group or belong sort of to one but also to another group, individual matrices do not allow for such detailed distinctions. One possibility to overcome this shortcoming is to employ a different method for assessing similarities in user studies. Examples that come to mind are direct similarity assessments, that is, a participant would rate the similarity of two icons at a time on a continuous scale. The disadvantage of this method is that to achieve the same number of similarity ratings as, for example, in our experiments, 2556 comparisons would have to be made (not counting symmetric and same icon comparisons). As we (and others) have run experiments with substantially more icons, this method becomes quickly infeasible as a certain number of repetitions (per TEC) is necessary to avoid influence of individual stimuli (icons). Other methods, such as selecting the most similar icon from a group of potential target icons have a similar problem of not creating enough data points for the similarity matrix. While it is possible to simulate data, we prefer methods that provide the respective data directly (Rogowitz, Frese, Smith, Bouman, & Kalin, 1998).

One way to obtain continuous similarity rankings would be to allow participants to place icons into piles on a continuous surface (e.g., a computer screen). The advantage would be that the Euclidean properties of such a surface could be used to derive continuous similarity measures by using Pythagoras theorem. This way every pair of icons would be assigned a similarity/dissimilarity value.

The data analysis using fusion coefficients has shown that topology is often but not always the main grouping criterion. This is reflected by a few missing values in Figure 5. This problem could be prevented by using averages of fusion coefficients for all instances within TECs (in relation to all other instances of a second TEC). The raw similarity matrix would basically be replaced by the cophenetic matrix. The downside of using this method would be that the purpose of clustering methods, revealing natural groupings, would be circumvented. The bigger issue that these missing values hint at is that in cases in which TECs are split up (e.g., several overlap and proper part relations in the cannon experiment/domain), topology simply is not the main grouping criterion and as such, it is difficult to derive weights for a topological CNG from this data. Our research focused primarily on topology and as revealed by other methods we used (raw similarities and indices), the results are reasonable with respect to shedding light on similarities/weights in CNGs. However, it would be necessary to conduct similar research on other aspects of spatial knowledge such as distance and directions (Bruns & Egenhofer, 1996; Li & Fonseca, 2006). An open question is whether individual similarities in a scene (or event) can be added or whether holistic methods are necessary to assess overall similarities.

4.3 Some Theoretical Thoughts

Hirschfeld and Gelman (1994), in their introduction to their book on *Mapping the Mind*, state that "[…] much of human cognition is domain-specific." (p. 3) While domain-specificity can have multiple meanings, the one meaning important for this paper is related to semantic domains (e.g., Guarino & Giaretta, 1995) and specifically addresses geographic domains. As such, the explanation of the behavioral data might be loosely related to the concept of *theory theory* (Gopnik & Wellman, 1994), approaches to model common-sense knowledge (Davis, 1990; Hobbs & Moore, 1985), and computational approaches to semantics such as FrameNet (Fillmore & Baker, 2010). However, the main focus of this paper has been on qualitative theories of geographic event conceptualization and how they can be grounded cognitively. To this end, it is important to understand that from an ontological perspective, this paper is much closer to upper level ontologies and addresses the problem that qualitative spatial relations have been considered largely applicable in a domain-independent fashion (with the exception that different formal models are suggested based on introspections of researchers or based on formal constraints). Keil (1994) noted that "The revival of interest in domains of cognition, especially in the contexts of cross-cultural and developmental studies, is a welcome new awareness of how different sorts of concepts and belief systems might become tailored to particular kinds of lawful regularities in our physical and social worlds." (p. 234). What is needed is indeed an approach on modeling geographic event conceptualization which systematically identifies regularities in the external world and allows for providing quantitative measures that will improve the cognitive adequacy of QSTR in several information processing tasks. In the spatial sciences, process models are being developed that capture domain specific information with the goal to characterize not only entities and their relations but, additionally, underlying processes (Torrens, 2012). These approaches should be explored for the modeling of behavioral data, too.

4.4 Some Application Oriented Thoughts

Last but not least, one of the next steps in our research will be the incorporation of the developed similarity models into spatial query processing and retrieval systems. On the one hand, this would improve the usability of such systems by empowering them with the ability to provide answers based on the relational similarity to the given query and make suggestions even when no exact match can be found or in query-by-example scenarios. The provided output could then, for instance, consist of a ranked set of alternatives. On the other hand, the implemented system would allow for performing a detailed evaluation and comparison of the developed similarity models and, hence, also the different methodologies for deriving weights—using human usability studies. We are currently aiming at an implementation in the form of a generic software module that can be turned into plugins to provide similarity-based querying capabilities within existing GIS software and query interfaces to spatial information on the semantic web. The module will be instantiated with a weighted conceptual neighborhood model for an arbitrary JEPD spatial calculus together with an implementation of predicates for the different relations applicable to geometric information. It will then be able to process queries over the defined set of relations and give a similarity-ranked set of instances as a result. To deal with configurations of more than two objects, the similarity values in the weighted conceptual neighborhood graph will have to be aggregated over several relations to yield an overall similarity assessment. Investigating different approaches for this aggregation step will also be a topic of future research.

References

Aldenderfer, M.S., Blashfield, R.K.: Cluster analysis. Sage, Newbury Park (1984)

Bruns, H.T., Egenhofer, M.J.: Similarity of spatial scenes. In: Kraak, M.J., Molenaar, M. (eds.) Seventh International Symposium on Spatial Data Handling (SDH 1996), Delft, The Netherlands, pp. 173–184 (1996)

Camara, K., Jungert, E.: A visual query language for dynamic processes applied to a scenario driven environment. Journal of Visual Languages and Computing 18, 315–338 (2007)

Clementini, E., Di Felice, P., van Oosterom, P.: A Small Set of Formal Topological Relationships Suitable for End-user Interaction. In: Abel, D.J., Ooi, B.-C. (eds.) SSD 1993. LNCS, vol. 692, pp. 277–295. Springer, Heidelberg (1993)

Cohn, A.G.: Conceptual neighborhood. In: Shekhar, S., Xiong, H. (eds.) Encyclopedia of GIS, p. 123. Springer, Boston (2008)

Cohn, A.G., Renz, J.: Qualitative spatial representation and reasoning. In: van Harmelen, F., Lifschitz, V., Porter, B. (eds.) Foundations of Artificial Intelligence. Handbook of Knowledge Representation, 1st edn., pp. 551–596. Elsevier, Amsterdam (2008)

Davis, E.: Representations of commonsense knowledge. Morgan Kaufmann Publishers, San Mateo (1990)

Dylla, F., Wallgrün, J.O.: Qualitative spatial reasoning with conceptual neighborhoods for agent control. Journal of Intelligent and Robotic Systems 48(1), 55–78 (2007)

Egenhofer, M.J., Franzosa, R.D.: Point-set topological spatial relations. International Journal of Geographical Information Systems 5(2), 161–174 (1991)

Everitt, B.S.: Cluster analysis, 4th edn. Arnold, London (2001)

Fillmore, C.J., Baker, C.F.: A frames approach to semantic analysis. In: Heine, B., Narrog, H. (eds.) The Oxford Handbook of Linguistic Analysis, pp. 313–339. Oxford University Press, Oxford (2010)

Freksa, C.: Using orientation information for qualitative spatial reasoning. In: Frank, A.U., Campari, I., Formentini, U. (eds.) Theories and Methods of Spatio-Temporal Reasoning in Geographic Space, pp. 162–178. Springer, Berlin (1992)

Galton, A.: Qualitative spatial change. Spatial information systems. Oxford Univ. Press, Oxford (2000)

Goldstone, R.L., Son, J.Y.: Similarity. In: Holyoak, K.J. (ed.) The Cambridge Handbook of Thinking and Reasoning, pp. 13–36. Cambridge Univ. Press, Cambridge (2005)

Gopnik, A., Wellman, H.M.: The theory theory. In: Hirschfeld, L.A., Gelman, S.A. (eds.) Mapping the Mind: Domain Specificity in Cognition and Culture. Cambridge University Press, New York (1994)

Guarino, N., Giaretta, P.: Ontologies and knowledge bases: Knowledge building and knowledge sharing. In: Mars, N. (ed.) Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, pp. 25–32. IOS Press, Amsterdam (1995)

Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Cluster validity methods: Part I. ACM SIGMOD Record 31(2), 40–45 (2002a)

Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Clustering validity checking methods: Part II. ACM SIGMOD Record 31(3), 19–27 (2002b)

Hirschfeld, L.A., Gelman, S.A.: Toward a topography of mind: An introduction to domain specificity. In: Hirschfeld, L.A., Gelman, S.A. (eds.) Mapping the Mind: Domain Specificity in Cognition and Culture, pp. 3–35. Cambridge University Press, New York (1994)

Hobbs, J.R., Moore, R.C. (eds.): Formal theories of the common sense world. Ablex, Norwood (1985)

Johnson, M.: The body in the mind: The bodily basis of meaning, imagination, and reasoning. University of Chicago Press, Chicago (1987)

Keil, F.C.: The birth and nurturance of concepts by domains: The origins of concepts of living things. In: Hirschfeld, L.A., Gelman, S.A. (eds.) Mapping the Mind: Domain Specificity in Cognition and Culture, pp. 234–354. Cambridge University Press, New York (1994)

Ketchen, D.J., Hult, T.M.: Validating cluster assignments. Psychological Reports 87, 1057–1058 (2000)

Klippel, A.: Spatial information theory meets spatial thinking - Is topology the Rosetta Stone of spatial cognition? Annals of the Association of American Geographers (67 manuscript pages) (accepted)

Klippel, A., Li, R.: The Endpoint Hypothesis: A Topological-Cognitive Assessment of Geographic Scale Movement Patterns. In: Stewart Hornsby, K., Claramunt, C., Denis, M., Ligozat, G. (eds.) COSIT 2009. LNCS, vol. 5756, pp. 177–194. Springer, Heidelberg (2009)

Klippel, A., Li, R., Yang, J., Hardisty, F., Xu, S.: The Egenhofer-Cohn Hypothesis: Or, Topological Relativity? In: Raubal, M., Frank, A.U., Mark, D.M. (eds.) Cognitive and Linguistic Aspects of Geographic Space - New Perspectives on Geographic Information Research (in press)

Klix, F.: Information und Verhalten: Kybernetische Aspekte der organismischen Informationsverarbeitung; Einführung in naturwissenschaftliche Grundlagen der allgemeinen Psychologie. Huber, Bern (1971)

Knauff, M.: The cognitive adequacy of Allen's interval calculus for qualitative spatial representation and reasoning. Spatial Cognition and Computation 1(3), 261–290 (1999)

Knauff, M., Rauh, R., Renz, J.: A cognitive assessment of topological spatial relations: Results from an empirical investigation. In: Hirtle, S.C., Frank, A.U. (eds.) Spatial Information Theory: A Theoretical Basis for GIS, pp. 193–206. Springer, Berlin (1997)

Kos, A.J., Psenicka, C.: Measuring cluster similarity across methods. Psychological Reports 86, 858–862 (2000)

Lewin, K.: Principles of Topological Psychology. McGraw-Hill, New York (1936/1966)

Li, B., Fonseca, F.: TDD: A comprehensive model for qualitative spatial similarity assessment. Spatial Cognition and Computation 6(1), 31–62 (2006)

Lu, S., Harter, D., Graesser, A.C.: An empirical and computational investigation of perceiving and remembering event temporal relations. Cognitive Science 33, 345–373 (2009)

Mark, D.M.: Spatial representation: A cognitive view. In: Maguire, D.J., Goodchild, M.F., Rhind, D.W., Longley, P.A. (eds.) Geographical Information Systems: Principles and Applications, 2nd edn., vol.1, pp. 81–89 (1999)

Mark, D.M., Egenhofer, M.J.: Topology of prototypical spatial relations between lines and regions in English and Spanish. In: Proceedings, Auto Carto 12, Charlotte, North Carolina, pp. 245–254 (March 1995)

Medin, D.L., Wattenmaker, W.D., Hampson, S.E.: Family resemblance, conceptual cohesiveness, and category construction. Cognitive Psychology 19(2), 242–279 (1987)

Milligan, G.W.: Clustering validation: results and implications for applied analyses. In: Arabie, P., Hubert, L.J., de Soete, G. (eds.) Clustering and Classification, pp. 341–375. Word Scientific Publ., River Edge (1996)

Nedas, K.A., Egenhofer, M.: Integral vs. Separable Attributes in Spatial Similarity Assessments. In: Freksa, C., Newcombe, N.S., Gärdenfors, P., Wölfl, S. (eds.) Spatial Cognition VI. LNCS (LNAI), vol. 5248, pp. 295–310. Springer, Heidelberg (2008)

Papadias, D., Delis, V.: Relation-based similarity. In: Proceedings of the 5th ACM Workshop on GIS, pp. 1–4. ACM, Las Vegas (1997)

Piaget, J., Inhelder, B.: Child's Conception of Space. Norton, New York (1948/1956/1967)

Pothos, E.M.: The rules versus similarity distinction. Behavioral and Brain Sciences 28(1), 1–49 (2005)

Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics 19(1), 17–30 (1989)

Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connections. In: Nebel, B., Rich, C., Swartout, W.R. (eds.) Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning, pp. 165–176. Morgan Kaufmann, San Francisco (1992)

Regier, T., Zheng, M.: Attention to endpoints: A cross-linguistic constraint on spatial meaning. Cognitive Science 31(4), 705–719 (2007)

Renz, J.: Qualitative Spatial Reasoning with Topological Information. LNCS (LNAI), vol. 2293. Springer, Heidelberg (2002)

Rissland, E.: AI and Similarity. IEEE Intelligent Systems 21(3), 39–49 (2006)

Rogowitz, B.E., Frese, T., Smith, J.R., Bouman, C.A., Kalin, E.: Perceptual image similarity experiments. In: Rogowitz, B.E., Thrasyvoulos, N.P. (eds.) Proceedings of SPIE. Human Vision and Electronic Imaging III, pp. 576–590 (1998)

Romesburg, H.C.: Cluster analysis for researchers. LULU Press (2004)

Roth, R.E., Finch, B.G., Blanford, J.I., Klippel, A., Robinson, A.C., MacEachren, A.M.: The card sorting method for map symbol design. Cartography and Geographic Information Science 38(2), 89–99 (2011)

Schultz, C., Amor, R., Guesgen, H.W.: Methodologies for qualitative spatial and temporal reasoning application design. In: Hazarika, S.M. (ed.) Qualitative Spatio-temporal Representation and Reasoning. Trends and Future Directions. IGI Global, Hershey (2011)

Schwering, A.: Semantic similarity of natural language spatial relations. In: Conference on Artificial Intelligence and Simulation of Behaviour (AISB 2007): Artificial and Ambient Intelligence. Symposium: Spatial Reasoning and Communication, Newcastle upon Tyne, UK, April 2-5 (2007)

Schwering, A.: Approaches to semantic similarity measurement for geo-spatial data: A survey. Transactions in GIS 12(1), 2–29 (2008)

Torrens, P.M.: Process Models and Next-Generation Geographic Information Technology - ArcNews Summer 2009 Issue (2012), http://www.esri.com/news/arcnews/summer09articles/process-models.html (retrieved February 11, 2012)

Tversky, A.: Features of similarity. Psychological Review 84, 327–352 (1977)

Tversky, A., Gati, I.: Studies of similarity. In: Rosch, E.L.B. (ed.) Cognition and categorization, pp. 79–98. Lawrence Erlbaum, Hillsdale (1978)

Wallgrün, J.O., Wolter, D., Richter, K.-F.: Qualitative matching of spatial information. In: Abbadi, A.E., Agarwal, D., Mokbel, M., Zhang, P. (eds.) GIS 2010, Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 300–309. ACM, New York (2010)

Yang, J., Klippel, A., Li, R.: Assessing the cognitive saliency of topologically identified discontinuities in earth dynamics. International Journal of Geographical Information Science (submitted)

The Mental Representation Derived from Spatial Descriptions is North-Up Oriented: The Role of Visuo-spatial Abilities

Chiara Meneghetti, Francesca Pazzaglia, and Rossana De Beni

General Psychology Department, University of Padua
{chiara.meneghetti,francesca.pazzaglia,rossana.debeni}@unipd.it

Abstract. This study aimed to investigate: (i) whether a mental representation derived from spatial descriptions was represented according to a specific orientation, and (ii) which kind of visuo-spatial ability dictate such a mental representation.

A sample of 148 participants listened to one of four descriptions combining spatial perspectives (survey vs. route) and orientations used to provide information (from south to north [SN] vs. from north to south [NS]). Then they performed pointing (SN- and NS-oriented) and map drawing tasks, and a series of visuo-spatial measures. The results showed that: (i) SN pointing performance was better after both SN and NS descriptions, indicating that information is preferentially represented in a mental map north up oriented even when descriptions are presented in the opposite direction (i.e. from south to north). Perspective-taking was the main spatial ability involved in sustaining the mental representation when participants were required to learn information and to adopt imaginary positions from north to south.

Keywords: spatial descriptions, spatial orientation, pointing task, perspective-taking ability.

1 Introduction

1.1 Spatial Descriptions and Orientation of Mental Representation

An environment can be learnt directly (by navigation) or indirectly using visual inputs (e.g. maps) or verbal inputs, such as descriptions. There is a large body of evidence to show that mental representations derived from learning environmental descriptions have spatial characteristics. Most studies on this topic were conducted within the framework of a mental model [1]. In the case of spatial texts, a reader builds a mental model (i.e. a referential representation of the meaning of the description) that includes scenes as well as representing the language (or text). These scenes preserve physical properties of space, such as relationships between objects [2, 3], positions [4] and spatial distances [5].

A question still debated in mental model studies is whether the mental representations derived from survey and route descriptions (two different ways to

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 262–278, 2012.

provide spatial information) are similar [3] or differ [2]. Route descriptions represent space from an egocentric perspective and use an intrinsic reference frame (e.g. “to your left”, “behind you”, etc.); survey descriptions represent the space from an allocentric perspective (a bird's-eye view) and use an extrinsic reference frame such as compass directions (north, south, east, west). One way to clarify this debated question is to focus on a specific feature of mental representation such as orientation, i.e. the preferred mode for representing information in memory.

Some studies have analyzed the orientation adopted by a mental representation drawn from spatial descriptions. For this purpose, Shelton and NcNamara [6, Exp. 1] asked participants to memorize environments by using descriptions (i.e. reading survey and route descriptions) or virtual exploration (i.e. watching survey and route videos). Their performance (tested using scene recognition in different orientations, 0°-315°) was related to both orientation and perspective. Recognition performance was best when the image perspective matched the learned perspective. On the other hand, images oriented to 0° were recognized better than in other orientations (orientation effect) after reading survey and route descriptions, and after watching a survey movie, but not a route movie (in the latter case, participants performed better with images consistent with the orientation of the legs on the path). In a subsequent experiment, the authors used the pointing task to test the orientation effect of mental representation derived from visual learning of environment in the route and survey perspectives. The pointing task involved imagining being at one landmark in the environment presented, facing a second landmark and pointing in the direction of a third. Different orientations were tested from 0° (corresponding to the initial path heading for the route perspective and the fixed heading for the survey perspective, both north-up oriented) to 315°. The results showed a better performance when pointing was 0° oriented (orientation effect) regardless of the type of perspective learned.

The orientation effect on learning spatial descriptions was further examined in a series of studies by Wilson and Wildbur [7, 8]. In Wilson and colleagues [7, Exp. 1], participants read descriptions of simple paths described from an aerial (corresponding to survey) or personal (corresponding to route) point of view: 0°-oriented pointing (where 0° corresponded to the initial view aligned with the learner) was better than 180°-oriented pointing (i.e. counter-aligned with the imaginary position of the learner), with no differences emerging between the survey and route groups. This result, in which the learners’ mental representation was aligned with the first perspective they had experienced, was also confirmed when other experimental manipulations were introduced, such as adapting the description of the path to a large-scale environment [7, Exp. 2], or adding salient landmarks outside the route, or employing cardinal directions [8]. These findings support the conviction that spatial information is organized mentally in a picture-like format that enables what is at the top (personal north) and what is at the bottom (personal south) to be defined according to the viewer’s position, where the first perspective taken defines a principal reference vector or “conceptual north”. This would mean that mental representations are memorized according to a specific orientation [9, 10] even when information is encoded using descriptions [6].

Although these findings suggest that mental representation is orientation-dependent (regardless of the spatial perspective taken), to our knowledge no studies have explored the preferred orientation of mental representations drawn from descriptions of large-scale environments using a typical spatial task, i.e. a pointing task. Shelton and McNamara [6, Exp. 1] compared the mental representation formed on the strength of survey and route descriptions using a visual test (a scene recognition test). Wilson and colleagues [7, 8] used a pointing task with descriptions of simplified paths, mainly from the route perspective. There is consequently still a paucity of knowledge on whether or not the mental representations derived from environment descriptions are orientation-dependent, and whether or not there is a similar orientation effect after survey and route descriptions have been learnt.

1.2 Spatial Descriptions and Visuo-spatial Abilities

It is well recognized that visuo-spatial abilities have a central role in environment learning [11, 12], even when spatial information is conveyed using descriptions [13, 14, 15]. Spatial ability is defined as the ability to generate, retain and transform abstract visual images [16], and it includes several factors such as visuo-spatial perceptual speed, spatial visualization and mental rotation [17]; the latter two are those most often used in investigations on the characteristics of mental representations drawn from spatial descriptions. Spatial visualization (SV) is the ability to perform a multistep manipulation of spatial stimuli, while mental rotation (MR) is the ability to mentally rotate 2- or 3-dimensional stimuli. Both these types of spatial ability are related to mental representations drawn from spatial descriptions [13, 14, 15, 18, 19, 20]. For example, MR is related to learning spatial descriptions in survey [18] and in route [19, 20] perspectives.

Another spatial ability involved in environment learning is perspective-taking (PT), which consists in imagining the appearance of objects from different orientations (or perspectives) misaligned with the observer's viewpoint; it is measured with the Perspective-Taking Task (PTT) [21]. It has been demonstrated that PT is dissociated from MR, the former requiring subject-centered rotations, the latter object-centered [22]. PT is relevant to environment learning [23] and mediates the relationship between general spatial ability and spatial environment learning (such as navigation [12]), although no investigations have so far analyzed its involvement in spatial description learning.

Although the above-mentioned studies have pointed to a relevant role of SV and MR abilities in learning spatial descriptions, and possibly of PT too (though no direct evidence has been collected), no studies have examined their combined role in sustaining mental representations derived from learning spatial descriptions when the orientation effect is tested.

Fields and Shelton provided evidence of the supporting role of spatial abilities in the orientation (and changes in the orientation) of mental representations acquired by virtual exploration [24]. They asked participants to learn an open environment by means of virtual exploration from the survey or route perspective. The results of pointing tasks showed the typical orientation effect: in particular, performance was better for 0°-oriented pointing than for the other orientations. Analyzing the visuo-spatial abilities involved in pointing task performance shed important light on the spatial competences that sustained this representation: the results showed that the best predictors of survey and route pointing accuracy were MR and PT, followed by spatial span and the ability to give left/right directions on a map. The key factor involved in the change of heading orientation (after both survey and route encoding) was the PT ability, however. These findings go to show that mental representation is orientation-dependent (after both survey and route learning), and supported primarily by the ability to imagine adopting new orientations.

On the other hand, it is hard to say whether some visuo-spatial abilities (such as PT) also have a major role in sustaining the orientation (and changes in the orientation) of mental representations derived from spatial descriptions (from the survey and route perspectives).

1.3 Aims

The present study aimed to investigate: (i) whether mental representations derived from spatial descriptions are presented according to a specific orientation, and whether this remains the same for survey and route descriptions; (ii) which visuo-spatial abilities support the orientation of mental representations.

To clarify these questions, four types of description were adopted, with different combinations of spatial perspective (survey vs. route) and orientation used to provide information (north to south vs. south to north). For a given figure depicting an environment, spatial information could be presented from the bottom upwards or vice versa. In other words, the way in which spatial information was conveyed could be conceptualized as being from south to north (SN) or from north to south (NS). This enabled us to understand whether the spatial information was represented in the participants' memory according to a specific orientation, and whether or not this depended on how the information had been presented. Participants listened to descriptions given from north to south (NS d) and from south to north (SN d) from the survey or route perspectives, and they completed a pointing task by adopting imaginary positions facing north (SN p) or south (SN p).

Predictions:
(i) Orientation of mental representation
In a preliminary study, we established that participants spontaneously represented environments with a north-up orientation after learning spatial descriptions [25], but we did not know whether this orientation in their memory was preserved even when information was explicitly presented in the opposite direction, i.e. when participants were obliged to imagine starting on a path or exploring an aerial view working from north to south. We wished to clarify whether (a) mental representation is north-up oriented even when information is encoded in the opposite direction; or on the contrary whether (b) the imaginary orientation adopted by the learner during their initial encoding influenced the orientation represented in their memory.

Thus:
(a) if information on the environment is memorized with a north-up orientation whatever the mode of orientation with which this information is presented, then we would expect to see the same results in SN and NS d, with a better performance for SN p than for NS p;
(b) if the initial imagery orientation adopted by learners is important in influencing how information is represented in their memory, then we would expect to see differences in SN and NS descriptions, i.e. a better performance in SN p than in NS p in the case of SN d [as suggested by 6, 24], and a better performance in NS p than in SN p in the case of NS d (previous studies did not test this latter condition directly; the starting point was north-up oriented in 6 and 24).

We explored whether or not the orientation effect changed as a function of the spatial perspective taken.

(ii) Visuo-spatial abilities and mental representation
We analyzed which visuo-spatial abilities are involved, and to what extent, in sustaining mental representations resulting from spatial descriptions, and whether their role changes as a function of the spatial perspective taken and the way in which the information was presented.

A set of visuo-spatial tasks was administered to measure SV and MR, abilities that have been shown to correlate with spatial description learning [13, 18, 19]. PT was also measured because of its central role in predicting visual environment learning [24], assuming that its involvement might also be extended to when an environment is learnt from descriptions. PT abilities were tested using the PTT [22] and route directions on a map (Spatial Indication Task, SIT). Working memory (WM) was also tested, in terms of spatial and verbal span, using the Corsi Blocks and the Digit span tasks, because previous studies had revealed the involvement of visuo-spatial WM in spatial description learning [19, 20].

2 Method

2.1 Participants

A sample of 148 (95 females) university students (mean age = 24.2, SD = 2.63) took part in the study, divided into four groups of 37 participants (24 females) each, labeled as: route-SN description; route-NS description; survey-SN description; survey-NS description.

2.2 Material

Individual difference measures
Mental Rotations Test (MRT) [26]. The MRT measures the ability to rotate 3-D objects (testing MR ability). The task involves finding two figures among four that are identical to the target, but rotated in space (20 items, time limit 8 min).

Embedded Figures Test (EFT) [27]. The EFT involves identifying simple pictures embedded in complex configurations (testing SV ability), i.e. participants have to find the simpler shapes contained (or embedded) in a composite figure (18 items, time limit 8 minutes).

Minnesota Paper Form Board (MPFB) [28]. The MPFB tests the ability to arrange separate objects to make up a complete figure (testing SV skills). Each item consists of one 2D target object and five alternatives (i.e. five sets of fragmented parts) and participants have to decide which set makes up the target object (31 items, time limit 8 minutes).

Perspective Taking Task (PTT) [21, 22]. The PTT involves taking a new imaginary perspective within a configuration of seven objects (testing PT ability). For each item participants imagine being at one object of the layout, facing another object, and pointing to a third using a circle for giving their response (12 items, time limit 5 min).

Spatial Indication Task (SIT) [29]. The SIT measures the ability to give instructions to turn left/right to cover a route on a map (testing PT ability), which goes in directions that are also counter-aligned with respect to the observer's view) (total turns 17).

Working Memory (WM) *tasks*. The Corsi Blocks task [30] consists of reproducing sequences of blocks arranged irregularly on a board. The Digit Span task [31] involves recalling a sequence of digits. Participants are asked to reproduce increasingly long sequences of blocks/numbers in forward or reverse order. In both tasks, the length of the sequence was varied from 2 to 9 blocks or digits (using two sequences for each length).

Sense of direction and spatial representation scale (SDSR) [32]. The SDSR comprises 11 items measuring five factors: general sense of direction, knowledge and use of cardinal points, and preference for survey, route or landmark-centered representations. Responses are given on a Likert scale (from 1= "not at all" to 5= "very good"); α= 0.75, [see also 33].

Spatial descriptions

Four descriptions were prepared of the same open environment ("The Zoo"). The four descriptions were of similar length (324 to 330 words, 11 sentences) and were equally recalled (when tested in a pilot study). The zoo was a square area containing nine landmarks (Entrance, Ticket booth, Elephants, Playground, Fountain, Ice-cream parlor, Chimpanzees, Lions and Dolphins): four landmarks were located in the corners, four halfway between each pair of corners and one in the center (the distance between any two landmarks around the perimeter was 100 meters). The landmarks were organized into three legs: Entrance, Playground and Chimpanzees (Lateral leg 1); Ticket booth, Fountain and Lions (Central leg); Elephants, Ice-cream parlor and Dolphins (Lateral leg 2). In all descriptions, general information was provided about the zoo area, then the orientation localizing the initial landmark, e.g. the entrance gate, was specified (see Table 1). Given a hypothetical picture of the environment, the entrance is located in the bottom left (route)/south-west (survey) corner in the SN descriptions, and in the top right (route)/north-west (survey) corner in the NS descriptions. It is essential to specify the starting positions for route descriptions and we opted to do the same for survey descriptions as well to guarantee a parallel version. A pilot study confirmed that participants correctly detected the starting location from a south-to-north or north-to-south orientation in both perspective versions, and that the four types of description were equally well recalled.

Route descriptions provided indications from a personal view point ("turn left", "go straight on" etc.); the path starts from the Entrance located in the "bottom left-hand corner of the zoo" (SN d), or in the "top right-hand corner of the zoo" (NS d); the two paths run mainly according to the arrows shown in Figure 1a (for SN d) and 1b (for NS d). The central leg of each path was described in the opposite direction to the lateral legs. The survey descriptions used canonical terms ("south", "north-east" etc.); in both versions, the entrance was defined first ("on the southern side of the zoo, in the western corner" -SN d- or "on the northern side of the zoo in the eastern corner" -NS d-), then the landmarks were presented, working gradually from south to north or from north to south (see parts of the text in Table 1).

Pointing task

The task consisted in participants imagining standing at one landmark at the zoo, facing another landmark and pointing to a third. Each sentence was written at the top of a sheet a paper (e.g. "Imagine you are at the Entrance facing the Playground, point to the Fountain") with a circle underneath the sentence showing an arrow going from the center towards the upper edge, which was used to give the answer: the center represented the point where participants imagined being (the Entrance in this example), and the tip of the arrow was the point they were facing (the Playground); the task consisted in indicating the direction of a third (target) landmark (e.g. the Fountain) by drawing a line from the center of the circle to a point on its circumference.

Fifty-two pointing tasks were prepared; 26 were SN oriented (testing imaginary positions going from south to north) and 26 were NS oriented (testing imaginary positions going from north to south). This distinction between SN and NS was reversed when the orientation changed, e.g. if we consider the pointing direction involved in the above example, this would be SN oriented for SN d, but NS oriented for NS d. The number of items was defined according to the possible orientation in the three legs.

Table 1. Parts of descriptions expressed from south to north (SN d), and from north to south (NS d) and presented from the route and survey perspectives

	SN descriptions	NS descriptions
Initial sentence	The Podana town zoo is square in shape and occupies a flat area of land. Now you will hear a description of what there is inside this area: (*route description*) from a personal viewpoint using egocentric terms (such as "on your left", "in front of you"); or (*survey description*) from an aerial viewpoint, using canonical terms (such as north, south-east) to locate elements that are 100 meters away from each other.	
Route descriptions	"You are at the Entrance gate in the bottom left-hand corner of the zoo (…); from the Entrance, you can start walking and you will find the Playground 100 meters away".	"You are at the Entrance gate in the top right-hand corner of the zoo (…); from the Entrance, you can start walking and you will come to the Playground 100 meters away".
Survey descriptions	"The Entrance is on the southern side of the zoo in the western corner (…); the Elephants pen is in the south-eastern corner (…); the Ticket booth is halfway along the southern side (…)".	"The Entrance is on the northern side of the zoo in the eastern corner (…); the Elephants pen is in the north-western corner (…); the Ticket booth is halfway along the northern side (…)".

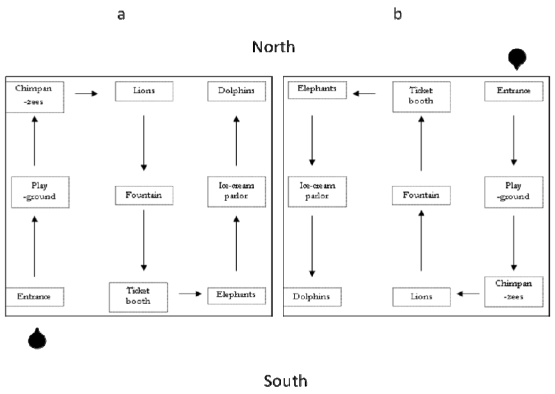


Fig. 1. The Zoo environment, showing the orientation of the information presented from south to north (panel a) or north to south (panel b). The arrows in panels a and b indicate the path covered in the route versions.

2.3 Procedure

The experimental session was divided into two parts (1 and 2) and lasted about two hours altogether. Participants were tested individually and could take a break between the two parts (only two of them did so). They were randomly assigned to one of the four types of description.

1. They listened to the spatial description twice (MP3 recordings six minutes long; hearing them twice ensured that they remembered most of the landmarks presented [19, 34]); then they completed the pointing task; each (randomly presented) item was shown on a sheet of paper containing the sentence and a circle; participants read the sentence and then indicated the direction of the landmark by drawing a line from the centre to the circumference of the circle. Then they drew a map of the environment.

2. In the second part, participants completed the visuo-spatial (MRT, PTT, EFT, SIT, Corsi Blocks, SDSR) and verbal (Digit span) measures, presented in random order.

3 Results

3.1 Scoring

Accuracy (total number of correct answers) was considered for the MRT, EFT, MPFB, SIT, and WM tasks. The degrees of error (the difference in degrees between the correct direction and the answer given by the participant) were considered for the PTT and pointing tasks.

3.2 Orientation of Mental Representation

Preliminary analyses
In the map drawing task, 3 participants drew ≤ 5 landmarks (in the NS descriptions) and were excluded from any further analyses. With the exception of these 3 cases, the participants revealed good map drawing scores (maximum score 9) with similar results for the four types of description (survey SN: M = 8.49, SD = 1.39; survey NS: M =8.59, SD = .86; route SN: M = 8.39, SD = 1.72; route NS: M =8.25, SD = 1.57).

Given that the information on the lateral and central legs was provided in the opposite direction in the route descriptions (i.e. SN-route d: lateral SN oriented vs. central NS oriented; and vice versa for NS-route d), comparisons of pointing task performance (degrees of error) were performed between the lateral and central legs in each description. No differences ($F_s < 1$) were found: the information in the central leg was represented in the same way as in the lateral legs, i.e. the central leg was mentally oriented in the same way as the lateral legs [33]. For the final analyses, the mean of all the degrees of error for the three legs was considered.

Pointing task performance
A mixed ANOVA was carried out, using 2 (spatial perspective: survey d vs. route d) x 2 (description orientation: SN d vs. NS d) as the between-participant factors – x 2 (pointing orientation: SN p vs. NS p) as the within-participant factor. The results showed significant main effects of perspective, F (1, 139) = 4.92 η2=.03 p= .03 – survey (M = 28.21 SE = 3.31) better than route (M = 38.63 SE = 3.33) - and of description orientation, F (1, 139) = 3.86 η2=.03 p= .05 – SN d (M = 28.80 SE = 3.31) better than NS d (M = 38.04 SE = 3.33) -. Only the description orientation x pointing orientation interaction was significant, F (1, 139) = 24.61 η2=.15 p≤.001. The post hoc comparisons (see means in Table 2) showed that SN p performance was better in SN d compared to NS p (p≤ .001); but SN p was also better than NS p in NS d (p= .01). The same interaction also showed that for counter-aligned pointing with description orientation (SN p in NS d vs NS p in SN d) no difference was found; while for pointing aligned with description orientation SN p in SN d were better respect to NS p in NS d (p≤ .001). These results indicate that the participants' mental representation was north-up oriented even when the descriptions presented the information going from north to south.

Table 2. Degrees of error in SN and NS pointing by orientation of the description (SN vs. NS)

Pointing orientation	Description orientation	
	SN d	NS d
	M (SE)	M (SE)
SN p	22.20 (3.33)	34.68 (3.89)
NS p	35.48 (3.85)	41.40 (3.55)

3.3 Visuo-spatial Abilities and Mental Representation

Given that the results of the ANOVA showed that the orientation of the description and of the pointing were relevant factors (while perspective did not interact with orientation), the relationships between visuo-spatial and verbal measures with NS and SN pointing in SN and NS descriptions were examined using correlations (1) and regression models (2).

1) The correlations (see Table 3) showed that both SN p and NS p (in both types of description) correlated significantly with MRT, EFT, and the backward version of the Corsi Blocks task (a higher accuracy correlated with a lower degree of pointing errors); the PTT correlated significantly with NS d (for both NS and SN p), and with SN d (only for NS p). For NS d, moreover, MPFB, sense of direction and SIT correlated with pointing performance (SIT only for NS p).

2) Regression models were used to see if the visuo-spatial measures could differently predict NS and SN pointing as a function of the type of description (SN vs. NS). Initially, the main predictors were selected using a stepwise regression, inserting pointing performance (degrees of error) as the dependent variable, and the measures significantly correlating with pointing as independent variables (see Table 3). The results showed that the measures selected were: PTT (R^2 = .22, F = 84.79 p ≤ .001; β = 47), MRT (R^2 = .04, F = 50.50 p ≤ .001; β = -.20), sense of direction (R^2 = .03, F = 35.88 p ≤ .001; β = -.12) and Corsi Blocks (backward version) (R^2 = .01, F = 28.26 p ≤ .001; β =-.11).

Then a hierarchical multiple regression model was used to ascertain how visuo-spatial measures change as a function of pointing and description orientation. In the first step, description orientation and pointing orientation (as dichotomous variables, see Table 4) and visuo-spatial measures – selected using stepwise regression - were inserted as independent variables (at continuous level). In the second step, the values corresponding to the two-way interactions between description orientation or pointing orientation and each visuo-spatial measure were inserted. In the third step, three-way interactions between description orientation, pointing orientation and each visuo-spatial measure were inserted. The results showed the significant effect of all three steps (Step 1: F = (5, 295) = 22.73 p ≤ .001; Step 2: F = (11, 295) = 11.78 p ≤ .001; Step 3: F = (14, 295) = 10.25 p ≤ .001) accounting for 28%, 3% and 3% of the variance, respectively.

In particular (as shown in Table 4), in the first step the description orientation (β = .13, p= .05) had a significant main effect, indicating that participants made more pointing errors in NS d; the PTT (β = .58, p≤ .001), MRT (β = -.23, p ≤ .05) and Corsi Blocks (β = -.26, p = .01) tasks also had a significant main effect, showing that higher scores for MRT and the Corsi Blocks, and smaller degrees of error in the PTT were associated with smaller pointing errors.

In the second step, the following interactions were significant:

- pointing orientation x PTT (β = -.36, p≤.01): performance for NS p in SN d and SN p in NS d (i.e. pointing counter-aligned with the orientation in which the descriptive information was conveyed) were more associated with PTT (fewer errors in PTT - fewer pointing errors);

- description orientation x Corsi Blocks (β= .21, p = .05): accuracy in the Corsi task was associated with pointing performance in SN d;

- description orientation x sense of direction – tendency- (β= -.16, p = .07): a higher rating for sense of direction tended to be associated with a lower degree of errors in pointing for NS d (but not in SN d).

In the third step, the only significant 3-way interaction involved pointing orientation x description orientation x PTT (β= .28, p = .01): NS p in NS d (i.e. pointing aligned with description orientation) was more associated with performance in the PTT (by comparison with SN p in SN d).

All interactions involving PTT (i.e. pointing orientation x PTT, pointing orientation x description orientation x PTT) showed that, when descriptions are presented from south to north, PT ability is involved only when counter-aligned imaginary positions are tested (with NS pointing), not the aligned positions (i.e. with SN pointing); conversely, when descriptions are presented from north to south, PT ability is involved in the imaginary positions both aligned and counter-aligned with the orientation of the description.

To ensure that the pattern of results generally overlapped in the survey and route descriptions, further regression analyses were carried out separately for two types of perspective. The results of regression models showed that the pointing orientation x PTT interaction was equally replicated for both survey and route descriptions ($p_s \leq$ .01). The pointing orientation x description orientation x PTT interaction emerged mainly for survey description (p ≤ .01) – route description (p = .07)-. For the other 2 way interactions (description orientation x Corsi Blocks, description orientation x sense of direction) no differences between survey and route descriptions were substantially found.

Table 3. Correlations of SN and NS pointing (degrees of error) in SN and NS descriptions with individual differences measures (*p ≤ .05, **p ≤ .01)

	SN descriptions		NS descriptions	
	SN pointing	NS pointing	SN pointing	NS pointing
Visuo-spatial measures				
Mental Rotation Test (MRT)	-.31**	-39**	-.38**	-.43**
Perspective Taking Task (PTT)	.19	.61**	.52**	.51**
Minnesota Paper Form Board (MPFB)	-.08	-.21	-.23*	-.32**
Embedded Figure Test (EFT)	-.23*	-.31**	-.25*	-.26*
Spatial Indication Task (SIT)	.04	-.18	-.19	-.30*
Forward Corsi Blocks task	-.16	-.16	-.04	-.08
Backward Corsi Blocks task	-.28*	-.39**	-.25*	-.32**
General sense of direction (SDSR)	-.07	-.20	-.24*	-.23*
Knowledge and use of cardinal points (SDSR)	-.05	-.04	-.05	-.05
Survey representation (SDSR)	-.14	-.20	-.08	-.10
Route representation (SDSR)	-.20	-.19	-.05	-.03
Landmark-centered representation (SDSR)	-.02	-.01	-.03	-.03
Verbal measures				
Forward digit span	-.15	-.03	-.12	-.08
Backward digit span	-.02	-.11	-.17	-.18

Table 4. Hierarchical multiple regression on pointing (degrees of error)

	Predictors	ΔR^2	β[a]	t	p
		.28 ($p \leq .001$)			
Step 1	Pointing orientation [b]		-.05	-1.01	.32
	Description orientation [c]		**.13**	**1.89**	**.05**
	PTT		**.58**	**5.15**	**≤ .001**
	MRT		**-.23**	**-2.03**	**≤.05**
	Sense of direction		-.05	<1	.56
	Corsi Blocks		**-.26**	**-2.55**	**.01**
Step 2		.03 (p = .04)			
	Pointing orientation x PTT		**-.36**	**-.3.28**	**≤ .01**
	Pointing orientation x MRT		.03	<1	.85
	Pointing orientation x Sense of direction		.03	<1	.67
	Pointing orientation x Corsi Blocks task		.09	<1	.97
	Description orientation x PTT		.09	1.10	.27
	Description orientation x MRT		.03	<1	.50
	Description orientation x Sense of direction		**-.16**	**-1.82**	**.07**
	Description orientation x Corsi Blocks		**.21**	**2.15**	**.05**
Step 3		.03 (p = .02)			
	Pointing orientation x Description orient. x PTT		**.28**	**2.25**	**.01**
	Pointing orientation x Description orient. x MRT		-.05	<1	.67
	Pointing orientation. x Description orient. x Sense of direction		-.02	-.19	.84
	Pointing orientation x Description orient. x Corsi Blocks		.11	-1.02	.31
Total R^2		.34			

[a]Standardized coefficients; [b] Dichotomous variable: 1, pointing aligned with description, i.e. SN p with SN d; NS p with NS d; 0, pointing counter-aligned with description; i.e. SN p with NS d; NS p with SN d; [c] Dichotomous variable: 1 NS d; 0 for SN d). PTT: Perspective Taking Task; MRT: Mental Rotation Test; Sense of direction: General sense of direction of SDSR. Significant results are given in boldface.

4 Discussion and Conclusions

This study investigated: (i) whether spatial descriptions (from the survey and route perspectives) are mentally represented according to a specific orientation; and (ii) which visuo-spatial abilities are mainly involved in sustaining this representation. From previous studies, we know that: (1) mental representations drawn from spatial descriptions are orientation-dependent, with the learner's initial viewpoint having a central role [6, 7, 8]. In previous studies information was mainly presented from south to north, however. In Shelton and colleagues [6, 24], for example, spatial information was presented with a north-up orientation - starting from the south-west corner - and we did not know whether or not the representation was preserved in memory with a north-up orientation when information was conveyed with the opposite orientation. Differences in the orientation effect on survey versus route descriptions have also been explored. We also know that: (2) visuo-spatial abilities sustain environment learning [11, 12], also when spatial descriptions are used [14, 18, 19], but what visuo-spatial abilities sustain this representation and the cost of maintaining a mental representation in a specific orientation remained unexplored. This study thus consisted of a novel joint analysis on the role of spatial orientation and perspective in mentally representing spatial information conveyed verbally (using descriptions), and the visuo-spatial abilities needed to sustain such representation.

Four versions of a description of a zoo environment were prepared, two mainly providing information extending from south to north in a hypothetical figure where north is at the top of the figure and south is at the bottom (SN versions), and the other two extending from north to south (NS versions). Each orientation version was presented from the survey (bird's eye view) and route (personal viewpoint) perspectives. After listening to the descriptions, participants completed pointing and map drawing tasks; the latter was carried out to check the accuracy of their mental maps that showed to be equally good in all descriptions.

(i) Orientation of mental representation. The results of the ANOVA showed that pointing performance was better in SN descriptions than in NS descriptions. It was also generally better for survey than for route descriptions (confirming previous results obtained using virtual exploration [24]). The spatial perspective did not interact with either the description orientation or the pointing orientation, however. Indeed, the latter variables only interacted with one another: in SN d, the SN p was better than the NS p; and in NS d, the SN p was better than the NS p. The SN p was therefore better than the NS p in both types of description, even if the best performance coincided with SN p associated with SN d.

These results indicate that the environment is represented in memory with a north-up orientation even when it is encoded in the opposite direction. When participants listened to spatial information that moved from north to south, their pointing northwards was still better than when they pointed southwards. In other words, the learners started their imaginary walk (in the case of a route description) or took an aerial view (in the case of a survey description) from north to south: they found the Entrance to the zoo in the upper right-hand corner (or in the north-eastern corner in survey version), then they encountered the Playground (or the playground was mentioned located halfway to east side in the survey version). Afterwards, when the orientation of their mental representation was tested by means of pointing, participants found easier to imagine themselves in a position facing north (e.g. "to be in the Playground facing the Entrance") than in a position facing south (e.g. "to be at the Entrance facing the Playground").

These results suggest that mental representations derived from spatial descriptions are like specifically oriented mental images: learners spontaneously create a north-up mental map where the top corresponds to a personal north and the bottom to a south. Even when they are obliged to encode information from north to south, the mental representation in their memory preserves its north-up orientation. It may be that simply by receiving information on the zoo's general layout (as a square in a flat area) - as provided in the first sentence in our descriptions-, participants could form in their minds a mental sketch of the zoo that was already oriented with the north at the top, in which the information they heard was subsequently placed according to the orientation they already had in their mind's eye. Further investigations will be needed, however, to see whether a north-up orientation effect is confirmed when other orientations are presented, in which information is conveyed in conjunction with other types of layout.

Taken as a whole, our results support the idea that mental representations derived from spatial descriptions (encoded from a survey or route perspective) are more like mental images that can be seen from a constant, north-facing viewpoint, rather than like mental models resembling an architect's 3D model, which can be viewed from many different point of view [1, 3]. Our results are not completely consistent with those of studies showing that a learner's initial orientation determined the whole representation [6,7,8]. Such previous studies did not manipulate the way in which the information was presented (indeed Shelton and McNamara [6] only considered paths or aerial views - in the route and survey versions – with a north-up orientation). Our findings suggest that individuals form a north-up oriented mental map even when the information they receive is encoded with the opposite orientation, from north to south. Overall our findings showed that the orientation of mental representations is at least partially influenced by the initial imaginary view (though SN pointing performance was best in the SN description) and it is uninfluenced by spatial perspective. The beneficial effect of perspective probably stems from the fact that pointing was better supported when an extrinsic frame of references was used to present the environment. Preserving the north-up orientation of a mental representation (when information is encoded with a NS orientation) is highly demanding, however, as shown by the involvement of visuo-spatial abilities (see next paragraph).

(ii) Visuo-spatial abilities and mental representation. The visuo-spatial abilities proved to have an important role in sustaining the mental representation derived from spatial descriptions, especially when the information was not encoded with a north-up orientation. This emerged from the results of correlations in which more visuo-spatial tasks were related with NS descriptions: MRT, VSWM (the backward version of the Corsi Blocks task), EFT and PTT correlated with both SN and NS descriptions, while MPFB, the sense of direction and the SIT were related only to NS descriptions, meaning that learning NS descriptions was a more (spatial) resource-consuming activity. The results of the regression models showed that the role of visuo-spatial abilities changes as a function of pointing and description orientation. In fact, MR, sense of direction and VSWM were relevant factors in predicting a good performance after learning spatial descriptions, as previously suggested [19, 34], but PT ability (which specifically interacts with pointing and description orientation) also had an important role.

The two-way interaction between pointing orientation and PTT in the regression model indicated that PT ability was the main predictor of performance for pointing with imaginary positions counter-aligned vis-à-vis the orientation proposed in the description, i.e. when participants performed NS pointing tasks after listening to SN descriptions, or when they performed SN pointing tasks in response to NS descriptions. This result is applied equally to survey and route descriptions. At the same time, the three-way interaction indicated that PTT was also the main predictor for NS pointing relating to NS descriptions especially for survey descriptions.

These results indicate that when participants listened to north-up descriptions and they, then, imagined occupying north-up positions (with SN pointing), i.e. when they imagined being in positions consistent with their spontaneously adopted orientation of their mental representation (north-up), the role of visuo-spatial abilities was marginal. Instead, the degree of involvement of the visuo-spatial skills, and of PT in particular, became relevant when: (i) participants performed NS pointing after learning SN descriptions; and when (ii) they learned spatial descriptions from north to south and performed NS and SN pointing. PT ability therefore became important in supporting imaginary positions taking a counter-aligned viewpoint with respect to the encoded orientation (i.e. for SN p in NS d and for NS p in SN d); in addition PT ability was also involved when descriptions were presented from north to south and tested imaginary positions in the same direction. This latter result indicates that participants obeyed the request to learn the information from north to south but pointing from north to south proved to be highly (spatial) resource-consuming.

Overall, the novelty of our findings lies in that they extend the role of PT ability, even when spatial information is encoded verbally (as well as visually [24]). The fact that the performance of PT task was a stronger predictor of pointing task performance than the other visuo-spatial tasks can be explained by the similarity between the two tasks. In both cases, participants were asked to imagine being on one object, facing another and pointing to a third, indicating the direction with the aid of a circle. The substantial difference lies, however, in that the PTT is perception-based (the arrangement of the objects remains in view while the learner is pointing, and all the items test imaginary positions not aligned with the learner's view, but at an angle of at least 90°); while the pointing task is memory-based, relying on the learner's memorized mental representation of the zoo's layout after listening to its description, and the pointing test is in aligned (0°) or counter-aligned (180°) views. Perception-based misaligned pointing thus predicted memory-based misaligned (i.e. counter-aligned) pointing with a description orientation, and pointing aligned with descriptions oriented from north to south. This latter result was particularly strong when the information was encoded from a bird's eye view (survey perspective), a condition in which the environment layout is more clearly presented as a picture (that is then committed to memory in north up orientation) than in the case of a route description. These results suggest that mental representations formed after learning a description have properties similar to those of actually perceived images, as suggested by previous studies [e.g. 35].

Further investigations could be conducted to clarify whether other visuo-spatial abilities are involved when the orientation effect is tested using other measures (instead of pointing task), such as the scene recognition task [10].

To sum up, the results of the present study provide novel evidence indicating that spatial information learned from descriptions is represented in the mind as a north-up oriented mental image. When spatial information was encoded and tested in a view counter-aligned with this spontaneous orientation, perspective-taking ability have a crucial role in compensating for those misalignments.

References

1. Johnson-Laird, P.N.: Mental models: towards a cognitive science of language, inference and consciousness, p. 513. Cambridge University Press (1983)
2. Perrig, W., Kintsch, W.: Propositional and situational representations of text. Journal of Memory and Language 24, 503–518 (1985)
3. Taylor, H.A., Tversky, B.: Spatial mental models derived from survey and route descriptions. Journal of Memory and Language 31, 261–292 (1992)
4. Bryant, D.J., Tversky, B., Franklin, N.: Internal and external spatial frameworks for representing described scenes. Journal of Memory and Language 31, 74–98 (1992)
5. Bower, G.H., Morrow, D.G.: Mental models in narrative comprehension. Science 247, 44–48 (1990)
6. Shelton, A.L., McNamara, T.P.: Orientation and perspective dependence in route and survey learning. Journal of Experimental Psychology: Learning, Memory, and Cognition 30, 158–170 (2004)
7. Wilson, P.N., Tlauka, M., Wildbur, D.: Orientation specificity occurs in both small- and large-scale imagined routes presented as verbal descriptions. Journal of Experimental Psychology: Learning, Memory, and Cognition 25, 664–679 (1999)
8. Wildbur, D.J., Wilson, P.N.: Influences on the first-perspective alignment effect from text route descriptions. The Quarterly Journal of Experimental Psychology 61, 763–783 (2008)
9. Shelton, A.L., McNamara, T.P.: Systems of spatial reference in human memory. Cognitive Psychology 43, 274–310 (2001)
10. McNamara, T.P.: How Are the Locations of Objects in the Environment Represented in Memory? In: Freksa, C., Brauer, W., Habel, C., Wender, K.F. (eds.) Spatial Cognition III. LNCS (LNAI), vol. 2685, pp. 174–191. Springer, Heidelberg (2003)
11. Hegarty, M., Montello, D.R., Richardson, A.E., Ishikawa, T., Lovelace, K.: Spatial ability at different scales: Individual differences in aptitude-test performance and spatial-layout learning. Intelligence 34, 151–176 (2006)
12. Allen, G.L., Kirasic, K.C., Dobson, S.H., Long, R.G., Beck, S.: Predicting environmental learning from spatial abilities: an indirect route. Intelligence 22, 327–355 (1996)
13. De Vega, M.: Characters and their perspectives in narratives describing spatial environments. Psychological Research 56, 116–126 (1994)
14. Bosco, A., Filomena, S., Sardone, L., Scalisi, T.G., Longoni, A.M.: Spatial models derived from verbal descriptions of fictitious environments: the influence of study time and the individual differences in visuo-spatial ability. Psychology Beitrage 38, 451–464 (1996)
15. Haenggi, D., Kintsch, W., Gernsbacher, M.A.: Spatial situation models and text comprehension. Discourse Processes 19, 173–199 (1995)
16. Lohman, D.F.: Spatial ability: a review and reanalysis of the correlational literature (Tech. Rep. No. 8). Stanford, CA: Stanford University School of Education (1979)
17. Linn, M.C., Petersen, A.C.: Emergence and characterization of sex differences in spatial ability: a meta-analysis. Child Development 56, 1479–1498 (1985)

18. Pazzaglia, F.: Text and picture integration in comprehending and memorizing spatial descriptions. In: Rouet, J.F., Lowe, R.K. (eds.) Understanding Multimedia Documents, pp. 43–59. Springer, NYC (2008)
19. Meneghetti, C., Gyselinck, V., Pazzaglia, F., De Beni, R.: Individual differences in spatial text processing: high spatial ability can compensate for spatial working memory interference. Learning and Individual Differences 19, 577–589 (2009)
20. Meneghetti, C., De Beni, R., Pazzaglia, F., Gyselinck, V.: The role of visuo-spatial abilities in recall of spatial descriptions: a mediation model. Learning and Individual Differences 21, 719–723 (2011)
21. Hegarty, M., Waller, D.: A dissociation between mental rotation and perspective-taking spatial abilities. Intelligence 32, 175–191 (2004)
22. Kozhevnikov, M., Hegarty, M.: A dissociation between object-manipulation spatial ability and spatial orientation ability. Memory and Cognition 29, 745–756 (2001)
23. Kozhevnikov, M., Motes, M.A., Rasch, B., Blajenkova, O.: Perspective-taking vs. mental rotation transformations and how they predict spatial navigation performance. Applied Cognitive Psychology 20, 397–417 (2006)
24. Fields, A.W., Shelton, A.L.: Individual skill differences and large-scale environmental learning. Journal of Experimental Psychology: Learning, Memory and Cognition 32, 506–515 (2006)
25. Meneghetti, C., Pazzaglia, F., De Beni, R.: Is mental representation derived from spatial descriptions north-up oriented? (under revision)
26. Vandenberg, S.G., Kuse, A.R.: Mental rotation, a group test of three-dimensional spatial visualization. Perceptual and Motor Skills 47, 599–604 (1978)
27. Oltman, P.K., Raskin, E., Witkin, H.A.: Group embedded figures test. Consulting Psychologists Press, Palo Alto (1971)
28. Likert, R., Quasha, W.H.: Revised Minnesota Paper Form Board. Psychological Corporation, New York (1941)
29. Nori, R., Giusberti, F.: Cognitive styles: errors in directional judgments. Perception 32, 307–320 (2003)
30. Corsi, P.M.: Human memory and the medial temporal region of the brain. Unpublished doctoral dissertation, McGill University, Montreal (1972)
31. Wechsler, D.: Wechsler Adult Intelligence Scale, rev. edn. Psychological Corporation, New York (1981)
32. Pazzaglia, F., Cornoldi, C., De Beni, R.: Differenze individuali nella rappresentazione dello spazio: presentazione di un questionario autovalutativo [Individual differences in representation of space: presentation of a questionnaire]. Giornale Italiano di Psicologia 3, 627–650 (2000)
33. Pazzaglia, F., De Beni, R.: Strategies of processing spatial information in survey and landmark-centred individuals. European Journal of Cognitive Psychology 13, 493–508 (2001)
34. Meneghetti, C., De Beni, R., Gyselinck, V., Pazzaglia, F.: Working memory involvement in spatial text processing: what advantages are gained from extended learning and visuo-spatial strategies? British Journal of Psychology 102, 499–518 (2011)
35. Denis, M., Cocude, M.: Scanning visual images generated from verbal descriptions. European Journal of Cognitive Psychology 1, 293–307 (1989)

Linguistic Principles for Spatial Relational Reasoning

Thora Tenbrink and Marco Ragni

SFB/TR 8 Spatial Cognition, University of Bremen/Freiburg, Germany
tenbrink@uni-bremen.de,
ragni@cognition.uni-freiburg.de

Abstract. Human spatial relational reasoning has been investigated by presenting participants with premises like: "The triangle is to the left of the circle, the circle is to the left of the square. Which relation holds between the triangle and the square?" Participants are expected to interpret the descriptions in a way that corresponds to the logical options that are theoretically available. Recent findings on spatial language usage highlight a range of pragmatic principles that speakers intuitively adhere to when producing and comprehending spatial relationship descriptions; these appear to contradict the principles used in relational reasoning studies. In order to clarify the relation between speakers' intuitions and the descriptions used in relational reasoning tasks, we present two studies in which linguistic representations of relevant configurations were elicited. Results highlight the systematic patterns speakers use in describing these configurations, adding new insights to research on spatial language usage across various levels of analysis. We argue that the identified principles may interfere with the reasoning processes investigated in earlier studies, and suggest that future studies should adequately account for the principles underlying intuitive spatial language use.

Keywords: Spatial language, linguistic preferences, N-term series problems, free production task, spatial reasoning.

1 Introduction

If your car is parked in front of a house, and your friend's car is parked behind yours, what is the relationship between your friend's car and the house? Relational reasoning of this kind plays an important role in everyday life (Johnson-Laird, 2001) and has been a central topic in cognitive psychology for a long time (De Soto, London, & Handel, 1965). Traditionally, relational reasoning is investigated using the paradigm of so-called *3-term series* tasks. An example for a 3-term series problem is given in the following:

(1) The triangle is to the left of the circle.
The circle is to the left of the square.
Which relation holds between the triangle and the square?

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 279–298, 2012.

3-term series tasks consist of two premises (given as statements) and a question about the conclusion to be drawn from the premises, such as the spatial relation between two objects as in this example. These problems are given as language-based descriptions of spatial (or other types of) relationships. Participants are expected to interpret the descriptions in a way that corresponds to the logical options that are theoretically available, typically assuming a spatial grid that only allows for specific equidistant positions. This type of reasoning with relations is expected to mirror reasoning processes in everyday life. The investigation of the cognitive processes involved in such tasks has concerned linguistic reasoning (e.g., Clark 1969a,b) just as well as non-linguistic transitivity inference (e.g., Halford et al., 1995). Research of this kind has led to the fundamental insight that human reasoning does not only rely on propositional thought processes but also on imagery-based mental models of the described situation (Johnson-Laird, 2001). In other words, people solve such problems by visualizing the relationships involved rather than by employing solely abstract logic reasoning.

Linguistic representation has long been recognized to be systematically related to thought (e.g., Fodor, 1975; Gerrig & Banaji, 1994). Therefore the ways in which humans intuitively employ language to represent relationships between entities should be a matter of central concern in this field. Surprisingly little, however, is known so far about the extent to which the abstract relationships described in 3-term series problems, presented to humans in a psychological laboratory setting, correspond to humans' intuitions about how such relationships should be described. The interpretation of relational reasoning experiments is heavily biased towards non-linguistic reasoning processes, basically ignoring the fact that these tasks rely on natural language comprehension processes just as well as using language in more natural contexts does. Superficially, in fact, the descriptions do resemble natural language usage; however, the extent to which speakers would actually be ready to use language in this way has not been questioned in any systematic way so far.

Early discussions about the possible impact of principles underlying language use (e.g., Clark, 1969a,b) on the comprehension of 3-term series problems centered around (now partly out-dated) linguistic theories of that time. Over the past decades, spatial language research has been highly active, providing ample grounds for reconsidering the situation particularly for the spatial domain. Further evidence about linguistic intuitions can be gained by eliciting spontaneous natural language descriptions, which is what we do in this paper by employing a spatial description task that uses the kind of configurations employed for 3-term series problems. We contend that if the descriptions used to invoke mental models violate general (implicit) principles of natural language usage, the validity and generalization of the results from relational reasoning tasks needs to be re-considered.

In the following, we will first introduce the theory of mental models, along with recent insights on preferred mental models. Then we turn to a brief account of what is known about spatial representation in language. Against this background we provide a categorical analysis of a range of principles underlying the linguistic descriptions used in N-term models, leading to qualitative hypotheses as to how speakers may intuitively describe spatial configurations of this kind. Our experimental study then sets out to test these hypotheses by eliciting written descriptions of spatial models. Results are discussed with respect to the consequences for research on relational reasoning that is based on natural language descriptions.

2 The Theory of Mental Models

Human spatial reasoning is nowadays generally interpreted in the framework of mental models theory (henceforth MMT; Johnson-Laird & Byrne, 1991). In MMT, a mental model of premise information is characterized as an integrated representation in which the premises are true. Spatial mental models are analogical representations of space (Knauff & Johnson-Laird, 2002). The objects are used to represent spatial relations by their position. Descriptions like “The triangle is to the left of the circle” (premise 1 in example 1 above) are simply represented by an arrangement (according to the relations) of the two objects. Mental models are abstractions and represent only the essential parts of the model. This implies that mental models are more abstract than visual models would be, and represent qualitative relations instead of metric information. Similarly, according to modern linguistic theories, spatial language does not represent quantitative or metric information by its lexicogrammar, but rather qualitative, function-based distinctions (e.g., Talmy, 2000).

Considering example (1) above, a number of observations and generalizations can be noted. Presenting the same information given in the second statement in a different way, such as in (2) below, is more difficult for the reasoning process (e.g., Knauff et al., 1998; Clark, 1969a,b):

(2) The square is to the right of the circle.

Both the early linguistic theories as well as current MMT approaches assume that the enhanced difficulty is due to a conceptual manipulation of the given information towards a prototypical form or deep structure resembling (1).

Furthermore, example (1) represents a case of a 3-term *relational verification problem*, which is just one typical format. First, there can be more than two premises, leading to a different number of terms (which is why the paradigm is sometimes referred to as N-term series problems). Second, the problems posed can be of a different nature. A *relational generation problem* consists of a set of premises without a concrete question about the conclusion. Here is a classical example of a 5-term relational generation problem:

(3) The hammer is to the right of the pliers.
The screwdriver is to the left of the pliers.
The wrench is in front of the screwdriver.
The saw is in front of the pliers.
What follows?

A conclusion has to be valid, i.e. it has to be true in all models consistent with the premises. A putative conclusion which is not satisfiable contradicts the set of premises, i.e., the negation of the conclusion follows from the set of premises. Such generation problems are of great importance for everyday life.

According to MMT there are three stages of thinking that reasoners go through during reasoning: In the *comprehension phase*, reasoners use their general knowledge and knowledge about the semantics of spatial expressions to construct an internal model of the "state of affairs" that the premises describe. This is the stage of the reasoning process in which the given premises are integrated into a unified mental model. According to this theory, only the mental model needs to be kept in memory, i.e. the premises may be forgotten (Mani & Johnson-Laird, 1982). Crucially, spatial descriptions may be vague; more than one possible model may be consistent with a given set of premises. Hence *determinate* task descriptions allowing for the construction of only a single model (based on the constraints of an underlying grid pattern) need to be distinguished from *indeterminate* task descriptions allowing for two or more models consistent with the premises (Rauh et al., 2005).

In the *description phase*, a parsimonious description of the mental model is constructed, including a preliminary conclusion. In other words, the mental model is inspected to find out relations that are not given explicitly.

In the *validation phase*, people try to find alternative models of the premises in which the conclusion is false. If they cannot find such a model, the conclusion must be true. If they find a contradiction, they return to the first stage – and so on until all possible models are tested (Johnson-Laird & Byrne, 1991). For this reason the validation phase could be viewed as an iteration of the first two phases in which alternative models are generated and inspected in turn.

An extension of the classical MMT is the so-called preferred mental model theory (PMMT, Knauff, Rauh, & Schlieder, 1995). Experimental studies addressing human reasoning with multiple model problems show a high consistency with respect to the generated conclusions. Thus, the majority of participants confronted with a particular set of premises tend to construct the same mental model. PMMT explains how such a first mental model is constructed, and why this model is "preferred" over others. For example, it appears to be easier to construct and to maintain just one model in working memory rather than all possible models simultaneously (Ragni et al., 2006).

Relational generation problems like our example 3 above involve complicated processes of inference particularly in the description phase, based on verbal data incorporated in the comprehension phase. To this date, it is not clear in what ways the verbalization itself supports or interferes with these processes. In particular, it is unknown to what extent abstract reasoning tasks like example 3 resemble more natural linguistic choices, for example, produced by speakers in everyday tasks involving reasoning processes. In this regard, the following aspects are particularly important: Which relations do humans use to describe positions of objects? If they have a (mental) model at hand – as is the case, according to PMMT, already in phase 2 – what are the linguistic principles underlying the description and encoding of such information? Before we present our linguistic free production experiment addressing these questions directly, we first review recent research in spatial language usage that provides relevant insights.

3 Spatial Language and Pragmatic Principles

In a heated controversy in the late 60-ies and early 70-ies of the last century, Clark (e.g., 1969a,b) and Huttenlocher (e.g., 1968), attempted to weigh the construction of spatial mental models against linguistic processes involved in comprehending the linguistic representations of 3-term series problems. Ormrod (1979) proposed to reconcile the two competing theories. Such a multi-layered view is representative of more recent work in this area; nowadays there is a high agreement concerning the cognitive complexity and the centrality of mental models involved in reasoning processes of this kind, even for non-spatial problems (Johnson-Laird, 2001). Nevertheless, Clark in his early work was right in at least one crucial respect: General linguistic processes and principles should not be ignored in research on relational reasoning. This is not only true for the kinds of phenomena that Clark pointed out at that time (which at least in part remain unchallenged), but even more so for the role of general action and situation contexts in using language that he identified in his later work (Clark, 1996). From a relevance-theoretic point of view, people confronted with abstract reasoning tasks assume that they are expected to find the most *relevant* conclusion (Van der Henst, Politzer, & Sperber, 2002). A considerable range of earlier findings concerning participants' determinate answers to indeterminate problems, and systematic errors with determinate problems, can be accounted for in this way. Here our focus lies on the insights to be gained by examining spatial language, which is known to be particularly central to mental imagery and is also a widespread medium for the investigation of relational reasoning problems.

We begin by addressing general principles concerning the mapping between spatial relationships and linguistic representation. Spatial relational reasoning tasks typically draw on a very small subset of relational terms out of a fairly wide repertory of spatial expressions offered in a language. In fact, entities may be linguistically related to one another in many ways. Following common terminology, a spatial relational description consists of a *locatum* that is being described, a *relatum* in relation to which the locatum is described, and a *spatial term* that describes the relation between locatum and relatum. All three of these *roles* may be filled in various ways, depending on a variety of factors. A number of principles underlying this mapping process have been proposed; for instance, Herskovits (1986) discusses how context-based relevance and salience phenomena constrain the use of spatial prepositions, and Coventry & Garrod (2004) address the impact of functional relationships between entities on spatial description choice. The specificity of a description with respect to the spatial relation (or direction) has been shown to vary systematically based on the nature of the spatial relationship and the number and location of objects in a configuration, current discourse goals, the requirements of the interaction partner, and many other factors (cf. Tenbrink, 2007). Talmy (2000) suggests that one entity (the locatum) is generally focused on while the other (the relatum) serves as background; this allocation may depend on the discourse history just as well as on the nature of the entities involved. In the following, we refer to this general mapping question as **spatial representation**.

Next, the question may be asked how objects need to be related to each other in order for a spatial term to apply. In this respect, spatial terms differ widely. There seems to be no particular constraint (except for spatial direction) on the usage of compass-based terms (*north of, south of*) as well as comparative terms such as *higher than*. In contrast, for the so-called *projective* terms (*left, right, front, behind, above, below*), which are often used in spatial reasoning tasks, it has repeatedly been noted that the objects in question need to be situated not only in close vicinity but actually *immediately adjacent* to each other. While there is some controversy concerning the potential flexibility of this principle and concerning the extent to which the principle holds across various syntactic forms available for projective terms (e.g., *on the left* vs. *to the left*), it is generally agreed that typically no other object should be situated between the locatum and relatum when these terms are used (Herskovits, 1986; Pribbenow, 1992; Talmy, 2000). Intuitively, this principle may be relaxed somewhat when many objects are collectively described in relation to another object. It may also play a role if the discourse task is to uniquely identify a referent or rather to describe an object's location (Vorwerg & Tenbrink, 2007; Carlson & Hill, 2009). When speakers describe more complex configurations in the real world, they tend to describe new objects in relation to adjacent objects that have already been described (Tenbrink, Coventry, & Andonova, 2011). We will refer to this potential constraint as **spatial directness**.

Much research on spatial language usage has focused on the description of only two objects relative to each other. Whenever more complex spatial configurations or scenarios come into play, speakers are confronted with the so-called linearization problem (Habel & Tappe, 1999); they need to provide a sequential account of a two- or three-dimensional situation. Only little is known about the principles governing this linearization in the absence of a natural order (such as temporal sequence in a route description). Spatial clusters (Ehrich & Koster, 1983) as well as functional relationships between objects (Andonova et al., 2010) have been found to affect the order of description for complex object arrangements. In the absence of these, speakers tend to follow a fairly regular, linear pattern, similar to the motion path of a person traveling through a spatial environment (Linde & Labov, 1975; Levelt, 1982). We will refer to the sequential order of successive spatial descriptions as **trajectory**.

Apart from the order of descriptions, the question arises as to how speakers frame the spatial descriptions syntactically. As such, the allocation of locatum and relatum does not yet determine the syntactic format, as there are various options:

(4) The table is to the right of the armchair.
(5) To the right of the armchair there is the table.
(6) The armchair is the object that the table is to the right of.

In a natural discourse context, speakers would typically know the position of one object and specify another object's position in relation to the known one. This would be reflected in the given-new distribution of the utterance. For example, if the speaker wishes to specify where the *table* is, the table (as locatum) would be a given entity, though its position is still unknown. As shown in example (4) the sentence therefore starts from the locatum and provides the new information – the table's location in relation to the relatum – in the latter part of the sentence, which is typically associated with new information (e.g., Halliday, 1994). In a situation in which none of the objects has yet been specified or is in the current focus of attention, it can be assumed that one object is first introduced by an indefinite article or even explicitly by stating its existence. We refer to these phenomena in terms of **syntactic format and information structure.**

4 Features of Linguistic Descriptions in Relational Reasoning Tasks

In the following, we examine how each of the four aspects of spatial language use identified in the previous section are (typically) represented in the spatial descriptions given to participants in relational reasoning studies. For illustration we use a typical five-term series problem as exemplified in Table 1. Note that the models assume an underlying grid pattern with equal distances between item positions.

Table 1. Relational reasoning task (Johnson-Laird & Byrne, 1991)

Model description	Possible models consistent with the description
A is to the left of B. C is to the right of A. D is behind C. E is behind A.	E D A B C
	E D A C B

4.1 Spatial Representation

The premises in relational reasoning tasks such as the one given in Table 1 always describe the relative position of two objects, using projective terms indicating a concrete spatial direction. Locatum and relatum are allocated without any discernible underlying principle. The descriptions do not contain indeterminate expressions such as *beside* or any other spatial terms (*beside, higher than*) that may also suitably represent the spatial situation (in part without indicating spatial directness to the same extent). Negations such as *not to the right of* were addressed in Schleipen et al. (2007), and Hörnig et al. (2006) used *between.*

4.2 Spatial Directness

In the problems given to experimental participants, spatial relationships are not necessarily intended to be direct. The premise "A is to the left of B" only indicates a general spatial direction; the objects could be direct neighbors, or a further entity might be situated between them, as in the second model represented as consistent with the description. This feature of relational reasoning problems is not an explicit part of the description itself, but needs to be derived from the fact that the problems could often not be solved otherwise. Additionally, this particular type of interpretation may be explicitly pointed out in the instructions to the experimental participants. Although Ragni et al. (2007:179) mention that humans tend to interpret projective terms as expressing direct relationships, they discard this aspect as having no influence on the results.

4.3 Trajectory

In contrast to the findings on spatial language usage described above, the descriptions do not follow any specific trajectory or pattern such as "from left to right" or "from top to bottom". Rather, the relationships are described in a seemingly random order, typically designed to raise certain expectations on the part of the problem solver.

4.4 Syntactic Format and Information Structure

The description above uses a uniform syntactic format, which may be characterized as LOCATUM SPATIAL-RELATIONSHIP RELATUM. Precisely the same format is used for the first sentence as for the following ones, without adjustments concerning information structure as described above. A broad range of earlier work in this area uses this syntactic pattern. However, as Hörnig et al. (2006) demonstrate, word order and information structure matter for the mental reasoning process.

4.5 Preliminary Conclusion

From this analysis it can be hypothesized that the linguistic descriptions used for spatial relational reasoning tasks do not necessarily correspond to speakers' intuitions. In this sense, one might say that the descriptions are 'linguistically naïve', and may lead to an unintended bias for the construction of mental models. However, so far no previous work has elicited speakers' natural language descriptions of the configurations used in relational reasoning problems. Therefore, it may still be the case that this particular type of abstract spatial scene might lead to different patterns in speakers' descriptions, at least if encouraged by the setting. In order to explore the extent to which this is the case, we designed two free production studies as described next.

5 Experimental Studies

In a study of 5-term series problems, resembling other relational reasoning studies, Ragni et al. (2007) had participants identify spatial configurations corresponding to sets of premises formulated in natural language. In our studies, the reverse procedure was applied; the participants were given spatial configurations and asked to describe them, using natural language. In order to explore speakers' intuitions about natural language use in this particular kind of situation, we asked naïve participants to write linguistic descriptions of spatial configurations resembling those used for 5-term series problems. Based on findings on spatial language, the following patterns could be expected with respect to the four types of phenomena addressed in the previous section.

- **Spatial representation.** Speakers would tend to use spatial relational terms that are sufficient and determinate enough to specify the spatial relationship between locatum and relatum qualitatively.
- **Spatial directness.** Speakers would be reluctant to use projective terms for non-direct spatial relationships (i.e., two objects not directly adjacent to one another).
- **Trajectory.** Speakers would use an orderly trajectory through the spatial model, and allocate locatum and relatum accordingly.
- **Syntactic format and information structure.** Speakers would adhere to syntactic formats appropriate for the information structure chosen for description.

Due to their nature, these expectations are necessarily formulated in qualitative terms, since natural language use is never entirely predictable. Nevertheless, to the extent that these patterns are identifiable in the descriptions produced by experimental participants, they do in some respects contradict the principles identified in the descriptions used in traditional relational reasoning research. Furthermore, a confirmation or rejection of these description patterns for spatial models of this kind adds to the body of research accumulating concerning spatial language use.

In order to encourage participants as much as possible to *deviate* from our expectations, we provided example descriptions that corresponded to the premises used in spatial reasoning tasks. As shown in section 4 these do not correspond to the patterns just outlined, and should therefore work directly against the emergence of the predicted patterns of spatial language use. Experimental participants are known to try to meet the experimenters' expectations (Orne & Whitehouse, 2000) and can be systematically primed by examples (Helfenstein & Saariluoma, 2007). Therefore, the example descriptions should substantially bias our informants towards using just those patterns that, according to our hypothesis, in subtle ways do not correspond to natural language usage. So if the patterns do nevertheless emerge in the data, this result would be stronger than without examples biasing towards different patterns.

We conducted two studies as follows. The aim of the first study was to gain insights on spontaneous descriptions for spatial models as such. For this purpose we asked the participants to provide two descriptions for each single spatial model. By asking for two descriptions we aimed to encourage creativity and raise awareness to the fact that any spatial configuration can be described in more than one way; again, this was intended to encourage speakers to deviate from the patterns we predicted for intuitive spatial language use. The second study addressed the requirement to be indeterminate, namely, to find a description that matches two spatial models at the same time, as is the case in traditional relational reasoning problems. For this purpose, participants were asked to provide one description that fits two models. We reasoned that this might induce the participants to slacken the principles typically employed for spatial description, since the descriptions needed to be flexible enough to support both models. The requirement to produce such indeterminate descriptions might lead to particular strategies of verbalization, diverging from those intended to fit just one model. In the following we first describe our study procedures and results, and then discuss both of them jointly.

5.1 Study 1: One Model – Two Descriptions

Participants. Eighteen German 12th grade high school students (age approx. 18 years; 6 of them male) and their teacher (female, age 35 years[1]) volunteered to participate in this study without payment.

Material, Design, and Procedure. We conducted a pencil-and-paper study. The material consisted of eight different spatial representations, presented to the participants in randomized order. The participants were told that the letters were supposed to represent names of fruit types (apples, dates, pears, etc.) and asked to describe the relative position of the fruits to each other for each of the models. The models were based on the same equidistant grid pattern as that assumed in relational reasoning studies. They were given the following example together with the German description:

```
        B

D   T   A
```

(7) The date is beside the grape, the apple is to the right of the date, the pear is not under the apple.
(German original: Die Dattel ist neben der Traube, der Apfel ist rechts von der Dattel, die Birne ist nicht unter dem Apfel.)

Here the syntactic structure of each spatial description corresponds to the strict syntactic format Locatum spatial-relationship Relatum. The example contains one occurrence of *not* and one of *beside* in order to introduce the option of using (directionally) underspecified spatial descriptions, as well as one description containing *to the right of* referring to a non-direct relationship. The trajectory is not orderly in the sense of a clear path through the model. Thus, participants should be encouraged (or biased) to use spatial language in just this way, deviating from the predicted patterns.

Participants were then asked to describe each of the eight models in two versions, each of which should correctly describe the model in its own way. However, descriptions did not have to be unambiguous (i.e., they might fit other possible

[1] This dataset would have to be excluded if there was a theoretical reason to assume age differences in spatial language use in such a setting, or if the data were found post-hoc to deviate from the remaining data (in which case the age deviation might have been a reason). Neither of these was the case; the data set was completely within the scope of the other results in all categories of analysis and thus did not affect the results in any particular way.

models as well), and each fruit represented in the model should be mentioned at least once. The participants could go through the material self-paced. Typically, they needed about 20 minutes altogether.

5.2 Linguistic Analysis and Results

The written data were segmented based on the spatial relationships described; each unit contained just one spatial relation. Thus, one syntactically complete sentence may be analyzed as two units: "rechts daneben eine Pflaume" (to the right of that a plum) "und darüber eine Birne" (and above that a pear). Segmentation was typically straightforward because descriptions were for the most part orderly and involved almost no hypotactic constructions such as "Der Apfel, unter dem die Pflaume liegt, befindet sich rechts von Orange und Kirsche" (The apple, which lies below the plum, is located to the right of orange and cherry). There were only five cases in which the description of one relationship was syntactically embedded in another; in those cases the embedded description was extracted into a separate unit. Altogether, 1228 units were collected. Next, we analyzed the data with respect to each of the linguistic factors described above, as follows.

Spatial Representation. We categorized the spatial terms used and identified the number of objects involved in each description. Apart from projective terms we identified occurrences of *beside, middle,* and *between* as well as negation. We distinguished between 1-object, 2-object, and multiple-object descriptions. For instance, "oben rechts ist die Birne" (on the top right there's the pear) refers to one object, and "rechts von ihr sind Orange, Dattel, Apfel und Pflaume" (to its right there are the orange, the date, the apple, and the plum) describes multiple objects. Descriptions that referred to a relatum implicitly, such as "darunter die Dattel" (there-below the date), were counted as involving two objects.

Most of the descriptions (79.32% of the 1228 units) referred to two objects (in relation to each other), and most (77.6%) contained at least one projective term. 8.14% referred to only one object, and 12.30% to more than two objects. 85.01% of the descriptions referring to two objects contained at least one projective term. Thus, describing the relationship of two objects using one projective term was a typical linguistic choice used by the participants.

However, there are other options that were also used regularly. 26.62% of all units contained "neben" (beside); 12.46% contained "neben" without a projective term (thus leaving the direction underspecified). 88.38% of all units containing "neben" referred to two objects. 44.44% of all occurrences of "neben" without a projective term referred to a position on the right side of the relatum, 30.72% on the left (the remaining ones referred to more complex relationships involving further objects). Here the variability between individuals was considerably high. Only 3.18% of all units contained "Mitte" (middle), 2.61% "zwischen" (between), and 3.66% of all units "nicht" (not).

8.88% of all units described groups of objects, i.e., more than one object as locatum, as in "P, A, B und D sind in einer Reihe" (P, A, B, and D are in one row). Of these, a clear majority (82.57%) used a left-to-right ordering of objects, as in the direction of reading.

Spatial Directness. We analyzed the extent to which 2-object descriptions referred to direct or indirect spatial relationships in the models. Results revealed that an overwhelming majority (95.69%) of the 974 units describing two objects referred to a direct relationship between the objects, i.e., the spatial term described the direct neighbor without any further objects or spaces in between. This fact was not made explicit except in just five occurrences of the term "direkt" (direct). Of those few units that did not describe direct relationships, there were two cases of a space in the grid pattern between objects described as *beside* each other, two cases of a diagonal relationship between objects described as *right* and *left* of each other, and 6 cases (0.62% of all units describing two objects) in which a further object was between the objects described as being *right* or *left* of each other. All of these 6 descriptions were produced by one single participant. In all other cases of a non-direct relationship, this fact was made explicit by using "nicht" (not), "diagonal", or some other relevant term. In three cases, the term "neben" (beside) was used together with "nicht" (not) when there was another object in between, i.e., the *beside*-relationship was described as *not* true. Altogether, it can be stated that, in spite of the example given to the participants, there was a very clear and strong tendency for 2-object descriptions to refer to direct relationships, contrary to the usage in relational reasoning studies.

Trajectory. In order to explore the extent to which participants' descriptions would adhere to certain patterns, we investigated the trajectory used in the descriptions, noting where the descriptions started and where they ended. Descriptions were investigated for both directions separately, left-right and top-bottom. Example (8) starts on the left top and ends on the right bottom of the given model (shown below the description):

(8) The mango is above the pear. Between the pear and the apple there is an orange. Below the apple there is a plum.
(German original: Die Mango ist über der Birne. Zwischen Birne und Apfel ist eine Orange. Unter dem Apfel ist eine Pflaume.)

M		
B	O	A
		P

There was a preference for starting on the left and ending on the right side of the model, as shown by almost half (45.64%) of the descriptions. In contrast, only 14.09% started on the right and ended on the left. For the vertical axis there was no such clear tendency; 20.13% started on the bottom and ended on the top of the model; and 25.17% started on the top and ended on the bottom. With respect to both axes, the remaining descriptions either ended up on the same side as they started, or they referred to the middle.

Syntactic Format and Information Structure. We investigated the allocation of the roles of locatum and relatum to linguistic structure as well as the general grammatical format. There was an overall tendency to start from one object, often without a relatum, and then move on from this (now known) object to the next (new) one. The evidence for this is as follows. 39.60% of all *initial* units lacked a relatum (to compare: 12.05% of *all* units lack a relatum). 76.45% of all non-initial units used an object as relatum that had already been mentioned earlier in the model description, i.e., a known object; but only 20% used an unknown object as relatum. The locatum, however, was only known in 11.51% of all non-initial descriptions. 10.97% of all non-initial descriptions related two or more unknown objects to each other.

22.23% of all units used the syntactic form SPATIAL-RELATIONSHIP LOCATUM, as in example (9) (where the implicit relatum is the model as a whole rather than another object). In contrast, 38.27% used the form SPATIAL-RELATIONSHIP RELATUM LOCATUM, as in example (10), and 32.17% used the form LOCATUM SPATIAL-RELATIONSHIP RELATUM, as in example (11). Only the last format corresponded to the examples given to them.

(9) On the top left there is the mango. (Oben links ist die Mango.)
(10) Below the pear there is the orange. (Unter der Birne ist die Orange.)
(11) The pear is above the plum. (Die Birne ist über der Pflaume.)

5.3 Study 2 – Two Models, One Description

Given the results of the first study, the impression emerged that people were not sufficiently encouraged by the examples to deviate from the predicted patterns of language use. We speculated that this might be due to an underlying aim to provide a complete description that could be used to unambiguously identify the configuration. People might pursue this aim in spite of the explicit statement in the instruction that this was not necessary, for example in order to make sense of the somewhat artificial task given to them. However, in relational reasoning studies, descriptions are generally underspecified so as to enable more than one correct solution. In order to account for this, we designed a second explorative study in which participants had to find one description that fit two models. Our aim, as before, was to explore if people would adhere to the predicted patterns of spatial language use in spite of encouragement to the contrary.

Participants. 16 speakers of German (14 high school and university students between 18 and 26 years with a mean age of 20, and two 60-year-olds[2]) volunteered to participate in this study without payment. None of the participants had taken part in the first study. Thus, we had a new chance to encourage speakers to deviate from the unconscious linguistic description principles identified in the spatial language literature.

[2] Again, these data sets were included in order to keep the database as large as possible, as we had no reason to assume a deviation because of age. The data sets were completely within the scope of the other results in all categories of analysis.

Material, Design, and Procedure. Materials and procedure were identical to the first study except that, this time, participants received two models at a time generated from an indeterminate description used in our earlier studies (Ragni et al., 2006; 2007). They were presented with the following example:

I.

		B
D	T	A

II.

	B	
T	D	A

(12) The date is beside the grape, the apple is right of the date, the pear is not under the apple.
(German original: Die Dattel ist neben der Traube, der Apfel ist rechts von der Dattel, die Birne ist nicht unter dem Apfel.)

This exemplifies a premise set as used in our studies leading to the two models above. All pairs of models were generated in this way by indeterminate premise sets. For each of the eight pairs of models given to them, participants were asked to write up to four sentences that accurately described both models at once, while avoiding phrases such as "either/or".

5.4 Linguistic Analysis and Results

The analysis was carried out in the same way as in study 1 as far as applicable. For example, we did not carry out a trajectory analysis since the two models to be described differed with respect to the position of a subset of the objects. Altogether, 488 units were collected. The following patterns emerged in the results.

Spatial Representation. As in the first study, describing the relationship of two objects using one projective term was a typical linguistic choice. Most of the descriptions (86.48% of the 488 units) referred to two objects (in relation to each other), 4.92% to only one object, and 8.40% to more than two objects. 82.79% contained at least one projective term. 83.89% of the descriptions containing two objects contained at least one projective term.

17.82% of all units contained "neben" (beside); 12.91% contained "neben" without a projective term. 89.66% of all units containing "neben" referred to two objects. 0.82% of all units contained "Mitte" (middle); 0.20% contained "zwischen" (between). 8.81% of all units contained "nicht" (not). 4.92% of all units described groups of objects, i.e., more than one object as locatum.

Spatial Directness. Again, there was a clear tendency for 2-object descriptions to refer to direct relationships, though not as strong as in the first study. 73.46% of the 422 units describing two objects referred to a direct relationship between the objects in *both* models. This was never made explicit; there were no occurrences of the term "direkt" (direct). 19.43% allowed for one or more objects between the locatum and relatum (in all except for three cases in just one of the models, but not the other). Furthermore, there were five cases of a space (in only one of the two models) between objects described as *right* of each other. Fourteen cases (3.32% of all units describing two objects) described a diagonal relationship between objects, eleven of these by using "nicht" (not) plus a projective term.

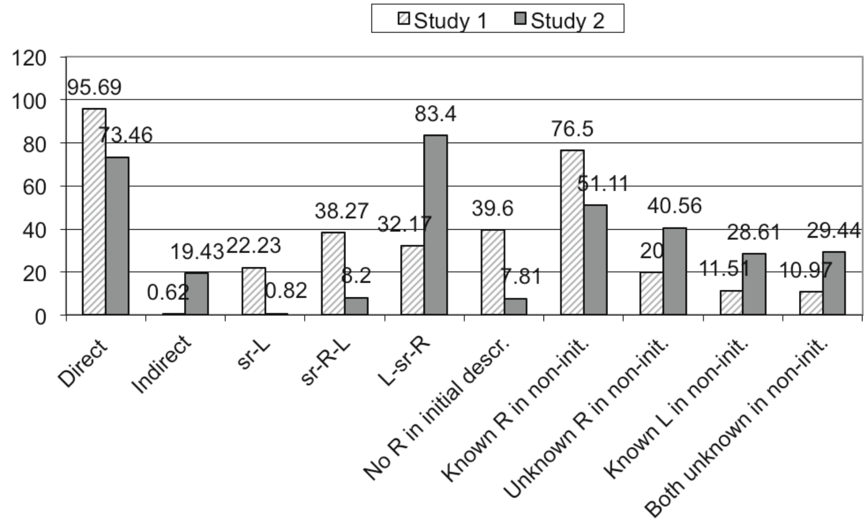
Syntactic Format and Information Structure. With respect to the distribution and givenness of spatial roles, the pattern of the first study did not emerge in this version. 7.81% of all *initial* units lacked a relatum, as compared to 6.56% of *all* units that lacked a relatum. 51.11% of all non-initial units used an object as relatum that had already been mentioned earlier in the model description, while 40.56% used an unknown object as relatum. The locatum was known in 28.61% of all non-initial descriptions. 29.44% of all non-initial descriptions related two or more unknown objects to each other. Thus, participants did not appear to start from one object and then move on to the next as in the previous study. Instead, the underlying strategy might have been to identify common spatial relationships that they could verbalize for both models at the same time, without a regular pattern of description.

The syntactic structure was rather uniform: 83.40% used the form LOCATUM SPATIAL-RELATIONSHIP RELATUM, as in example (11) above, corresponding to the example given to the participants. Only 8.20% used the form SPATIAL-RELATIONSHIP RELATUM LOCATUM, as in example (10), and 0.82% of all units used the syntactic form SPATIAL-RELATIONSHIP LOCATUM, as in example (9). These structures are represented in Table 2 below as L-sr-R, sr-R-L, and sr-L, respectively (see next section).

5.5 Qualitative Comparison of Results

Since our studies were designed to explore speakers' intuitive principles when describing spatial configurations of the kind used in relational reasoning research, we did not aim at a direct comparison between the results. Both studies had the same aim of exploring the extent to which speakers would deviate from the principles predicted from spatial language research. They differed with respect to the means by which we attempted to encourage such a deviation: in the first study, we aimed to support creativity by having participants produce two descriptions, and in the second study, we aimed to enhance underspecification by asking for one description for two models. A qualitative comparison is nevertheless informative at this point. In Study 2 participants seemed to be somewhat less reluctant to use projective terms to refer to indirect relationships between the described objects, there was a decrease in syntactic variation in Study 2 as compared to Study 1, and the distribution of known and new relata and locata varied considerably. These different tendencies are represented in Table 2, which illustrates the patterns found in the two studies summatively.

Table 2. Summary of patterns in Study 1 and Study 2



6 General Discussion

Robust empirical findings (e.g. Byrne & Johnson-Laird, 1989; Johnson-Laird & Byrne, 1991; Knauff et al., 1995; Rauh et al., 2005; Ragni et al., 2006; 2007) show that humans reason by the construction and manipulation of mental models. In order to examine the extent to which linguistic descriptions that are used in this line of research correspond to human natural language usage patterns, we proceeded in four steps. First, we summarized previous findings on spatial language usage with respect to systematic principles that speakers may not be aware of. Second, we analyzed the linguistic patterns of relational reasoning tasks, leading to a range of expectations concerning the ways in which spontaneous descriptions by naïve speakers might diverge from the tasks used in research settings. Third, we had participants write such descriptions without constraints, using an example intended to bias them towards the kind of linguistic patterns used for relational reasoning tasks. Rather than examining to what extent the resulting patterns were based on the example given to participants, we were interested in the *limits* of this bias. In our first elicitation study, additional conceptual flexibility was induced by producing two descriptions for one single configuration. Fourth, in our second linguistic elicitation study we introduced a further requirement of relational reasoning research, namely to produce indeterminate descriptions that are applicable for two spatial models at the same time.

A number of fairly clearcut patterns could be identified in our data (steps 3 and 4), confirming for the most part our expectations (motivated from the literature in steps 1 and 2) about (qualitative) tendencies in the use of spatial language. Some of these patterns correspond to the configuration descriptions used in relational reasoning tasks. In both studies there was a clear preference for descriptions of two objects in relation to each other using projective terms (rather than *beside* or negations, which they saw in the examples, or alternative spatial terms such as *middle* or *between*). Furthermore, in Study 1 three types of syntactic format were regularly employed, one of which corresponds to the form used in reasoning tasks, namely LOCATUM SPATIAL-RELATIONSHIP RELATUM. This syntactic format was predominant in Study 2.

However, our results also highlight principles of language use that were rarely broken in spite of the example given in the instruction, indicating the limits of the bias. Most crucially, the data collected in Study 1 revealed a very strong, almost unbroken preference for *direct* spatial relationships (i.e., involving no other objects between the two objects described), confirming earlier suggestions in this respect (Herskovits, 1986; Talmy, 2000), including real-world object configurations (Tenbrink et al., 2011). In Study 1, participants were decidedly reluctant to abandon this principle. Study 2 revealed that participants given the task of producing indeterminate descriptions were somewhat less reluctant to break the principle. Nevertheless, the preference for direct relationships was still overwhelming in these data.

It can be concluded from this result that a strong default assumption for natural language usage is that objects in a configuration are directly adjacent to one another, even if the instruction to study participants suggests the contrary. As a consequence, the position directly beside an object described using a projective term will be strongly preferred and remain prominent even in the face of conflict. This may constrain the ways in which preferred mental models are generated so as to compensate for limitations of working memory resources (Knauff, Rauh, & Schlieder, 1995). So far, the construction of preferred mental models has been explained by insertion principles in the visual spatial sketchpad, assuming that participants tend to keep the partially constructed mental models. Our results rather suggest that there may be a conflict already at an earlier stage than suggested by Knauff et al. (1995), namely in the premise interpretation phase. Each premise must be understood *before* it can be integrated in the mental model; this is difficult to achieve if the preferred location in the model is already filled. This conflict leads to the need for a re-interpretation of the spatial term according to the artificial definition in the experimental setting. This translation process has mostly been assumed to be automatic; however, this remains to be empirically tested. In fact, the extent to which the preferred interpretation of a direct adjacency of spatial relations is consciously accessible to speakers, and can therefore be easily and immediately changed according to the experimental requirements, is not known. As a consequence, a possible influence on the deduction process cannot be excluded. So far, the mechanisms involved in the re-interpretation of spatial terms have been largely disregarded in this respect. To avoid such conflicts in interpretation, the participants should receive more natural premise descriptions. If no translation process is necessary, the linguistic burden of (re-)interpretation can be reduced.

The trajectory preferences found in Study 1 reveal default assumptions about spatial patterns, and they highlight how humans scan through and analyze a built (mental) model, if there is only one to begin with. A continuous processing of the visual scanning had been already assumed in the literature (Rauh et al., 2005). Our results support this continuous process pattern and specify it in more detail, showing how the trajectories deviate from the structure used in relational reasoning tasks – again contrary to the example given in the instruction. In particular, Study 1 revealed a clear preference for orderly trajectories, starting from one object (often without relatum) and then moving from known to new objects step-by-step. This finding is consistent with results on object configuration descriptions in more naturalistic settings (Ehrich & Koster, 1983; Andonova et al., 2010). Also, there was a preference for left-to-right ordering (corresponding to the direction of reading in German) with respect to overall trajectories and descriptions of object groups. No corresponding preference for bottom-to-top or top-to-bottom was identified.

Study 2 revealed that the tendency towards using a continuous trajectory pattern appeared to be constrained by the discourse task. Given the task of formulating one description fitting two models, the participants focused on identifying joint spatial relationships, rather than producing continuous descriptions. As a consequence, these description patterns more closely resembled the premises in reasoning tasks (as in the examples given to our participants), diverging from the orderly trajectory and information structure based principles that are typical for descriptions of spatial configurations. These results shed new light on findings concerning the effects of continuity in spatial relational reasoning problems, for instance by Knauff et al. (1998). According to their analysis, continuous descriptions are easier because of the cognitive processes involved. These effects may be further supported by the nature of the discourse and its constraints, which provide a conceptual framework for the reasoning task. As with directness, the suggestion of a continuous description may well happen without the reader's conscious awareness. With discontinuous descriptions, the interpretation of ambiguous relationships may come into focus to an increased degree; i.e., a description perceived as discontinuous may subtly suggest an indeterminate model.

Altogether, although the surface of the linguistic descriptions clearly resembles natural language usage, the meaning to be conveyed by them – the spatial relationships described – do not correspond to the implicit pragmatic rules of language usage (and, quite possibly, default interpretation) identified here. The language used in N-term series experiments contradicts preferences and patterns in natural language usage in various ways. Participants of such studies are therefore confronted with an *artificial language* that they first need to learn to interpret, which requires additional mental effort. The fact that relational reasoning tasks deny fundamental principles of spatial language use may have a greater impact on the processing and reasoning involved when solving these tasks than has hitherto been acknowledged. To remedy this, future studies on spatial relational reasoning might use premise sets that are not associated with direct spatial relations, for instance by using negation ("C is not left of B" rather than "B is left of C"), and they might account for the natural preference for orderly, left-to-right trajectories as well as principles concerning information structure in premise formulation.

To conclude, future experimental studies in relational reasoning research should focus more on natural, intuitive rather than artificial relations, i.e., those resembling spontaneous descriptions used by humans in everyday life to describe spatial arrangements. This can avoid a bias in the results. Moreover, it may be advantageous to have participants generate the relation between objects, rather than asking them to validate conclusions. Conclusion generation problems should result in more natural results. To avoid linguistic load, and assuming that participants tend to analyze their internal models continuously, a continuous presentation of the premises is preferable, along with direct spatial relations between objects or formulations that do not intuitively suggest such a direct relationship.

Acknowledgments. Funding by the Volkswagen Foundation (Tandem grant to Jan Wiener and Thora Tenbrink) as well as by the DFG (SFB-TR 8 Spatial Cognition Bremen / Freiburg) is gratefully acknowledged.

References

Andonova, E., Tenbrink, T., Coventry, K.R.: Function and Context Affect Spatial Information Packaging at Multiple Levels. Psychonomic Bulletin & Review 17, 575–580 (2010)

Carlson, L.A., Hill, P.L.: Formulating Spatial Descriptions across Various Dialogue Contexts. In: Coventry, K., Tenbrink, T., Bateman, J. (eds.) Spatial Language and Dialogue, pp. 89–103. Oxford University Press, Oxford (2009)

Clark, H.H.: Influence of language on solving three-term series problems. Journal of Experimental Psychology 82(2), 205–215 (1969a)

Clark, H.H.: Linguistic processes in deductive reasoning. Psychological Review 76(4), 387–404 (1969b)

Clark, H.H.: Using Language. Cambridge University Press, Cambridge (1996)

Coventry, K.R., Garrod, S.C.: Saying, seeing and acting: The psychological semantics of spatial prepositions. Psychology Press, Hove and New York (2004)

De Soto, C.B., London, M., Handel, S.: Social reasoning and spatial paralogic. Journal of Personality and Social Psychology 2, 293–307 (1965)

Ehrich, V., Koster, C.: Discourse Organization and Sentence Form: The Structure of Room Descriptions in Dutch. Discourse Processes 6, 169–195 (1983)

Fodor, J.A.: The language of thought. Harvard, New York (1975)

Gerrig, R.J., Banaji, M.R.: Language and Thought. In: Sternberg, R.J. (ed.) Thinking and Problem Solving, pp. 233–261. Academic Press, San Diego (1994)

Habel, C., Tappe, H.: Processes of segmentation and linearization in describing events. In: von Stutterheim, C., Klabunde, R. (eds.) Processes in Language Production, pp. 117–153. Westdeutscher Verlag, Opladen (1999)

Halford, G.S., Smith, S.B., Dickson, J.C., Maybery, M.T., Kelly, M.E., Bain, J.D., Stewart, J.E.M.: Modeling the Development of Reasoning Strategies: The Roles of Analogy, Knowledge, and Capacity. In: Simon, T.J., Halford, G.S. (eds.) Developing Cognitive Competence: New Approaches to Process Modeling, pp. 77–156. Erlbaum, Hillsdale (1995)

Halliday, M.A.K.: An Introduction to Functional Grammar, 2nd edn. Arnold, London (1994)

Helfenstein, S., Saariluoma, P.: Apperception in primed problem solving. Cognitive Processing 8, 211–232 (2007)

Herskovits, A.: Language and Spatial Cognition. Cambridge University Press, Cambridge (1986)

Hörnig, R., Oberauer, K., Weidenfeld, A.: Between reasoning. Journal of Experimental Psychology 10, 1805–1825 (2006)

Huttenlocher, J.: Constructing spatial images: A strategy in reasoning. Psychological Review 75, 550–560 (1968)

Johnson-Laird, P.N.: Mental models and deduction. Trends in Cognitive Science 5, 434–442 (2001)

Johnson-Laird, P.N., Byrne, R.M.J.: Deduction. Erlbaum, Hillsdale (1991)

Knauff, M., Johnson-Laird, P.N.: Visual imagery can impede reasoning. Memory & Cognition 30, 363–371 (2002)

Knauff, M., Rauh, R., Schlieder, C.: Preferred mental models in qualitative spatial reasoning: A cognitive assessment of Allens calculus. In: 17th Annual Conference of the Cognitive Science Society, pp. 200–205. Erlbaum, Mahwah (1995)

Knauff, M., Rauh, R., Schlieder, C., Strube, G.: Continuity effect and figural bias in spatial relational inference. In: 20th Annual Conference of the Cognitive Science Society, pp. 573–578. Erlbaum, Mahwah (1998)

Levelt, W.J.M.: Linearization in describing spatial networks. In: Peters, S., Saarinen, E. (eds.) Processes, Beliefs and Questions. Essays on Formal Semantics of Natural Language and Natural Language Processing, pp. 199–220. Reidel, Dordrecht (1982)

Linde, C., Labov, W.: Spatial networks as a site for the study of language and thought. Language 51, 924–939 (1975)

Mani, K., Johnson-Laird, P.N.: The mental representation of spatial descriptions. Memory and Cognition 10(2), 181–187 (1982)

Ormrod, J.E.: Cognitive Processes in the Solution of Three-Term Series Problems. The American Journal of Psychology 92(2), 235–255 (1979)

Orne, M.T., Whitehouse, W.G.: Demand characteristics. In: Kazdin, A.E. (ed.) Encyclopedia of Psychology, pp. 469–470. American Psychological Association and Oxford Press, Washington, D.C. (2000)

Ragni, M., Fangmeier, T., Webber, L., Knauff, M.: Complexity in Spatial Reasoning. In: Sun, R., Miyake, N. (eds.) Proceedings of the 28th Annual Cognitive Science Conference, pp. 1986–1991. Erlbaum, Mahwah (2006)

Ragni, M., Fangmeier, T., Webber, L., Knauff, M.: Preferred Mental Models: How and Why They Are So Important in Human Reasoning with Spatial Relations. In: Barkowsky, T., Knauff, M., Ligozat, G., Montello, D.R. (eds.) Spatial Cognition V. LNCS (LNAI), vol. 4387, pp. 175–190. Springer, Heidelberg (2007)

Rauh, R., Hagen, C., Knauff, M., Kuß, T., Schlieder, C., Strube, G.: Preferred and alternative mental models in spatial reasoning. Spatial Cognition and Computation 5, 239–269 (2005)

Schleipen, S., Ragni, M., Fangmeier, T.: Negation in Spatial Reasoning: A Computational Approach. In: Hertzberg, J., Beetz, M., Englert, R. (eds.) KI 2007. LNCS (LNAI), vol. 4667, pp. 175–189. Springer, Heidelberg (2007)

Talmy, L.: Toward a Cognitive Semantics, 2 vols. MIT Press, Cambridge (2000)

Tenbrink, T.: Space, time, and the use of language: An investigation of relationships. Mouton de Gruyter, Berlin (2007)

Tenbrink, T., Coventry, K.R., Andonova, E.: Spatial strategies in the description of complex configurations. Discourse Processes 48, 237–266 (2011)

Van der Henst, J.B., Politzer, G., Sperber, D.: When is a conclusion worth deriving? A relevance-based analysis of indeterminate relational problems. Thinking and Reasoning 8, 1–20 (2002)

Vorwerg, C., Tenbrink, T.: Discourse Factors Influencing Spatial Descriptions in English and German. In: Barkowsky, T., Knauff, M., Ligozat, G., Montello, D.R. (eds.) Spatial Cognition V. LNCS (LNAI), vol. 4387, pp. 470–488. Springer, Heidelberg (2007)

Extended Verbal Assistance Facilitates Knowledge Acquisition of Virtual Tactile Maps

Kris Lohmann and Christopher Habel

Department of Informatics
University of Hamburg
Vogt-Kölln-Straße 30
22527 Hamburg, Germany
{lohmann,habel}@informatik.uni-hamburg.de

Abstract. We report on an experiment testing the VAVETaM (Verbally-Assisting Virtual-Environment Tactile Maps) approach for an intelligent multimodal tactile-map system, which was proposed to support blind and visually impaired people in acquiring survey knowledge. In the experiment, participants received two types of assisting utterances while exploring virtual tactile maps in a repeated-measures experiment: (1) only names of map objects and (2) additional information, for example, about spatial relations between the objects. The latter type of verbal assistance was similar to that which humans give when they are asked to verbally assist a map explorer. The virtual tactile maps were presented using a device for haptic human-computer interaction. The data indicate that the spatial knowledge map users acquire consists of two subtypes: knowledge of the structure of map entities that represent objects enabling locomotion (such as streets) and knowledge of the configuration of potential landmarks. Regarding both subtypes together, participants performed significantly better after learning the map with additional verbal information compared to receiving only information about the proper names of objects. A more fine-grained analysis shows that this improvement is only based on knowledge of the configuration of potential landmarks.

Keywords: Spatial Knowledge Acquisition, Virtual Tactile Map, Audio-Tactile Map, Maps for People with Low Vision, Multimodal Interface, Verbal Assisting Utterances, Virtual Haptics.

1 Introduction

Maps are a major means for providing spatial knowledge of the environment, such as for getting a first overview of a university campus or for planning a route in an unknown environment. Since blind and visually impaired people do not have any—or only unsatisfactory—access to visual maps, tactile maps are proposed as substitutes for acquiring spatial knowledge [3]. However, these maps lead, compared to visual maps, to disadvantages regarding speed and accuracy of the knowledge-acquisition process.

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 299–318, 2012.

Exploring a tactile map has to be done sequentially [1, 3]. Therefore, comprehending even a sparse map with relatively few details (such as used in the experiment reported) imposes large demands to integrate the stream of knowledge entities provided over time. In contrast, on a visual map of the same complexity, an overview of the map including objects and spatial relations among them is perceived with less effort, due to the highly parallel character of visual perception.

To overcome these problems, several multimodal systems that use sounds or prerecorded speech when objects on tactile maps are touched have been developed [e.g., 15, 30, 33, see 6 for a recent overview]. These systems improve the effectiveness of tactile map use, compared to a unimodal tactile map, by associating information-bearing sounds or descriptions to map objects. Our goal in developing VAVETaM (Verbally-Assisting Virtual-Environment Tactile Maps) is to further reduce or to overcome the drawbacks of tactile maps that are due to sequential haptic exploration. Therefore, we suggest extending audio-tactile map approaches by providing an extended set of assisting utterances given in natural language similar to what a human assistant can give.

1.1 VAVETaM: Verbally-Assisting Virtual-Environment Tactile Maps

With the VAVETaM system, we aim for a system that is capable of generating situated and coherent assisting utterances including further information such as relations between map objects. In the following, this type of assistance is called ‘extended assistance’. Generating extended assistance is enabled by observing and interpreting the map-exploration movements the user performs instead of associating areas or objects on the (virtual) tactile representation to fixed audio information. Therefore, the system is, in comparison to the existing approaches, capable to act more like an assisting human giving information to a visually impaired user of a tactile map [see 20, 13, 19, for a detailed discussion of the VAVETaM conception and 18, for a discussion of a prototype implementation of the language-generation components of the VAVETaM system].

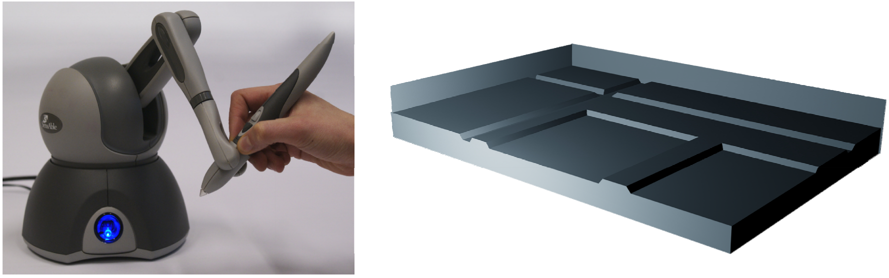


Fig. 1. (a) The haptic device used for the study and the VAVETaM system, (b) a cross-section through a model of a virtual tactile map (the borders at the back and left part visualize the frame of the map)

VAVETaM utilizes the Sensable Phantom Omni haptic device[1] for human-computer interaction to enable perception of the tactile map (see Fig. 1 (a)). This device is used in the experiment reported in Section 2. It consists of a pen-like handle attached to a moveable arm. Sensors register the position of the axis of the handle and the arm. Servomotors in the device enable the perception of virtual objects. A 3D model that is stored on a connected computer is used to calculate the forces generated by the servomotors in the arm depending on the current position of the handle. When the user moves the handle towards an area that is modeled as solid-object, a force is generated. As the device generates the force depending on the user's hand position, a virtual haptic perception is enabled, so the user can explore and feel virtual objects with impenetrable surfaces using the device. In the context of VAVETaM, the virtual objects model tactile maps, in which streets and potential landmarks, such as buildings, are marked as concave areas. A cross-section through such a virtual 3D map is shown in Fig. 1 (b).

Users explore the maps by performing movements on the surface of the map and especially by following (concave) lines representing streets and exploring the other map objects represented by concave regions with movements that enable them to detect the shape of the objects. In their sequential exploration, they focus haptically on objects or parts of objects about which they wish to acquire knowledge; for the respective dynamically determined region we use the term 'haptic focus' in the following [19].

1.2 Spatial Knowledge Acquisition Using VAVETaM

Increasingly, the benefits of different unimodal or multimodal sources for spatial knowledge have become in the focus of research. For example, Giudice, Betty, and Loomis [12] show evidence for the functional equivalence of spatial knowledge derived from haptic representations and visual representations. Giudice, Bakdash, and Legge [11] discuss that situated verbal information can potentially be helpful for acquiring spatial knowledge in the absence of vision in large-scale indoor environments that are learned by direct exploration. Blades [3] and Espinosa and colleagues [9] discuss empirical evidence that tactile maps can be an efficient means to communicate spatial knowledge.

VAVETaM are intended to provide blind and visually impaired people with a more adequate means for acquiring spatial knowledge of their environment. The goal of the experiment presented in Section 2 is to test whether extended verbal assistance facilitates the process of acquiring spatial knowledge from tactile maps. Humans' spatial knowledge of their environment is highly differentiated. A well-established distinction is that between route knowledge and survey knowledge [cf. 26, 32, however, see 23 for an alternative approach]. Route knowledge is characterized as sequential and view-point-dependent knowledge whereas survey knowledge is viewpoint-independent knowledge providing an overview about a real-world area.

[1] Phantom devices are manufactured by Sensable (`http://www.sensable.com`) and especially used in the area of 3D modeling and manipulation.

Maps and tactile maps are especially adequate to gain survey knowledge [cf. 29]. Survey knowledge is needed in a multitude of scenarios—for example, to plan a route that later can be learned from the map. Hence, our research focuses on the acquisition of survey knowledge, which is represented as spatial mental model by the map user (the term ‘spatial mental model’ is used in this paper to avoid implications of the term ‘cognitive map’ [28]).

In order to test spatial mental models that participants gained by their exploration of the multimodal maps, appropriate tests have to be applied. To our knowledge, no standard assessment methods for survey knowledge exist [see 24, 16, for an overview of testing methods for spatial knowledge]. We used different, mutually supportive tests [as suggested by 14, 16]. Three tasks were designed and used after each map-learning condition, which are described in Sections 1.3 and 2.2. The first task consisted of answering a number of questions concerning the spatial layout of and between objects on the map. In a second task, participants were asked to produce a sketch map. The third task was a recognition test, in which participants had to determine correct parts of a visualization of the map they had explored among different incorrect ones. Section 2.2 discusses the tests used in detail.

1.3 Goals of the Experiment

We evaluate the potential of using situated assisting utterances for a human computer interaction system in a Wizard-of-Oz-like experiment [see 8, for a discussion of Wizard-of-Oz studies]. We tested two types of verbal assistances to tactile map explorations, which differed with respect to provided information. In the baseline condition, called ‘simple-assistance condition’, verbal utterances that only inform about proper names of the haptically explored objects were given. The type of corresponding utterances is similar to those proposed in other multimodal systems [e.g., 6, 15, 30]. This type was compared to the knowledge gained in the extended-assistance condition, which included additional information. In the extended-assistance condition, participants additionally received information about map objects that do not carry proper names (such as intersections) and information about relations between map objects (e.g., which buildings are located close to a street).

Our hypothesis was that verbal utterances in the extended-assistance condition and the perception of the virtual tactile map are successfully integrated in the process of spatial-knowledge acquisition. Therefore, more precise spatial mental models are acquired compared to spatial mental models build up by learning the tactile map under the simple-assistance condition.

2 Method

2.1 Participants

Twenty-four participants initially participated in the experiment. Two participants had to be excluded due to technical problems. One was not able to pass the test that concluded the training procedure and was therefore also excluded. Data of three additional participants were collected, leading to a total of 24 evaluated participants (14 males, mean age: 24.7 years, *SD*: 3.3 years). All participants were compensated by partial course credit or on a monetary basis. They were naïve about the purpose of the experiment. All participants gave written informed consent and reported to speak German on a native-speaker level. All participants used their self-reported primary hand for the exploration of the virtual tactile map.

The choice for a setting with blindfolded sighted participants rather than with blind or visually impaired participants was made for two reasons. Firstly, reading a map is a complex skill that both, sighted and visually impaired people have to learn [cf. 22, 16]. Blind and visually impaired people are not always familiar with maps. Since the effects of (un)familiarity with map-like representations cannot be anticipated, testing blindfolded sighted participants who are familiar with maps ensures homogeneity among the participants. Secondly, blind and visually impaired people are a small group and, with respect to their visual experience, heterogeneous group. This group is already to a large degree involved in experimental testing. Besides providing insights concerning the integration of haptic and natural language representation by sighted participants, one goal of the experiment reported was to test and refine the experimental methods before testing visually impaired people. A follow-up experiment with blind and visually impaired people is ongoing.

2.2 Materials and Procedure

Material. Two different virtual tactile maps and corresponding utterances with two different sets of names were created to avoid carry-over effects in the repeated-measures design. The utterances were started by the experimenter using a custom-built interface.

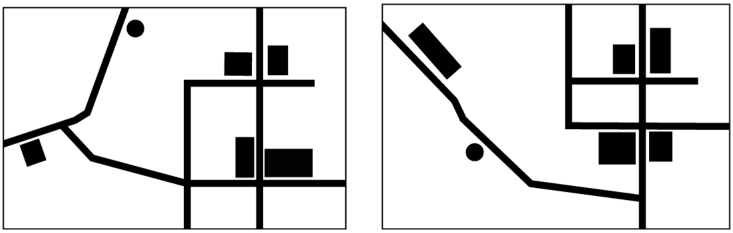


Fig. 2. Visualization of the two maps (map 1 to the left and map 2 to the right) used in the experiment

Maps. The virtual tactile maps were haptically explored by the participants using the Sensable Phantom Omni device attached to an Apple iMac. The maps were of similar complexity regarding the number of intersections, the amount of parallel streets, potential landmarks, and dead ends. Both maps included five tracks ('track' is a more general term for street-like structures and similar to the term 'path' introduced by Lynch [21]) and six potential landmarks (one tree and five buildings). See Fig. 2 for a visualization of the maps used. A pre-study showed that visualizations of the maps could be completely remembered when learned visually.

The maps were modeled for haptic interaction using Autodesk's 3D Studio Max. The 3D models were presented with the Phantom device using the Sensable OpenHaptics toolkit[2].

Assisting Utterances. The assisting utterances were recorded before the study. They were given in German, spoken by a 26-year-old male native German speaker. The participants heard the utterances via headphones.

Two name sets were created that could be used for both maps (the amount of tracks and potential landmarks were identical on both maps). The name sets consisted of names for the tracks, i.e., street names (S), and for potential landmarks. The later were of the following types: (LM1) names signifying the function of the potential landmark, (LM2) individual names such as brand names for chains of stores, and (LM3) class names. See Table 1 for an overview of the name sets used.

Table 1. Name sets used in the experiment

	Set 1	Set 2
S	*Poststraße, Humboldtstraße, Lärchenweg, Goethestraße, Hegelstraße*	*Hochstraße, Dorfstraße, Amselweg, Blumenstraße, Bergstraße*
LM1	*Hauptbahnhof [main station], Universität, Christuskirche, Bertolt-Brecht-Schule*	*Rathaus [town hall], Gedächtniskirche, Anne-Frank-Schule, Museum*
LM2	*Aldi*	*Lidl*
LM3	*Eiche [oak]*	*Buche [beech]*

We developed a set of assisting utterances inspired by utterances that occurred in a corpus of human assisting utterances. To collect this corpus, several human participants were asked to verbally assist a blindfolded map user. The assistant saw a visualization of a blindfolded map user's exploration of a virtual tactile map of the type described above. The visualization of the tactile map that was shown on a computer screen was similar to Fig. 3. A red dot moving corresponding to the map user's exploration movements visualized the haptic focus of the map user for the assistant. The assistant was instructed to help the blindfolded map user comprehending the tactile map. Both, the assistant and the map user, were informed that only the assistant should talk; that is, that the co-operative action should not be performed in a dialogical manner. The frequently occurring utterances in the corpus were grouped to messages classes. Message classes are defined by the type of information of the corresponding utterances [19]. In the context of the present paper, the identification message class is particularly important. By stating an utterance of

[2] `http://www.sensable.com`

the identification message class, the assistant informs the user of the tactile map about the identity of the map object that is explored. Usually, this is done by stating the proper name of the map object in combination with a demonstrative. If an object (e.g., an intersection) does not have a proper name, it can often be identified by referring to objects with proper names. For the intersection example, these are the names of the streets that form the intersection (see example (3a) below). A detailed discussion of the other message classes is out of the scope of this paper [see 19, for a discussion]. In the following, the assisting utterances that were included in the two different assistance conditions are discussed.

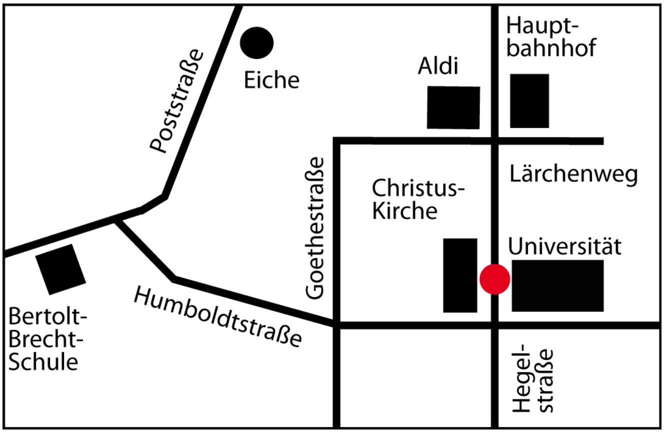


Fig. 3. Visualization of one of the maps used with one of the name sets; the dot between the buildings 'Christuskirche' and 'Universität' indicates a map user's exploration position

In the *simple-assistance condition* that provided the baseline for the described study, information about the names of objects in the haptic focus was given. Only utterances of the identification message class for objects with proper names were included. Consequently, no information for map objects without a proper name was given. In the maps used for the study, this affected intersections and dead ends, which were not verbally identified in the simple-assistance condition. Example (1a) is a translation of the assisting utterance that was given when the track 'Hegelstraße' was explored. Examples (1b) and (1c) are translations of identification messages for the frame of the map and the building called 'Bertolt-Brecht-Schule'. Note that utterances that use deictic reference are time critical; that is, they should only be given when the map user actually explores the object that is talked about. For example, a human assistant would give the assistance (1a) when the user is exploring the right track parallel to the map frame. Figure 3 shows a position on this track with a dot. Considering this position is the map user's exploration position, giving utterances such as (1b) or (1c) would be inappropriate.

(1a) This is Hegelstraße.[3]

(1b) This is the left map frame.

(1c) This is Bertolt-Brecht-Schule.

In addition to utterances such as (1a)–(1c), information that a human assistant would potentially include was given in the *extended-assistance condition*. (2a)–(3c) are examples for translations of assisting utterances given in this condition for track objects. The assisting utterances (2a)–(2d) are suitable when a user explores the track 'Hegelstraße', for example, at the position marked with the dot in Fig. 3. As can be seen from the examples, the extended-assistance condition included verbal information about the extent of tracks (i.e., what determines the end of a track, see example (2b)). The set of utterances included information about the intersections a track had (2d) and information about spatial and geometric relations with other tracks and landmarks (see examples (2a), (2b), and (2c)). These utterances, which do not include deictic references, are not as time critical as the ones containing deictic references; however, they were only given when the participant explored the part of the map that they were about, as well.

(2a) Hegelstraße is parallel to Goethestraße.

(2b) Hegelstraße ends to the left at an intersection with Lärchenweg and to the right in a dead end.

(2c) Below Hegelstraße, there are Rathaus and Museum.

(2d) Hegelstraße intersects with Goethestraße.

Furthermore, for parts of tracks that were close to landmarks or between landmarks, assisting utterances were given that stated this relation (see (3c)). Additionally, intersections and dead ends were identified (see examples (3a) and (3b)).

(3a) This is the intersection between Goethestraße and Humboldtstraße.

(3b) This is the dead end that forms the right end of Lärchenweg.

(3c) Now, you are between Christuskirche and Universität.

(4a) and (4b) are translations of assisting utterances for a potential-landmark object. For these objects, the set of utterances in the extended-assistance condition included utterances that stated the relation to other map objects (see (4a)) and, if appropriate, the global location in the map (see (4b)).

(4a) Bertolt-Brecht-Schule is located below Poststraße.

(4b) Bertolt-Brecht-Schule is located in the upper part of the map.

Control of the Assisting Utterances. The experiment was performed as Wizard-of-Oz-like experiment. The experimenter started the playback of the utterances using

[3] The purpose of the translations in this paper is to illustrate the content of the examples. We purposefully ignore article conventions.

custom-built software developed for this purpose. The experimenter looked at an extended visualization (similar to Fig. 3) in which, additionally, buttons were located close to the map objects. As in the corpus-collection study, a dot moving corresponding to the participant's exploration movements visualized the haptic focus of the participant for the experimenter. With these buttons, the experimenter was able to start the assisting utterances of the classes described above. Both, the map objects and the corresponding buttons, were represented in the same color. The experimenter started an utterance when the participant explored the corresponding map object.

Pre-studies indicated that if a map object is explored, giving information of the identification message class should precede any other information. Hence, the experimenter started an utterance of the identification class prior to any other utterance in the extended-assistance condition. The other utterances for that object were given in the order that the experimenter found most appropriate. However, we avoided unnatural repetitions of utterances. To facilitate this, the buttons for utterances that were played once were marked in the interface.

Assessment Methods. Following the map exploration with one of the assistance conditions described, the spatial knowledge that the participants gained by the exploration of the map was tested using three methods: (1) asking questions about spatial relations of objects on the map, (2) sketch mapping, and (3) a task in which participants needed to recognize correctly visualized parts of the map in a puzzle-like setting.

Relation Questions. First after the map exploration, participants answered questions concerning the spatial layout of the map. A similar approach has been used previously to test spatial mental models resulting from different types of spatial descriptions [4, 5, 27]. The first test consisted of 20 questions on spatial relations between objects. Where it was possible, these questions asked for relations between objects that were not explicitly stated in the prerecorded assisting utterances for the extended-assistance condition. The experimenter asked the questions in an individual random order. 10 questions involved spatial relations including potential landmarks and 10 were only about the track configuration. The answering options were 'yes', 'do not know', and 'no'. 10 questions were answered correctly with a 'no' and 10 questions with a 'yes'. A correct 'yes' and correct 'no' were evaluated as correct answers. A wrong 'yes', a wrong 'no', and 'do not know' were evaluated as wrong answers.

In the experimental procedure, each set of names occurred with each map. Therefore, sets of relation questions were developed for each combination of a map and a name set. This resulted in four sets of 20 questions. The questions of these sets were matched with each other. For example, question 1 always asked for the relation of two landmarks that were relatively distant to each other and was always correctly answered with 'yes' and question 17 always asked if two streets are parallel and was always correctly answered with 'no'. Two sets of questions were created for each map. This resulted from the use of two different sets of names. The individual questions for each set of names always asked for a spatially equivalent fact. For example, in the first map for the first name set the first question was 'Is Eiche left of Hauptbahnhof?' The map object that was called 'Eiche' in the first name set was called 'Buche' in the second name set. The map object that was called

'Hauptbahnhof' in the first set was called 'Anne-Frank-Schule' in the second set. Consequently, the first question for the first map with the second set of names was: 'Is Buche left of Anne-Frank-Schule?'

As reported in Sect. 2.3, a subsequent analysis revealed that the questions involving potential landmarks and those involving only tracks constitute two subscales.

Sketch Task. After completing the relation-questions, participants were asked to sketch the map on a sheet of paper. For a discussion of the validity and reliability of sketch maps as assessment methods for spatial knowledge see Lohmann [17] and Blades [2]. The frame defining the dimensions of the map was printed on the paper for sketching.

All 48 sketches were evaluated by the researcher and an independent rater. Both raters were uninformed about the condition in which the sketch map was produced and about whether sketch maps were created by the same participant. The rating was performed in two respects, reflecting the two knowledge types identified in the principal component analysis for the relation questions: Firstly, raters evaluated how well the sketch resembled the original map concerning the course of tracks, their parallelism, and the junctions they have. Secondly, raters evaluated how well potential landmarks were represented at their correct positions. The rating was performed on a 5-point Likert-type scale. A rating of 1 is associated with 'does not reflect the original' and 5 is associated with 'reflects the original precisely'.

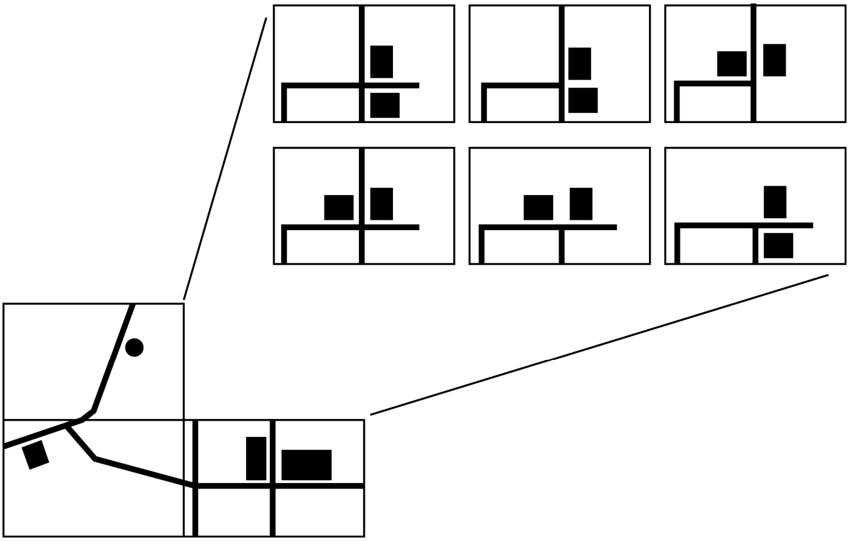


Fig. 4. Example for the options for the upper-right puzzle part

Recognition Puzzle. The third test performed was a recognition test. A visualization of the map that was explored previously was split into quadrants. Participants were given a set of possible map parts for each quadrant and were asked to decide which one is correct.

The goal positions and orientation of the parts were given. For each position, there were six options: the correct solution and five parts with a wrong spatial layout of a track and/or potential landmark. Each potential part fitted to each potential part of the other quadrants: there were no potential landmarks or tracks on any part that would have led to an inconsistent picture when combined. Refer to Fig. 4 for an example of the options for the upper-right part of the visualization of the map.

Procedure. The experiment was performed as repeated-measures experiment with the assistance condition as within-subject variable. This design was chosen to overcome the problem of individual differences in the ability to understand virtual tactile maps resulting from differences in spatial abilities biasing the results (see Wen, Ishikawa, and Sato [32], for a discussion of individual differences in spatial abilities).

To avoid carry-over effects, two different maps and sets of names were used for the different experimental conditions. To control for order effects that bias the main research question, the experiment was fully counterbalanced concerning the order of assistance conditions, maps, and the set of names. Accordingly, the second learning procedure was performed with the map, the assistance condition, and the set of names that were not used in the first condition. Consequently, the extended-assistance condition occurred as often as first learning condition as the simple-assistance condition. Furthermore, the first map and the first set of names occurred as often as the second map and the second set of names. Following this principle, each assistance condition was performed as often with each map and each set of names as the other one.

To ensure that participants understood the interaction with the multimodal system, they were trained in different aspects of it. Firstly, they were introduced to the haptic device and explored some standard examples of the Chai 3D[4] haptic toolkit. Then, they were interactively introduced in virtual tactile maps. Therefore, they explored a map, first assisted by the experimenter who gave assisting utterances and then by prerecorded utterances controlled by the experimenter, as used in the experimental conditions. The map used for training purposes was different from the two maps used in the experimental conditions. In the training procedure, the map objects had artificial names such as A-Building or Alpha-Street. To conclude the training, participants were tested for their ability to identify the shape of objects (such as a triangle and a square) and to follow a complex track structure without leaving it while receiving and following assisting utterances given by the experimenter. The training procedure including the training test took 30–60 minutes.

After a short break, participants read written instructions about the time they had for exploration and the tasks they had to solve after learning. Information about the tasks was included to minimize order effects resulting from knowledge of the tasks from the first condition. For the same purpose, an example map similar to the ones used in the experimental conditions was printed on the instructions. Participants were instructed to read the map in a way that they would find the route from each potential landmark to each other.

[4] http://www.chai3d.org

Each exploration was limited to eight minutes. After the first exploration, first the relation-questions task was performed, then participants created the sketch map, and finally did the recognition-puzzle task. A ten-minute break was made before the participants performed the second exploration with the same time constraint as the first exploration.

2.3 Results

Relation Questions. Taking all questions into account, the average result showed that participants were able to correctly answer significantly more questions ($t(23) = 8.08$, $p < .001$) when they learned the map under the extended-assistance condition ($M = 14.04$, $SE = .61$) than when they learned the map in the simple-assistance condition ($M = 8.46$, $SE = .42$) and the effect is large ($r = .86$) [according to 7].[5]

Table 2. Component loadings of the questions on the two knowledge dimensions

Question Number	Component 1 (Landmark Knowledge)	Component 2 (Track Knowledge)
1	**.776**	.125
2	**.784**	–.038
3	**.669**	.121
4	**.546**	.246
11	**.832**	.028
12	**.656**	–.029
13	**.749**	.382
14	**.685**	.426
15	**.684**	.247
6	–.319	**.483**
7	–.362	**.334**
8	.298	.020
9	–.204	**.527**
16	–.380	**.669**
17	–.071	**.384**
18	–.175	**.261**
19	–.204	**.335**
20	–.342	**.643**

In the subsequent analysis, question 5 was excluded due to a somewhat unclear formulation of this question. For the following analysis of the data, we assumed that the set of names and the maps do not have an effect on the items (no significant effects of the map or the set of names used were found in the analysis of the resulting 19 questions).

[5] An α-level of .05 was used for all calculations reported in this paper.

Screening the data, it was obvious that the assistance condition had a strong effect on some but not on all questions. The effect on the answers of questions involving potential landmarks differed from those questions about tracks. Whereas the answers to the landmark-related questions showed large differences between the conditions, the answers to the questions that only involved track knowledge did not. To support this theory, a principal component analysis was calculated. Two components were extracted, corresponding to the two types of questions. We assumed that the components are not independent. Therefore the rotation method we chose was oblique rotation. Interestingly, there was a negative correlation between the components (component 1 correlated with −.202 with component 2). The component loadings are shown in Table 2. Those questions involving potential landmarks loaded highly on component 1, those that do not involve knowledge of potential landmarks loaded highly on component 2. We found no explanation for the fact that question 8 did not load as expected on component 2, which reflects track knowledge. The factors together explained about 40.59 percent of the variance.

Table 3. Translation of the questions for map 1 with the first set of names constituting the landmark-knowledge subscale

Question Number	Translation of the Question
1	Is Eiche left of Hauptbahnhof?
2	Is Eiche left of Hegelstraße?
3	Is Bertolt-Brecht-Schule left of Christuskirche?
4	Is Hauptbahnhof above Universität?
11	Is Eiche right of Christuskirche?
12	Is Hegelstraße left of Bertolt-Brecht-Schule?
13	Is Bertolt-Brecht-Schule above Eiche?
14	Is Hauptbahnhof below Aldi?
15	Is Hauptbahnhof right of Universität?

Table 4. Translation of the questions for map 1 with the first set of names constituting the track-knowledge subscale

Question Number	Translation of the Question
6	Do Humboldtstraße and Poststraße form a T-intersection?
7	Are Goethestraße and Hegelstraße parallel?
8	Does Lärchenweg form a dead end to the right?
9	Do Hegelstraße and Lärchenweg form an intersection?
10	Do Goethestraße and Lärchenweg meet?
16	Do Poststraße and Hegelstraße form a T-intersection?
17	Are Humboldstraße and Goethestraße parallel?
18	Does Goethestraße form a dead end?
19	Do Poststraße and Lärchenweg form an intersection?
20	Do Humboldtstraße and Lärchenweg meet?

For the two resulting factors, two subscales were constructed. Since the questions involving landmarks differed from those involving only tracks, these subscales are named the ‘landmark-knowledge subscale’ and the ‘track-knowledge subscale’. Table 3 contains translations of the questions of landmark-knowledge subscale with the remaining 9 items. Table 4 contains translations of the questions of the track-knowledge subscale. Figure 5 shows the mean number of correct answers after the learning conditions and for the two subscales. The figure indicates that extended assistance led to a larger amount of correctly localized landmarks. In contrast, the knowledge of the track structure was not affected by the learning condition.

Separate mixed-design ANOVAs were calculated for each of the subscales. In both subscales, there were no significant main effects of gender, handedness, the order of conditions, the order of the maps used, or the order of the set of names. Therefore, these variables were stepwise excluded from the model for further analysis.

Concerning the track subscale, participants were not able to answer significantly more questions after learning the map under the extended-assistance condition and there were no interaction effects. A paired-sample t-test of equivalences was used to check whether the number of correct answers on the track subscale can be considered equivalent [31]. In the analysis, we used a liberal symmetrical equivalence interval of .50 and an alpha level of .05 (two tailed), resulting in the following t value: $t = .23$. This value is lower than the critical constant, therefore, the number of answers can be considered statistically equivalent among the conditions.

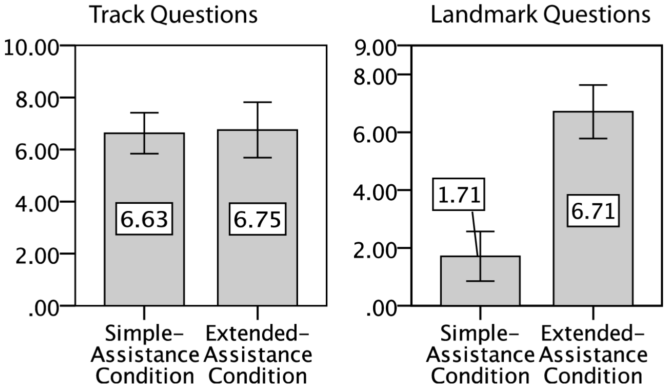


Fig. 5. Mean number of correct answers of the track subscale and the landmark subscale of the relation-question task (Error bars represent the 95% confidence interval of the mean)

In contrast to the equivalent number of correct answers on the track-knowledge subscale in the two conditions, there was a highly significant and large effect of assistance condition on the amount of correct answers for the landmark-knowledge subscale ($F(1, 23) = 80.23$, $p < .001$, $r = .88$[6]). Participants were able to answer significantly more questions correctly after having learned the map with the extended assistance condition.

[6] All effect sizes reported are based on planned contrasts [10].

Sketch Mapping Task. Two raters evaluated the sketches. The intraclass-correlation coefficient was calculated to test the agreement between the raters. The track ratings (*ICC(3,1)* = .68) and the potential-landmark ratings (*ICC(3,1)* = .72) agreement was calculated separately [25]. Both provided a fair agreement. The values of the raters were averaged for further analysis. Figure 6 shows an example of a participant's sketch map.

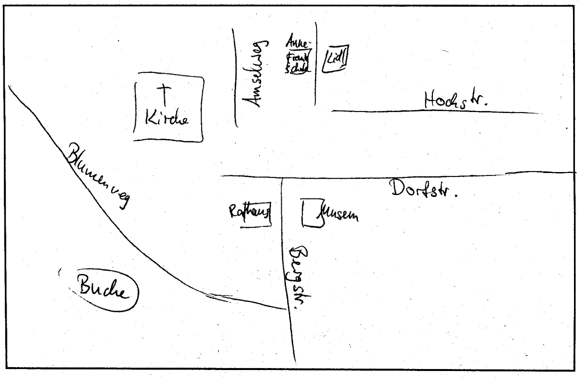


Fig. 6. Example of a participant's sketch of map 2 after learning the map under the extended-assistance condition

Figure 7 shows the mean ratings of the sketches. The figure suggests that potential landmarks depicted corresponded better to the tactile map when it had been learned under the extended-assistance condition.

Separate mixed-design ANOVAs were calculated for the landmark-configuration ratings and the track-structure ratings. In both ratings, there were no significant main effects of gender, handedness, the order of the maps used, the order of the conditions, or the order of the set of names. Therefore, these variables were stepwise excluded from the model for further analysis.

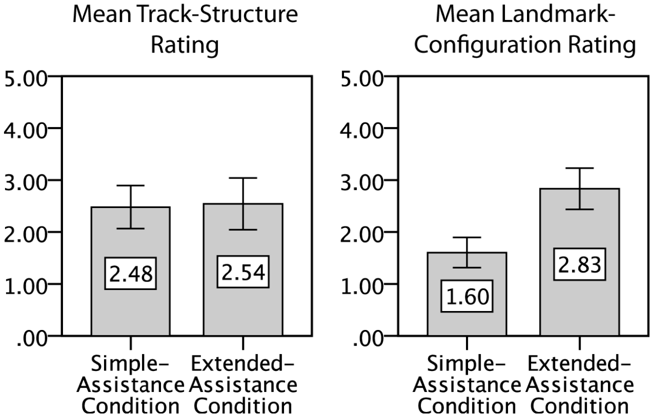


Fig. 7. Mean ratings of the track structure and the landmark configuration on the sketches (Error bars represent the 95% confidence interval of the mean)

Concerning how well the track structure as sketched corresponded to the tracks in the tactile map, no significant effects of the assistance condition were present. As for the analysis of the results of the relation questions, a paired-sample t-test of equivalences was calculated to control whether the mean ratings can be considered equivalent. Again, we used a liberal symmetrical equivalence interval of .50 and an alpha level of .05 (two tailed), resulting in the following t value: $t = .24$. This value is lower than the critical constant, therefore, the mean ratings of how well the track structure is sketched can be considered statistically equivalent between the conditions.

In contrast, ratings for how well potential landmarks are depicted corresponded to landmarks on the tactile map were significantly affected by the assistance condition. Participant sketched potential landmarks better when the map was learned with the extended-assistance condition ($F(1, 23) = 21.39$, $p < .001$, $r = .69$).

Recognition Puzzle. As described, each part of the recognition puzzle except from the correct part had either an incorrect part of a track or an incorrectly placed landmark, or both. We evaluated separately how well the solution of the participants reflected the structure of the tracks and the position of potential landmarks. The map was split into quadrants. For each quadrant six options were given: the correct solution and five parts with a wrongly depicted track structure and/or incorrectly placed landmarks. From the number of mistakes, we calculated the number of puzzle parts that showed the correct track structure and the number of puzzle parts that showed the correct landmark positions for each participant.

Figure 8 shows the number of parts that correctly reflected the track structure and the landmark configuration under each assistance condition. The figure indicates that participants benefited from extended assistance concerning their acquired knowledge of the landmark configuration. The figure also shows a slight improvement of the mean for the choice of parts that correctly reflect the track structure.

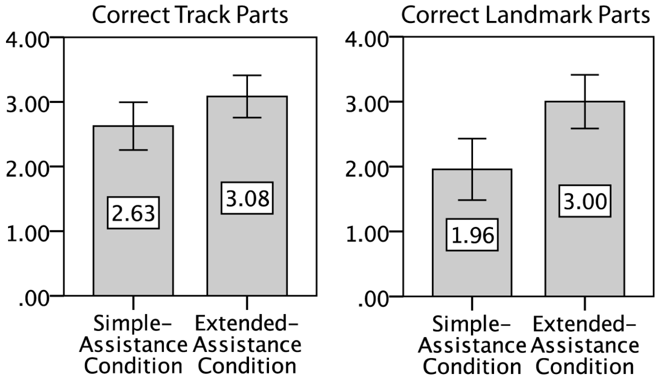


Fig. 8. Mean number of correctly identified puzzle parts (Error bars represent the 95% confidence interval of the mean)

Two separate mixed-design ANOVAs were calculated for the number of parts chosen that correctly reflected the track layout and the number of parts chosen that correctly reflected the landmark layout.

Concerning the number of parts that correctly showed the track structure, no main effects of gender, handedness, the order of the maps used, the order of the conditions, or the order of the set of names were present. Therefore, these variables were stepwise excluded from the model for the further analysis. Like in the two other tests, there was no significant effect of assistance condition. However, there was a tendency suggesting that participants performed better under the extended-assistance condition ($F(1, 23) = 3.87$, $p = .061$). Consequently, the paired-samples t-test of equivalence, using the same equivalence interval (.50) and alpha level (.05) as above, resulted in a value that is above the critical constant ($t = 1.97$).

The number of parts that correctly reflect the landmark layout showed a significant main effect of gender ($F(1, 20) = 8.20$, $p < .05$, $r = .54$). Males chose more elements with a correct landmark configuration. Furthermore, there was a significant interaction effect of gender and the order of assistance conditions ($F(1, 20) = 6.39$, $r = .49$, $p < .05$), males, in contrast to females, performed better when the extended-assistance condition preceded the simple-assistance condition. However, Levene's test indicated that the assumption of equal error variances was broken ($p = .002$). The following analysis was performed taking effects of gender, the ordering of the conditions and their interaction into account.

The data show a significant and large effect of the assistance condition ($F(1, 20) = 7.74$, $p < .05$, $r = .53$). Participants chose a higher amount of puzzle parts with a correct landmark setup after having learned the map under the extended-assistance condition.

3 Discussion and Summary

The experiment was performed to test whether a multimodal system that gives extended verbal assisting utterances for explorations of virtual tactile maps facilitates spatial knowledge acquisition and by this, improves existing multimodal approaches. We asked participants to explore maps in two experimental conditions: one with a restricted set of verbal information (simple assistance) and one condition with additional verbal information (extended assistance). The verbal information included in the latter condition was inspired by the information human assistants gave in a corpus study that we had previously made.

As expected, the experiment reported shows that, altogether, spatial knowledge is acquired more efficiently when (virtual) tactile maps are explored with extended assistance than when they are explored with simple assistance.

The data show that knowledge gained from virtual tactile maps consists of two subtypes: (1) knowledge of the track structure ('track' is a term for entities enabling locomotion, such as streets) and (2) knowledge of the configuration of potential landmarks. Contrary to our expectations, the overall increase in knowledge is only based on an increase in knowledge of potential landmarks. While the data show a strong increase in knowledge of potential landmarks, participants did not acquire significantly more knowledge of the track structure. The number of correct answers to questions about coarse-grained spatial facts about tracks (relation questions) and the mean ratings of the track structure reflected in sketch maps is statistically equivalent between the two assistance conditions (shown by paired-sample t-tests for equivalence).

In addition to the main finding that learning virtual tactile maps with additional verbal assistance significantly facilitates acquisition of knowledge of potential landmarks, the study also indicates that all three tests are applicable to assess survey knowledge acquired by tactile map explorations, especially the relation questions and sketch-mapping task, which have shown a perfect match concerning both statistically significant difference and equivalence. All three tests support the same result, except from the fact that the data for knowledge of tracks of a recognition task (recognition puzzle) are not statistically equivalent among the experimental conditions.

The experiment reported was performed in the context of the development of the VAVETaM system, which is intended to help blind and visually impaired people by providing efficient multimodal external survey-knowledge representations. Therefore, an experiment with blind and visually impaired people and adapted tests of spatial survey knowledge is ongoing.

Overall, the results encourage the development of the multimodal system approached. Further research has to investigate the reasons for the absence of an improvement of knowledge of the track structure. For example, for sighted travelers, information about potential landmarks might be more important. Therefore, it is possible that sighted participants focus their attention on verbal information regarding potential landmarks. However, other possible explanations cannot be ruled out with the current data.

Although further work is required to gain a more complete understanding of the effect of verbal assisting utterances on the knowledge acquisition process of (virtual) tactile maps, our findings show a clear improvement when receiving verbal assistance in the style of the extended-assistance condition on the knowledge of potential landmark and encourage the development of the multimodal human-computer-interaction system.

Acknowledgements. The research reported in this paper has been partially supported by DFG (German Science Foundation) in IRTG 1247 ‘Cross-modal Interaction in Natural and Artificial Cognitive Systems’ (CINACS). We thank Stephanie Badde, Carola Eschenbach, Christian Floß, Benjamin Haarhaus, Matthias Kerzel, and Mark May for discussion. We thank the anonymous reviewers for their highly useful commentaries.

References

1. Bérla, E.P.: Tactile Scanning and Memory for a Spatial Display By Blind Students. The Journal of Special Education 15(3), 341–350 (1981)
2. Blades, M.: The Reliability of Data Collected from Sketch Maps. Journal of Environmental Psychology 10(4), 327–339 (1990)

3. Blades, M., Ungar, S., Spencer, C.: Map Use by Adults with Visual Impairments. The Professional Geographer 51(4), 539–553 (1999)
4. Brunyé, T.T., Taylor, H.A.: Extended Experience Benefits Spatial Mental Model Development with Route but Not Survey Descriptions. Acta Psychologica 127, 340–354 (2008)
5. Brunyé, T.T., Taylor, H.A.: Working Memory in Developing and Applying Mental Models from Spatial Descriptions. Journal of Memory and Language 58(3), 701–729 (2008)
6. Buzzi, M.C., Buzzi, M., Leporini, B., Martusciello, L.: Making Visual Maps Accessible to the Blind. In: Stephanidis, C. (ed.) HCII 2011 and UAHCI 2011, Part II. LNCS, vol. 6766, pp. 271–280. Springer, Heidelberg (2011)
7. Cohen, J.: A Power Primer. Psychological Bulletin 112(1), 155–159 (1992)
8. Dahlback, N., Jonsson, A., Ahrenberg, L.: Wizard of Oz Studies—Why and How. Knowledge-based Systems 6(4), 258–266 (1993)
9. Espinosa, M.A., Ungar, S., Ochaíta, E., Blades, M., Spencer, C.: Comparing Methods for Introducing Blind and Visually Impaired People to Unfamiliar Urban Environments. Journal of Environmental Psychology 18, 277–287 (1998)
10. Field, A.P.: Discovering Statistics Using SPSS. Sage, London (2005)
11. Giudice, N.A., Bakdash, J.Z., Legge, G.E.: Wayfinding with Words: Spatial Learning and Navigation Using Dynamically Updated Verbal Descriptions. Psychological Research 71(3), 347–358 (2007)
12. Giudice, N.A., Betty, M.R., Loomis, J.M.: Functional Equivalence of Spatial Images From Touch and Vision: Evidence From Spatial Updating in Blind and Sighted Individuals. Learning, Memory 37(3), 621–634 (2011)
13. Habel, C., Kerzel, M., Lohmann, K.: Verbal Assistance in Tactile-Map Explorations: A Case for Visual Representations and Reasoning. In: Proceedings of AAAI Workshop on Visual Representations and Reasoning 2010, Menlo Park, CA, USA (2010)
14. Jacobson, R.D.: Cognitive Mapping Without Sight: Four Preliminary Studies of Spatial Learning. Journal of Environmental Psychology 18, 289–306 (1998)
15. Jacobson, R.D.: Talking Tactile Maps and Environmental Audio Beacons: An Orientation and Mobility Development Tool for Visually Impaired People. In: Proceedings of Maps and Diagrams for Blind and Visually Impaired People: Needs, Solutions, and Developments, Ljubjiana, Slovenia (1996)
16. Kitchin, R.M., Blades, M.: The Cognition of Geographic Space, I. B. Tauris, London, New York (2002)
17. Lohmann, K.: The Use of Sketch Maps as Measures for Spatial Knowledge. In: Wang, J., Brölemann, K., Chipofya, M., Schwering, A., Wallgrün, J.O. (eds.) Understanding and Processing Sketch Maps – Proceedings of the COSIT 2011 Workshop, pp. 45–54. IOS Press, Belfast (2011)
18. Lohmann, K., Eichhorn, O., Baumann, T.: Generating Situated Assisting Utterances to Facilitate Tactile-Map Exploration: A Prototype System. In: Proceedings of the NAACL Workshop on Speech and Language Processing for Assistive Technologies 2012, Montreal, QC, Canada (2012)
19. Lohmann, K., Eschenbach, C., Habel, C.: Linking Spatial Haptic Perception to Linguistic Representations: Assisting Utterances for Tactile-Map Explorations. In: Egenhofer, M., Giudice, N., Moratz, R., Worboys, M. (eds.) COSIT 2011. LNCS, vol. 6899, pp. 328–349. Springer, Heidelberg (2011)

20. Lohmann, K., Kerzel, M., Habel, C.: Generating Verbal Assistance for Tactile-Map Explorations. In: van der Sluis, I., Bergmann, K., van Hooijdonk, C., Theune, M. (eds.) Proceedings of the 3rd Workshop on Multimodal Output Generation 2010, Dublin (2010)
21. Lynch, K.: The Image of the City. MIT Press, Cambridge (1960)
22. Millar, S.: Understanding and Representing Spatial Information. British Journal of Visual Impairment 13(1), 8–11 (1995)
23. Montello, D.R.: A New Framework for Understanding the Acquisition of Spatial Knowledge in Large-scale Environments. In: Kahana, M.J., Golledge, R.G. (eds.) Spatial and Temporal Reasoning in Geographic Information Systems, Oxford, New York, pp. 143–154 (1998)
24. Newcombe, N.: Methods for the Study of Spatial Cognition. In: Cohen, R. (ed.) The Development of Spatial Cognition, pp. 277–300. Lawrence Erlbaum, Hillsdale (1985)
25. Shrout, P.E., Fleiss, J.L.: Intraclass Correlations: Uses in Assessing Rater Reliability. Psycholocial Bulletin 86(2), 420–428 (1979)
26. Siegel, A.W., White, S.H.: The Development of Spatial Representations of Large-scale Environments. Advances in Child Development and Behavior 10, 9–55 (1975)
27. Taylor, H.A., Tversky, B.: Spatial Mental Models Derived from Survey and Route Descriptions. Journal of Memory and Language 31(2), 261–292 (1992)
28. Tversky, B.: Cognitive Maps, Cognitive Collages, and Spatial Mental Models. In: Frank, A.U., Campari, I. (eds.) Spatial Information Theory: A Theoretical Basis for GIS, pp. 14–24. Springer, Berlin (1993)
29. Ungar, S.: Cognitive Mapping Without Visual Experience. In: Kitchin, R., Freundschuh, S. (eds.) Cognitive Mapping: Past, Present and Future, pp. 221–248. Routledge, London (2000)
30. Wang, Z., Li, B., Hedgpeth, T., Haven, T.: Instant Tactile-audio Map: Enabling Access to Digital Maps for People with Visual Impairment. In: Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 43–50. ACM, Pittsburg (2009)
31. Wellek, S.: Testing Statistical Hypotheses of Equivalence. CRC Press, Boca Raton (2003)
32. Wen, W., Ishikawa, T., Sato, T.: Working Memory in Spatial Knowledge Acquisition: Differences in Encoding Processes and Sense of Direction. Applied Cognitive Psychology 25, 654–662 (2011)
33. Zeng, L., Weber, G.: Audio-Haptic Browser for a Geographical Information System. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) ICCHP 2010, Part II. LNCS, vol. 6180, pp. 466–473. Springer, Heidelberg (2010)

A Linguistic Ontology of Mode: The Use of Locations in Spatial Language*

Sander Lestrade[1,2]

[1] Research group Sprachkontakt und Sprachvergleich
[2] SFB/TR 8 Spatial Cognition, Universät Bremen

Abstract. This paper discusses the linguistic ontology of mode (also known as *directionality*). It is argued that mode is best analyzed in terms of temporally linked locations and that our understanding of a corresponding path of motion probably follows from the conceptual enrichment of the semantics of a motion expression.

1 Introduction

Basic spatial expressions locate things at locations. These locations are named regions identified by means of the *configuration* function, which can be understood as defining a region with respect to a reference object or *ground* (latter term by Talmy [1]; for present purposes, it does not matter whether this region is geometrically or functionally defined and whether it should be described in points or vectors, but see for example [2, 3, 4] for discussion). In the example in (1), the configuration term *under* identifies the region under the ground *the table* that contrasts with, for example, the locations *on* and *next to it.*

Jackendoff [5, p. 169]
(1) *The mouse ran from under the table.*

As things may change location over time, configurations are often linked to a restricted interval only. This temporal range is specified by the *mode* function, as it will be called here (following Kracht [6]; mode is better known as *directionality* which is an unfortunate choice as will be explained below). According to the currently prevalent understanding of mode, the mode expression *from* in (1) describes the set of paths that have the location specified by the configuration *under* as their starting points [5, 7, 8, 9]. Alternatively, however, mode distinctions can be characterized in terms of locations that are mapped to an ordered domain such as (event) time: the mouse being under the table only at the beginning of the event [10, 11, 6, 12], or, as a third option, in terms of an abstract phase transition that only by the incidental application to the spatial domain expresses a transition of location [13]: the mouse in the first phase 'being under the table' and in the second phase '*not* being under the table'.

* Many thanks to John Bateman, Lotte Hogeweg, Susanne Hackmack, and three anonymous reviewers for valuable comments that helped to improve this paper.

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 319–337, 2012.

This paper will argue that the correct ontological characterization of mode is in terms of temporally linked locations, as the second type of analysis has it. In Section 2 it will first be explained what is meant by "ontological characterization", as this crucially distinguishes between different research programs with completely different research goals, which of course should be evaluated according to their own merits. Section 3 briefly discusses the three different types of analyses mentioned above. Section 4 shows why an analysis in terms of change of locations is preferable, at least for our present ontological purposes.

2 A Linguistic Ontology (for Spatial Language)

Our perception of the world is largely determined by our phylogenetic makeup and therefore probably universal (to give only one example, we can all roughly perceive about the same part of the electromagnetic spectrum). As is well-known, however, languages differ in the way they subsequently package these analog perceptual differences into discrete lexical items [cf. 14]. And as generally recognized too, languages differ in their "choice" of the subset of semantic distinctions that become *grammatically* relevant, that is, relevant to the system instead of the lexicon only.

Along these lines, the study of meaning can be subdivided into three levels: At the (universal) perceptual extreme, a *conceptualization* is the rich collection of sensory inputs that may or may not be expressible in language. Next, the *semantics* is the collection of meaning aspects that can be expressed in language and by means of which characterizations or definitions of lexical items can be given. At the grammatical extreme is the *linguistic ontology*, which describes those aspects of the semantics that become grammatically relevant. Whether or not it makes sense to distinguish between conceptualization and semantics and what exactly may count as "grammatically relevant" is not always consensual however, and probably the distinction should be considered a continuum rather than clear-cut.

Differences in labeling aside, this three-way division of meaning is not new but recognized by other scholars too [cf. e.g. 15, 16, 17, and references cited therein]. For example, Jackendoff [15, p. 291] distinguishes between a *conceptual, lexical* and *linguistic semantics*,[1] the latter is the equivalent of what is here called an *ontology* and is defined as "those semantic features that are mentioned in phrasal interface rules – the part of conceptualization that is 'visible' to these rules [i.e. the rules of grammar; SL]". Similarly, Grimshaw [16, Chapter 3] splits "the analysis of a verb like *write* in the following way: that *write* means to do something and not to be something is linguistic; that it means what it means and not what *draw* means, is not." The aspect of lexical meaning that is grammatically relevant she calls *semantic structure* or *linguistic information*, the aspect that is not she calls *semantic content* or *cognitive information*. Obviously, our ontology is the same as Grimshaw's semantic structure.

[1] In fact, Jackendoff proposes an additional level of meaning between perception and conceptualization, viz. *Spatial structure* or *SpS*.

The goal of a linguistic ontology is to identify the minimal semantic commitments of a language system as evidenced in concrete instances of language use [17]. The goal of this paper, accordingly, is to identify the minimal semantic commitments of the spatial system of languages that differentiate between mode expressions. That is, the purpose of the present paper is not to study the conceptual representation of motion expressions as stored and dealt with in the human brain, but instead to determine the grammatically relevant distinctions only. (Note that for engineering purposes, it is only this information that we need. In order to construct a system capable of interacting with natural language about space, it is only necessary to distinguish the semantic commitments embodied in language, independently of their instantiation in cognitive systems.)

One of the main reasons for the diverging analyses of mode expressions is probably the mixing of the three levels of meaning mentioned above. The prevalent understanding of mode expressions in terms of paths seems to be mostly motivated by a conceptual semantic intuition. Indeed, if we think about motion events we cannot but think of paths of motion too, for how else could some object go from a to b than by going through all intermediate points? Crucially, this is not necessarily how language goes about this, as will be shown in Section 4: The ontology of mode seems to concern a discrete change of locations only. First, however, the different types of analysis of mode that were mentioned above will be introduced in more detail.

3 Different Understandings of Mode

In this section a number of accounts of mode will briefly be discussed, divided into three larger types, *viz.* path, location, and phase accounts. Not all approaches discussed below adopt the distinction between semantics and ontology mentioned in the previous section and therefore not all analyses are strictly concerned with the ontology proper of mode. However, as the goal of this section is merely to show the variation in analyses that exists at different levels, the discussion is not limited to ontological accounts only. In the next section, it will be shown that for ontological purposes the location type of account qualifies best.

According to the first and presently most dominant view, mode is concerned with (sets of) paths, to be understood as ordered set of points in space. In fact, this first major type consists of two subtypes, depending on the role that is given to time. A path (and therefore mode) can either be considered to be atemporally ordered or thought to have some sort of temporal trace that links each point in space to a corresponding point in time.

The proposed organizing principles in the atemporal subtype of the path analysis may differ between accounts. According to Jackendoff [5, p. 165], there are five categories of mode, divided into three broad types.[2] The first class of

[2] Surprisingly, Jackendoff never seems to properly define *paths* or *path functions*. His use of the notion, for example in discussing the different mode classes and the possible roles paths may play in an event or state, suggests that he takes paths to be ordered sets of points in space. He does explicitly note, however, that paths are spatial and atemporal notions [5, p. 170].

bounded paths includes source paths and goal paths, FROM and TO. In bounded paths, the reference object or region with respect to this object specified by the configuration function is the beginning or endpoint of the path. The paths in the second class of *directions* do not include this reference object or region, but would do so if the path were extended by some unspecified distance. The members of this class are AWAY FROM and TOWARD. The third and final class of *routes* consists of one category only, *viz.* VIA. The reference object or region is related to some point in the interior of the path and nothing is specified about the endpoint of the motion.

As another, formally more developed, example of this subtype, Zwarts [8, 9] defines a path as a continuous function *p* from the real interval [0,1] to a domain *S* of places. A path has a starting point p(0) and an endpoint p(1) and for every $i \in$ [0,1], p(i) is an intermediary point of the path. [9] distinguishes four fundamental mode types: *Transitions, Cycles, Progressions,* and *Continuations.* Transitions involve paths that go from one spatial domain to a different, complementary domain. The difference between Transitions (Jackendoff's bounded paths) and Progressions (Jackendoff's directions) is that only the latter are *adjacent* and *cumulative.* In Zwarts' analysis, two paths are adjacent ("connect") if one starts where the other ends, i.e. p(1)=q(0). A set of paths X is connected iff there are p ∈ X with a connecting q ∈ X. A set of paths X is non-connected iff there are no p ∈ X with a connecting q ∈ X. A connected set of paths X is cumulative iff for all p, q ∈ X, if p+q exists, then p+q ∈ X. Since Transitions are defined as having either a starting or end point in a different spatial region, they are necessarily non-connected. Progressions such as *toward*, on the contrary, are connected and cumulative. You can add another four steps toward the house to four steps toward the house and still go toward the house (cumulativity) and therefore divide eight steps toward the house in two times four steps toward the house (connectivity).

Cycles and Continuations are different from Transitions and Progressions in that they are *reversible.* The path operation of *reversal* is defined as follows: the reversal of p is the path which assigns to every $i \in$ [0,1] the position that p assigns to $1 - i$. A set of paths X is reversible if and only if for every p, if p ∈ X then reversal (p) ∈ X. For example, Jackendoff's VIA is reversible (both 'jumping over the fence from left to right' and 'from right to left' are VIA), but the reversal of TO is FROM. The difference between cycles and continuations is that only the latter have the property of cumulativity: The concatenation of two cycles, e.g. *around*, is said not to yield a new cycle, but two cycles in sequence; in contrast, the concatenation of two continuations, e.g. *through*, does yield a new continuation.

In the *temporal* variant of path analyses [18, 19, 20], paths are elaborated with a temporal trace function that maps each point of the path to a point in the running time of the event. For example, Miller and Johnson-Laird say that

> "[t]he conceptual core of the system for indicating movement is the path, which usually has a distinctive beginning and end. As an object traverses a path it passes each successive location at a later moment in time, so time indices can be associated with each location." [18, p. 406]

Thus, the sequence of temporally marked locations constitutes a path. For example TO path is defined as the location y where a locatum x is at the end point of an interval in time with the additional assertion that x was not at y at the beginning of this time interval [18, p. 406]. The origin (FROM) and terminus (TO) of the path are said to have a distinctive status without further motivation. Next to these two points, the moving referent can be said to be at intermediate points of the path (VIA).

In a temporal version of the path analysis, the continuous function p in the atemporal proposal of Zwarts could be said to map from a *time* interval (rather than from the real interval) to a domain of places. This changes Zwarts' typology of mode, as [20] argues. Because time is directed, the property of reversibility disappears. As a result, only three basic distinctions of mode remain: *Place*, in which a path remains in the same location, *Goal*, in which a path ends up in some location, and *Source*, in which a path starts in some location. All other distinctions are argued to be derived from this basic set in this proposal. For example, VIA is analyzed as the combination of Goal and Source and TOWARD is said to be an atelic variant of Goal.

In the second major view, mode is characterized in terms of locations that only hold for a part of event or a restricted time period within some relevant time span [10, 11, 6, 12, 21]. The number of distinctions that are made in the location type of analysis again may differ between authors. For example, Schank [10] argues for a two-way distinction. He proposes that, what are here called, Source and Goal of a physical motion are two arguments of the conceptually primitive act PTRANS that "involves a change of state of something that is a location" [10, p. 225]. Source is the location where the object is located at the beginning of a PTRANS act and Goal is another location where the object ends up.

In a similar vein, Wunderlich [11] distinguishes between locative and dynamic prepositional phrases (PPs). Both are seen as one-place predicates that express the property of being located in some region. Dynamic PPs inherently involve time. Within dynamic PPs, only those prepositional phrases that combine with verbs of placement are called *directional.* Non-directional dynamic PPs combine with motion verbs and can be described in terms of paths, extended regions that unify the position of an object at different times [11, p. 602]. Instead, directional PPs are variants of locative PPs that crucially involve the basic predicate CHANGE, which expresses a transition from one region into another. CHANGE is related to parameters that have to be instantiated by the context, for example to a motion verb. Just like in Schank's analysis, mode only has two distinctions, Goal and Source.

Also Kracht [6, 12] proposes an account of mode in terms of locations that are true for specific time intervals of the running time of the event. Rather than developing mode distinctions that follow logically from his theoretical assumptions, Kracht seems to try to account for those distinctions that have been identified as such in the literature already. As a result, he ends up with five mode distinctions (or *modes*, in his terminology), depending on the time point at which some location holds. In the *static* mode, the entire event is located in the given location; in the *coinitial* mode (cf. FROM) the event starts out at a given location; in the *cofinal* mode (TO) it ends there; the *transitory* (VIA) describes a location in between (but not at) the begin and end point, and finally, there is the *approximative* mode (TOWARDS).[3] This mode is more complex as it involves the addition of a distance function to evaluate it. The location at the end point of the event must be closer with respect to the location for which the approximative mode holds than the location at the beginning of the event.

As a final example of the location analysis, and the one that is implicitly assumed here, Lestrade [21] argues that mode links configuration expressions to an extended event structure of the verb such as proposed by [22] ([cf. also 23]). Pustejovsky shows that Davidsonian event arguments may have internal structure. For our present purposes, only the structure in which there is a strict partial order between the two subevents is relevant: An event e_3 is a complex event structure that consists of two subevents, e_1 and e_2, where e_1 and e_2 are temporally ordered such that each is a logical part of e_3, the first subevent precedes the second, and there is no other event that is part of e_3 [22, p. 69]. Formally:

(2) a. $[_{e_3}\ e_1 <_\alpha e_2\ ] \ =_{def} \ <_\alpha (\{e_1, e_2\}, e_3)$

b. $\forall e_1, e_2, e_3[<_\alpha (\{e_1, e_2\}, e_3) \leftrightarrow e_1 \preceq e_3 \wedge e_2 \preceq e_3 \wedge e_1 < e_2 \wedge$
$\forall e[e \preceq e_3 \rightarrow e = e_1 \vee e = e_2\ ]]$

For a non-spatial example, the verb *build*, describing e_3, can be analyzed into a development process e_1 and a resulting state e_2.

Pustejovsky [22, p. 74] explicitly allows for adverbial phrases to take scope over both the entire event and over individual subevents. [21] argues that there are only three logical possibilities for spatial modification of motion verbs then: the spatial modification of the matrix event is called *Place* directionality;[4] the modification of the first subevent is called *Source*, and the modification of the second subevent is called *Goal*. For example, depending on the type of directionality that is imposed by the spatial modifier and assuming the structure in (2), a walking event e_3 of subject x modified by location y can be decomposed as follows: $[walk(e_3, x) \wedge locate(e_3, x, y)]$ for Place, $[walk(e_1, x) \wedge locate(e_2, x, y)]$ for Goal, and $[locate(e_1, x, y) \wedge walk(e_2, x)]$ for Source.

According to the third major type of analysis, finally, mode has to be understood as a more abstract category that has different, domain specific instantiations. Fong [13] analyzes mode expressions as ordered structures that are interpretable in any domain that is *diphasic*, that is, involving two abstract, complementary states. Only in spatial domains, mode expressions denote a change in

[3] The notions *approximative* and *TOWARDS* are in fact not comparable, as the latter, at least in the usage of Jackendoff, combines different ontological domains, only one of which is equivalent to Kracht's approximative, the other one in fact concerning orientation (cf. Section 4.1).

[4] For readers familiar with the framework of Jackendoff: Note the different use of the term *Place* here.

place. The spatial interpretation follows from the process of *co-compositionality* [22], in which an underspecified semantic form becomes contextually enriched by its composition. Following Löbner [1989, cited in 13, p. 29], the notion of *admissible phase-interval* is formulated as an interval that starts with a phase that is not-p and is followed by a phase that is (and stays) p. That is, it starts with times t for which $p(t) = 0$, it extends to later times t' with $p(t') = 1$, and there is no later time t'' with $p(t'') = 0$ again. The strict development from not-p to p is given up in [13], also allowing for changes in the opposite direction, from p to not-p. The crucial point remains the monotonicity of a change.

Unfortunately, the choice between an abstract phase or location type of analysis cannot be made on the basis of a study of spatial language, as the former analysis, once applied to the spatial domain, makes the same predictions as some of the analyses of the location type. This choice calls for a more philosophical discussion going beyond the scope of this paper, and therefore the third type of analysis will not be further considered here.

The two remaining options may be graphically represented as in Figure 1, illustrated for Goal mode only for reasons of space. In the path analysis, illustrated at the left, Goal mode is a continuous and ordered set of points in space that ends in a named region, represented by the circle. In the location analysis on the right, mode is a discrete notion expressing the link, represented by the arrow, between a location and a time interval: For a given interval, some locatum is said either to be or not to be located in some location (hence its discreteness).

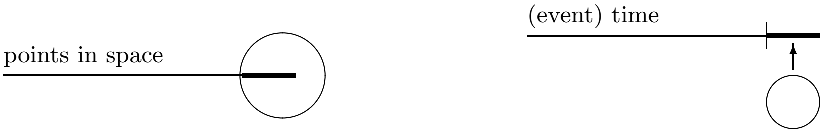


Fig. 1. Graphical representation of different analyses of Goal mode. Path analysis (*left*): The (heavy) end part of the ordered set of points in space is the subset that is described by some configuration expression, represented by the circle. Location analysis (*right*): The heavy part of the time line represents the final interval for which some location, again represented by a circle, is said to hold for a locatum. Goal mode, represented by the arrow, establishes this link.

In the next section, it will be argued that, of these remaining two options, the analysis in terms of temporally linked locations is to be preferred from the point of view of a linguistic ontology, as it best describes the basic set of distinctions that are made in a cross-linguistic sample of mode systems (Section 4.2) and some uses of mode expressions (Section 4.3). Also, it will be explained why the prevalent path analysis of mode need not be correct (Section 4.1).

4 Mode in Terms of Temporally Specified Locations

Defining mode, or any other semantic category for that matter, often is a circular procedure: To collect the instances to characterize one first needs an idea of what it is that is to be collected. In this section, we will first consider an example of how this arguably has led to a wrong sample of alleged mode expressions. Since the path analysis is almost exclusively motivated on the basis of this data set, it follows that the jury is still out and that the location analysis still stands a chance. Next, in Section 4.2, a more sound procedure of data collection is discussed. Following a procedure proposed by Corbett [24], first the kernel of mode distinctions is established on the basis of which a location type of analysis in fact seems to be much more plausible. Subsequently, in Section 4.3, a number of observations will be discussed that further strengthen the case for analyzing mode in terms of locations.

But before we begin, it should again be stressed that the purpose here is *not* to argue that motion paths cannot be referred to by language nor to argue that analyses of paths as such are wrong. The point to be made is that *mode* is indifferent toward paths.

4.1 Leaving the Path Analysis

The crucial examples in Jackendoff's choice for paths are given in (3). He argues that any alternative to his path analysis of mode cannot account for all nine possibilities given [5, p. 168–170], in which especially (3-b,c) are supposed to be problematic. Instead, a path, be it concrete or metaphorical, can be identified in all of them and it is this path, it is claimed, that is referred to by the prepositional phrase.

(3) a. ([THING] traverses [PATH])
1 – bounded path: John ran into the house.
2 – direction: The mouse skittered toward the clock.
3 – route: The train rambled along the river.
b. ([THING] extends over [PATH])
1 – bounded path: The highway extends from Denver to Indianapolis.
2 – direction: The flagpole reaches (up) toward the sky.
3 – route: The sidewalk goes around the tree.
c. ([THING] is oriented along [PATH])
1 – bounded path: The sign points to Philadelphia.
2 – direction: The house faces away from the mountains.
3 – route: The canons aim through the tunnel.

Note however that this argument only goes through if we first assume that paths are the primitives of mode (and if we accept the identification of a path in the sentences in (3-c)): It is assumed that mode is about paths; next, examples of situations in which paths can be identified are given, and then it is said that these examples, and therefore mode, can only be uniformly described in terms of paths. Crucially, however, there is no independent evidence that these sentences actually are examples of *mode*.

The focus on (English) spatial adpositions, of which we have just seen only one example but which is characteristic for the vast majority of undertakings in spatial semantics, is probably the main reason that a path analysis has not previously been questioned [but cf. 13, 6, 25, 21]. It suffers from two problems, however. First, by only looking at English we miss potentially relevant information from other languages. It may be that a cross-linguistic data sample suggest a different characterization. For example in the configuration domain, the comparison of two rather closely related languages like English and Dutch already calls for a quite different ontology than one may have come up with for English alone [cf. 26]. Secondly, Bateman et al. [17, p. 1035] argue that lexical items tend to be too idiosyncratic in their bundling of semantic properties to reveal generic semantic ordering principles. Languages can be expected to vary at more lexical levels of spatial organization, combining semantic features from domains that are not necessarily basic to the actual domain of interest. Although, in comparison with nouns, prepositions are rather grammatical, they probably still are too lexical a category to give a clean picture of the mode domain. This can easily be illustrated: Since languages generally have dozens of non-synonymous prepositions, they necessarily make more than just the five-way mode distinction that theories of mode allow at the maximum (cf. Section 3). As a result, a semantic characterization of mode on the basis of prepositions may include meaning contrasts that do not really belong to the domain proper. In the remainder of this section, it is shown how the class of prepositions indeed mixes ontological categories. In Section 4.2, an alternative, more principled method of data collection will be shown to yield a different and smaller domain of analysis.

In the proposal of Jackendoff, TOWARD belongs to the type of paths called *directions*, which, unlike a *bounded* path such as TO, do not include (the region with respect to) the reference object but would do so if the path were extended by some unspecified distance. In a non-trivial sense, we probably only want to allow for extensions in approximately the same direction (otherwise, any direction could be turned into a TO path).[5] Now, consider an enclosure around point A with an opening at its south side and point B to its north, as illustrated in Figure 2. Because of the enclosure, one can only go from *A to B* going southwards, through the opening. To go from *A toward B*, however, one should go north. Crucially, the TOWARD path in this situation cannot be extended in the same direction to become a bounded *to B* path.

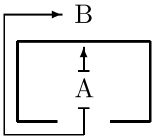


Fig. 2. TO vs. TOWARD

It is proposed here that *toward* expresses something categorically different from *to*. Instead of mode, *toward* expresses orientation. When modifying a motion event with an orientation, the moving object generally ends up closer to the ground in which direction it is moving. And by continuing along this direction,

[5] Thanks to Emar Maier (p.c.) for this observation.

the moving object will probably end up at this ground too. But this need not be, as this example shows. (Whether “towardlike” expressions in other languages need to be analyzed as orientation or in terms of the combination of Goal mode with ‘near’ configuration must be determined on a case-for-case basis.) To avoid any further confusion between an orientation and location interpretation of *directionality*, it is proposed to use this notion for orientation only, using *mode* for the temporal linking of locations instead [following 6].

4.2 The Kernel of Mode

As said above, an important problem for the characterization of mode is that we do not really know what the data to account for are. Some prepositions in English may express it, but apparently not all do, as just shown for *toward*. So how do we know which items to include in our analysis? A possible solution proposed by Corbett [24] is to first establish and describe a kernel of distinctions that we surely want our ontology to account for. Other distinctions then either are less basic to our domain of interest or simply alien to it.

We can establish this kernel for mode by studying spatial-case inventories from a cross-linguistic perspective. It is well-known that more grammatical means of expression tend to make less idiosyncratic or language-particular meaning distinctions [1, 27, 17], which, for ontological purposes, is the reason we are interested in such grammatical distinctions in the first place. Morphological case should be especially interesting in this respect, as it is one of the most grammatical means available in languages. And if we find the same type of distinctions in language after language, we can be more certain that this set is basic indeed (cf. [28]). Our kernel of mode, then, will consist of those distinctions that systematically show up in the spatial-case systems of different languages.

Lestrade [30] describes a cross-linguistic study of spatial-case inventories that was thus motivated [cf. 31, 33, 32, for similar enterprises, with consistent results]. A summary of the results is given in Table 1. A few explanatory notes are necessary for correct interpretation. If a spatial case is used as a general marker of a location function that does not distinguish between modes, the meaning column remains empty. If a basic level of mode is missing from the mode column, this generally means that it is expressed by different means than case markers. For example, in Malayalam, a postposition *ninnə* that combines with a locative case marked noun is used to mark Source. Syncretism between, or underspecification of, certain meanings is marked with a slash (“/”). The difference between the absence of a certain mode distinction from this table and an underspecification/syncretism analysis is that in the former case a different means of expression is (systematically) used, whereas in the latter case one spatial case marker really seems to express two mode distinctions.

As can be seen, Place, Source, and Goal distinctions are always in place before other mode-like distinctions such as ‘via’ or ‘toward’ enter the stage [for a more elaborate discussion of these data, cf. 30]. Thus, the kernel of mode as established via a cross-linguistic study of spatial-case inventories consists of Place, Goal, and Source. For concreteness, an example of this distinction is given for

Table 1. Mode distinctions made by spatial case inventories. Abbreviations: P Place, S Source, G Goal.

Language	Mode(-like)	Language	Mode(-like)
Ainu	P, G, S, ‘via’	Ket	P1, P2, S, G, ‘via’
Alamblak	P, P/G, G, ‘via’	Koasati	P(/G), G
Aymara	P, G, S	Lithuanian	P
Basque	P, G, S, ‘up to’	Malayalam	P/G, G
Cahuilla	P, G, S	Mangarrayi	P(/G), G
Dyirbal	P, G, S	Maricopa	P/G, ‘via’
ES Nivkh	P/S, G, ‘up to’	Meithei	P/G, S
Evenki	P, G, S, ‘toward’, ‘via’, ‘along’, ‘from the direction of’	Mundari	P/G, S, ‘up to’
Finnish	S, P, G	Nez Perce	P, G, S
Harar Oromo	P/G, S	P-Maliseet	
Hua	P/G, S	Tarma Quechua	P, G, S, ‘via’, ‘up to’
Hungarian	S, P, G, ‘toward’	Tswana	
Hunzib	P, G, S, ‘via’	Tundra Yukaghir	P/G, P, G, S, ‘via’
Ika		Warao	P, G/‘via’, S
Imonda	P, G, S	West Greenlandic	P, G, S, P/‘via’
Kanuri	P/S/‘via’, G	Yasin-Burushaski	P, G, S

Dyirbal, in which allative case marks Goal (4-a), the ablative marks Source (4-b), and locative case marks Place (4-c).

Dyirbal [34, p. 57]

	a. *miɖa-gu*	b. *miɖa-ŋunu*	c. *miɖa-ŋga*
(4)	camp-ALLATIVE	camp-ABLATIVE	camp-LOCATIVE
	‘to the camp’	‘from the camp’	‘at the camp’

Differently from Goal and Source, Place involves the absence of a change of location. As a result, Place is not always accepted as a distinction of mode. Due to space limitations, a discussion of its mode status has to be left for another occasion and only Goal and Source distinctions will be considered as mode distinctions in the remainer of this paper. (Note however that including it would not undermine the present argument but in fact would only strengthen the case for a location analysis, especially in the form of [21] in which Place simply links the location to the matrix event.)

The question of our present concern is whether the (remaining) basic distinctions Source and Goal follow from the proposals. That is, are these mode distinctions in any way privileged over other less-basic distinctions by the different types of analysis, when understood as ontologic proposals?

As discussed in Section 3, different accounts of the path type of analysis come up with different classifications of mode. Although a *temporal* version of the path analysis could predict the kernel of mode [20], there is no inherent ordering in the typology of mode distinctions proposed by the *atemporal* path accounts. From this perspective, the accounts of Jackendoff [5, 7] and Zwarts

[8, 9] overgenerate: They do not predict the kernel of mode. Instead, infrequent Path types such as Jackendoff's VIA, AWAY FROM or Zwarts's continuation and cycles are predicted to be just as basic as FROM and TO paths.

Although the absence of a hierarchy of mode distinctions could of course be due to the specific formulations of Jackendoff and Zwarts only, it seems to be inherent instead: Although path analyses can be made to account for the kernel when understood as a temporal notion, this in fact implies implementing a location analysis. For this, we have to stipulate a defining importance of the location of the starting and end point of the path. That is, we have to ignore all intermediate points of the path and reduce its notion to something that basically says: first the locatum was not at this location, in the end it was (for Goal, and the other way around for Source). In other words, we have to reduce the set of points described by the path to a set of two, one that is and one that is not at some location. In again other words, we have to assume a change of location analysis.

Interestingly, Zwarts [9] eventually does end up with a distinction between more and less "directional" paths in which Source and Goal are the "most directional" expressions. These basic mode options are non-cumulative and reversible prepositions (cf. Section 3). Together, these properties define his class of Transitions, which are changes of location. Thus, indeed, prototypical mode needs to be described in terms of a change of location in a path analysis too. In fact, Zwarts [9, p. 95] even acknowledges the option of path prepositions lacking mode, thereby showing that paths and mode are different things indeed. Only by arguing that locations at the begin and end point of a path have special relevance, a path analysis can predict the kernel. But probably, such argumentation always involves an argument that really favors the location analysis.

In contrast with path analyses, analyses in terms of locations can predict the kernel of distinctions. In most general terms, according to this type of analysis, mode is about a configuration that is linked to some ordered dimension (mostly time; the ordering is necessary to distinguish Source from Goal). Depending on the ordering dimension chosen and on the way in which this link is established, location analyses correctly predict two (*viz.* Goal and Source; [10, 11]) or three ([21]; additionally including Place) basic types. This type of analysis may overgenerate too, however, as evidenced by [6, 12], in which four or five distinctions are predicted, depending on how we appreciate the approximative mode.

In sum, the most succinct ontological analysis of the two (three, actually) basic distinctions of mode as identified in a cross-linguistic study of spatial-case inventories seems to be in terms of locations. In the next section, additional evidence for this analysis will be discussed.

4.3 Additional Support for a Location Analysis

Above, it was first shown that we do not necessarily need to adopt a path analysis as the argumentation for it is based on circular reasoning (Section 4.1) and, next, that when considering the kernel of mode distinctions, an analysis in terms of locations in fact seems preferable (Section 4.2). In the remainder of this section, a number of observations will be discussed that corroborate this proposal or at least make more sense when mode is thought of as the temporal linking of locations.

First, and most importantly, even if we wanted to say that motion expressions ontologically involve paths, the question is whether the encoding of this path is to be found *at the ground*, i.e., as a function of the reference object. Traditionally, the focus of theorizing about spatial meaning is on its local encoding, offloading a large part of the motion semantics to prepositions. But spatial meaning rather seems distributed over the whole sentence [cf. [35], [36], [37]]. As a result, the ontology of an adposition may be much "lighter" than usually argued for. Indeed, according to the present proposal, this path, if present, would instead be encoded by the verb, the mode expression only modifying a specified part of this path. Consider the following example to make things more concrete:

(5) *During his holidays, John walked every day.*

Assuming that John did not go on vacation to walk on a treadmill, one could argue that a path is already present in (5), necessitated by a sensible interpretation of the verb *to walk* in this context. What a mode expression adds to such an utterance, then, is the linking of a location to the begin or end point of this path (possibly established via the alignment between path and event time), in case of Source or Goal mode, or the location of the path as a whole, in case of Place.

As another argument in favor of a lighter semantics of mode expressions consider the following example from Hungarian:

Hungarian (T. Bos, p.c.)
(6) *János *(ki)-megy a haz-ból.*
John out-goes the house-ELATIVE
'John is leaving the house.'

Hungarian Source mode distinctions (and, to a lesser extent, Goal; cf. [[38], p. 229]) need to be "licensed" by verbal prefixes. That is, without *ki-* prefixed to the verb the use of elative case, which expresses Source, is ungrammatical. If elative case expressed a Source path, one would think it should have been possible to combine it with a simple verb like *menni* 'to go/walk', like it is possible to insert *to the shop* in (5). Instead, if elative case links its location to a subinterval of the event only, it needs a biphasic event or translational motion to provide the subeventual structure to link to. Apparently, in Hungarian, a simple motion verb may not sufficiently provide this structure or motion component, neither does the mode expression itself. Instead, an additional prefix is necessary for this.

It may seem hard to believe that the motion verb 'walk' does not sufficiently express a change of place from an English perspective, but we find a very similar situation in Yukatek Maya. The combination of manner of motion verbs with spatial adjuncts leads to a Place interpretation, as illustrated in (7-a). Only with a specific class of “inactive” verbs, *hem* ‘descend’ in (7-b) and *hna’k* ‘ascend’ in (7-c), a change of location can be expressed:

Yukatek Maya [25, ex. 1]

(7) a. *Le=ch’íich’-o’ túun xíiknal y-óok’ol le=che’-o’.*
DEF=bird-D2 PROG:A3 fly A3-top DEF=tree-D2
‘The bird is flying [i.e., circling] above the tree.’

b. *Le=ch’íich’-o’ h-em u=xíiknal te=che’-o’.*
DEF=bird-D2 PRV-descend(.B3.SG) A3=fly LOC:DEF=tree-D2
‘The bird flew down from the tree [lit. it descended from the tree flying].’

c. *Le=ch’íich’-o’ h-na’k u=xíiknal te=che’-o’.*
DEF=bird-D2 PRV-ascend(.B3.SG) A3=fly LOC:DEF=tree-D2
‘The bird flew up to the tree [lit. it ascended the tree flying].’

In Yukatek Maya adjuncts themselves apparently do not make a mode distinction. Their interpretation as a Place, Goal, or Source is dependent on the verb construction. In a clause with a manner of motion verb only, i.e. without the additional use of an inactive verb, a spatial adjunct is interpreted as a Place, no matter how plausible a Goal or Source interpretation would be. Only in combination with a verb from the inactive class, a change of location can be expressed, as illustrated in (7-b,c).

These data are meant to show that spatial adverbial expressions may only provide a location. In the case of Yukatek Maya, that is, without the help of further pointers, only if the verb sufficiently provides a path or change of location, this location can be interpreted as a Goal or Source. Thus, if paths are to be part of our analysis, they may belong to the verb semantics rather than to that of the location expression.

Of course, in our eventual interpretation of the whole utterance the different parts have to be integrated. That is, the location that is specified by the configuration expression needs to be linked to the semantics of the motion verb in some way. Mode expression can be thought of as pointers that facilitate this integration. For example, Goal mode says that its configuration should be linked to the end state (in case of a change of state analysis of motion) or end point (in case of a path analysis) of some motion event. In both options, crucially, mode expressions themselves need not refer to paths.

Note that the above discussion does not mean that motion verbs necessarily involve paths. It could be that the notion of paths is altogether outsourced to the conceptualization part of meaning (cf. Section 2). Probably, this is a language-particular matter. For example, Bohnemeyer [25] argues on the basis of examples like the above that in Yukatek Maya, motion is not framed as a moving object traversing a path, but as a discrete location change with respect to single grounds instead. The traversal of a corresponding path is left to implicature. Also if the subject/relatum does not move but the ground is moved instead, change of location verbs, which in English would suggest motion of the subject/relatum, are applicable (cf. [39] for similar observations in Japanese). Similarly, if an object has "beamed" into a ground, one can say that it *entered* it in Yukatek Maya.

But note the felicitous use of *into* in combination with the verb *beam* in the previous sentence. Apparently, in English too we can use Goal (and Source, as will be shown now) expressions in combinations with verbs that express a discrete change of location. Another example, taken from the Wikipedia entry for *teleportation*, which is arguably a better example as *to beam* could be said to involve some sort of a path (John Bateman, p.c.), is given in (8):

(8) *Teleportation is a term that refers to a number of theories and notions concerning the transfer of matter* ***from*** *one point* ***to*** *another* ***without traversing the physical space between them****, similar to the concept apport, an earlier word used in the context of spiritualism*

(consulted on September 9, 2011; emphasis mine, SL)

Whereas a path analysis probably needs to analyze this example in terms of a metaphorical use in which the hypothesized essence of the mode expressions, *viz.* the path, is ignored, a change of location analysis can straightforwardly account for such uses. Obviously, given the infrequency of this phenomenon in our daily lives, there are not too many examples to illustrate this use and to tell apart metaphorical from standard use. The important point here is the naturalness with which such verbs, if they occur, combine with Source and Goal expressions, suggesting mode does not express paths but rather a discrete change of location.

When we actually do use a metaphor, we highlight a common structure in two different domains. We use something we know about a well-understood domain (the *vehicle*), to understand something else. Now, if the essence of mode was about paths, we should be able to identify a metaphorical path in non-spatial uses of mode expressions too – assuming that there is a non-arbitrary relation between spatial and non-spatial uses of an item. This is not what we find however, as shown in the following example:

(9) *The traffic light turned from red to green.*

Rather than following a continuous path, the transition between red and green in (9) seems to be very discrete. Obviously, if mode is understood as a change of location instead of paths, this non-spatial use is straightforwardly explained as a metaphorical change of location, i.e. from 'being in a red state' to 'being in a green state'. One may argue that (9) is not an example of the metaphorical use of mode expressions, reserving the notion for more creative uses only. The point then still stands, however: A unified semantics for these and spatial motion uses of *from* and *to* suggests a discrete rather than continuous ontology.

Finally, another line of reasoning that shows that paths may not be inherent to mode expressions goes as follows. If a path analysis of mode were correct, it should be possible to align an elongated or distributed figure with a mode expression by means of an existential verb. For a path, at least for an atemporal one, it should not matter whether something is covering it at a single moment in time or whether something is subsequently going through its parts: In both cases, the path is an ordered set of points in space. As examples (10-a,b) show, the use of an existential verb is indeed possible for prepositions that in fact refer to paths, but as evidenced by (10-c,d) this does not hold for the canonical mode distinctions Goal and Source (since intuitions are subtle here, I use examples from my native language Dutch):

(10) a. $^{\surd}$*De kabel ligt door de sloot.* ‘The cable is through the ditch.’
b. $^{\surd}$*Er ligt zand door de kamer.* ‘There is sand through the room’
c. **De kabel ligt de sloot in.* ‘The cable is into the ditch.’
d. **Er ligt zand de kamer uit.* ‘There is sand from the room.’

Under a location analysis, contrastively, mode expressions that link a location to a subinterval of the event only need a motion verb or at least a context on which a change of location reading can be imposed. This excludes their combination with a stative verb like *liggen* ‘to lie’ as indeed observed in (10-c,d).

4.4 Conclusion

In this section it was shown that locations rather than paths should be considered as the primitives of mode. Although paths are very likely to play a role in our eventual conceptualization of motion events and may be linguistically encoded by some motion verbs or prepositions indeed, they are probably alien to the ontology of mode. That is, a path interpretation of mode probably follows from the semantic integration of mode expressions in the larger linguistics environment or from the addition of world knowledge to the semantics proper.

5 Discussion

It was argued above that the ontology of mode should be formulated in terms of temporally linked locations, and that paths need not, and probably do not, characterize the mode domain, as the prevalent analysis of mode has it. Configurations, on the contrary, may very well involve paths. In fact, it is hard to think of a different way of analzying for example the semantics of prepositions such as *around* or *across*, which are, in the present proposal, considered to be configurations indeed.

The following examples may serve to illustrate how modes and (path) configurations can be analyzed (using the technicalities of [21] and assuming the strict partial order structure described in Section 3):

(11) a. *The boy ran* $\emptyset$ *around* *the house.*
locatum motion verb (mode:Place) configuration:around ground
$[run(e, boy) \wedge locate(e, boy, around\ the\ house)]$
b. *The cat came* *from* *under* *the table.*
locatum motion verb mode:Source configuration:under ground
$[locate(e_1, cat, under\ the\ table) \wedge come(e_2, cat)]$

c. *The bird flew* *into* *the house.*
locatum motion verb mode:Goal&configuration:in ground
$[fly(e_1, bird) \wedge locate(e_2, bird, in\ the\ house)]$

Place mode could either be said to be covertly expressed or the default interpretation in the absence of Source and Goal markers, hence the use of the parentheses in the glossing in (11-a). The formula shows how the location *around the house* in this example is understood as applying throughout the running event. This is contrasted with the Source use of *under* in (11-b). Here, the formula makes explicit the subeventual structure in which a location event precedes one of coming. In (11-c), Goal mode, which can be considered the opposite of Source, is illustrated with the 'in' configuration.

Finally, an anonymous reviewer suggests that the difference between the two views opposed in the previous section lies in the assumption on what a path is (which, as said above, is indeed not always made explicit). If a path is taken to be "a position relative to an ordered dimensions that allows the distinction of Source and Goal (and might consist of nothing but an ordered pair of points/positions)", this too, the reviewer suggests, would explain the above findings and provide a uniform model for all "directional" expressions, including notions such as *around.* Although I am not sure if this idea could indeed account for all observations made here (something that I will leave to the reviewer to demonstrate), I think we maybe should not even try to bring the two together and instead appreciate the possible merits of the view of mode as proposed here, such as its compatibility with both geometrical and functional characterizations of locations, the very strong predictions it makes about the kernel of mode (and hence the straightforward account of it), and the clear distinction between ontological categories (whereas paths sometimes seem to straddle the distinction between "directionality" and configuration, no such confusion is possible here: mode is the link between location and event time).

References

[1] Talmy, L.: Toward a Cognitive Semantics. MIT Press, Cambridge and London (2000)

[2] Zwarts, J.: Vectors as relative positions: A compositional semantics of modified PPs. Journal of Semantics 14, 57–86 (1997)

[3] Zwarts, J., Winter, Y.: Vector space semantics: A model theoretic analysis of locative prepositions. Journal of Logic, Language and Information 9, 169–211 (2000)

[4] Garrod, S.C., Sanford, A.J.: Discourse models as interfaces between language and the spatial world. Journal of Semantics 6, 147–160

[5] Jackendoff, R.: Semantics and Cognition, 8th edn. MIT Press, Cambridge (1983)

[6] Kracht, M.: On the semantics of locatives. Linguistics and Philosophy 25(2), 175–232 (2002)

[7] Jackendoff, R.: Semantic Structures. MIT Press, Cambridge (1990)

[8] Zwarts, J.: Prepositional Aspect and the Algebra of Paths. Linguistics and Philosophy 28, 739–779 (2005)

[9] Zwarts, J.: Aspects of a typology of direction. In: Rothstein, S. (ed.) Theoretical and Crosslinguistic approaches to the Semantics of Aspects, pp. 79–106. John Benjamins, Amsterdam (2008)

[10] Schank, R.: Identification of conceptualizations underlying natural language. In: Schank, R., Colby, K. (eds.) Computer Models of thought and Language, pp. 187–248. W.H. Freeman and Co., New York (1973)

[11] Wunderlich, D.: How do prepositional phrases fit into compositional syntax and semantics? Linguistics 29, 591–621 (1991)

[12] Kracht, M.: The Fine Structure of Spatial Expressions. In: Asbury, A., Dotlačil, J., Gehrke, B., Nouwen, R. (eds.) The Syntax and Semantics of Spatial, pp. 35–62. John Benjamins, Amsterdam (2008)

[13] Fong, V.: The order of things: What directional locatives denote. PhD Thesis, Stanford University (1997)

[14] Gärdenfors, P.: Conceptual spaces. The geometry of thought. MIT Press, Cambridge (2004)

[15] Jackendoff, R.: Foundations of language. Brain, meaning, grammar, evolution. Oxford University Press, Oxford (2002)

[16] Grimshaw, J.: Words and Structure. CSLI, Chicago (2005)

[17] Bateman, J.A., Hois, J., Ross, R., Tenbrink, T.: A linguistic ontology of space for natural language processing. Artificial Intelligence 174(14), 1027–1071 (2010)

[18] Miller, G., Johnson-Laird, P.: Language and perception. Harvard University Press, Cambridge (1976)

[19] Piñón, C.J.: Paths and their names. In: Beals, K., Cooke, G., Kathman, D., Kita, S., McCullough, K.-E., Testen, D. (eds.) Papers from the 29th Regional Meeting of the Chicago Linguistic Society, CLS, Chicago, vol. 2, pp. 287–303 (1993), http://pinon.sdf--eu.org/covers/ptn.html

[20] Lestrade, S.: The space of case. PhD Thesis, Radboud University Nijmegen (2010)

[21] Lestrade, S.: Analyzing directionality: From paths to locations. In: Hois, J., Ross, R.J., Kelleher, J., Bateman, J. (eds.) Proceedings of the 2nd Workshop on Computational Models of Spatial Language Interpretation and Generation (CoSLI–2), Boston, MA (2011)

[22] Pustejovsky, J.: The generative lexicon. The MIT Press, Cambridge (1995)

[23] Moens, M., Steedman, M.: Temporal ontology and temporal reference. Computational Linguistics 14(2), 15–28 (1995)

[24] Corbett, G.G.: The penumbra of morphosyntactic feature systems. Morphology 21(2), 445–580 (2010)

[25] Bohnemeyer, J.: The pitfalls of getting from here to there: Bootstrapping the syntax and semantics of motion event expressions in Yucatec Maya. In: Bowerman, M., Brown, P. (eds.) Cross–Linguistic Perspectives on Argument Structure: Implications for Learnability, pp. 49–68. Lawrence Erlbaum, NJ (2007)

[26] Bowerman, M.: Learning how to structure space for language: A cross–linguistic perspective. In: Bloom, P., Peterson, M.A., Nadel, L., Garrett, M.F. (eds.) Language and Space, pp. 385–436. MIT, Cambridge (1996)

[27] de Hoop, H., Zwarts, J.: Case in formal semantics. In: Malchukov, A., Spencer, A. (eds.) The Oxford Handbook of Case, pp. 170–184. Oxford University Press, Oxford (2009)

[28] Gentner, D., Bowerman, M.: Why some spatial semantic categories are harder to learn than others: The typological prevalence hypothesis. In: Guo, J., et al. (eds.) Crosslinguistic approaches to the Psychology of Language, pp. 465–480. Psychology Press, New York (2009)

[29] Wilson, M.: Six view of embodied cognition. Psychonomic Bulletin and Review 9(4), 625–636 (2002)

[30] Lestrade, S.: The kernel of mode: A cross-linguistic study of spatial-case inventories. University of Bremen (2011) (manuscript)

[31] Kilby, D.: Universal and particular properties of the Ewenki case system. Papers in Linguistics 16(3/4), 45–74 (1983)

[32] Creissels, D.: Spatial cases. In: Malchukov, A., Spencer, A. (eds.) The Oxford Handbook of Case, pp. 609–625. Oxford University Press, Oxford (2009)

[33] Stolz, T.: Lokalkasussysteme. Aspekte einer strukturellen Dynamik. Gottfried Egert Verlag, Wilhelmsfeld (1992)

[34] Dixon, R.M.W.: The Dyirbal language of North Queensland. Cambridge University Press, Cambridge (1972)

[35] Sinha, C., Kuteva, T.: Distributed spatial semantics. Nordic Journal of Linguistics 18, 167–199 (1995)

[36] Levinson, S.C.: Space in language and cognition: Explorations in cognitive diversity. Cambridge University Press, Cambridge (2003)

[37] Zlatev, J.: Holistic spatial semantics of Thai. In: Palmer, G., Casad, E. (eds.) Cognitive Linguistics and Non–Indo European Languages, Mouton de Gruyter, Berlin (2003)

[38] Hegedűs, V.: Hungarian spatial PPs: Nordlyd: Tromsø Working Papers in Linguistics 33(2), 220–233 (2008)

[39] Kita, S.: Japanese ENTER/EXIT verbs without motion semantics. Studies in language 23(2), 307–330 (1999)

Detecting Events in Video Data Using a Formal Ontology of Motion Verbs[*]

Tommaso D'Odorico and Brandon Bennett

School of Computing, University of Leeds, Leeds, LS2 9JT, United Kingdom

Abstract. In this paper we formalise an ontology for motion verbs based on classical logic, event calculus, supervaluation and standpoint semantics. We present a theoretical account and a logic-programming implementation of the ontology, which fits within a system aimed at detecting event occurrences in video scenes. Our purpose is to build a bridge between Computer Vision and Knowledge Representation and Reasoning, and to address the issue of concept vagueness in formal ontologies.

1 Introduction

This paper describes an *ontological approach* for the detection of event occurrences in video scenes. We introduce a formal semantics for vague terms and its implementation ProVision, a logic-programming event detection system. Motivation stems from our intent to tackle the Mind's Eye Challenge described in Sec. 2, a project involving the detection of 48 motion verbs over a large dataset of video scenes. We believe that an ontological approach based on formal definitions is able to semantically characterise motion verbs more comprehensively than machine learning based approaches.

Our research also concerns the general subject of *vagueness*, introduced in Sec. 3. Vague concepts are ubiquitous in natural language, but they prove to be problematic in formal ontologies as the boundary for their applicability is not easy to establish, hence hampering the provision of a clear, classical formal definition.

The specific domain of *motion verbs* on which we focus our attention is outlined in Sec. 2, together with the data available for our experimental section. Our long-term goal is to formalise most of these concepts within an ontology for vague motion verbs, ready for implementation in our event detection system. A preliminary version of this ontology is laid out in Sec. 4, with a specific focus on the verbs 'approach' and 'hold'.

In Sec. 5 we report some experimental results given by the inferred event occurrences returned by ProVision, which contains an implementation of the ontology. Detection statistics for the verbs 'approach' and 'hold' are discussed in Sec.6.

[*] This work has been partially funded by the DARPA Mind's Eye Program (project VIGIL, W911NF-10-C-0083), whose support is gratefully acknowledged.

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 338–357, 2012.

2 Application Domain

The interest in the development of a logical formalism for reasoning about vague motion verbs has begun with our involvement in the DARPA Mind's Eye challenge [8,9], a long-term project aimed at automatically recognising 48 types of actions from video sequences, listed in Table 1.

Table 1. Motion verbs list

Approach	Arrive	Attach	Bounce	Bury	Carry
Catch	Chase	Close	Collide	Dig	Drop
Enter	Exchange	Exit	Fall	Flee	Fly
Follow	Get	Give	Go	Hand	Haul
Have	Hit	Hold	Jump	Kick	Leave
Lift	Move	Open	Pass	Pick Up	Push
Put Down	Raise	Receive	Replace	Run	Snatch
Stop	Take	Throw	Touch	Turn	Walk

DARPA also provided an extensive collection of videos, each of which contain examples of one or more occurrences of actions that can be described by the verbs in Table. 1.

The Mind's Eye challenge is composed of several tasks, mainly *recognition*, *gap filling* and *anomaly detection*. The recognition task, aims at recognising any occurrence of an event within a particular video sequence among the ones listed in Table 1. The gap filling and anomaly detection tasks, outside the scope of the work presented here, respectively aim at inferring event occurrences over the gaps of an incomplete video sequence, and at detecting unconventional or singular occurrences of events.

A number of approaches to the recognition of events from video sequences are based on a combination of Machine Learning and Inductive Logic Programming [12,11,23], where event models are learnt through the analysis of qualitative spatio-temporal relations between objects in the videos.

The approach to the recognition task we are presenting in this paper is based on a formal ontology of motion verbs. The verbs describing the events to be recognised are very particular, and we believe that this fundamentally different approach allows for a greater specification and understanding of each verb's semantic characteristics, which may not be completely grasped by learning techniques.

Our event detection system ProVision is an implementation of this formal ontology. It is designed as a module of a wider framework for event analysis and detection, whose input is a video and whose output is a high-level description of the events occurring in it. Within this framework, the initial processing of video frames, not described in this paper, is performed by trackers and classifiers that output a structured description of the relevant objects. This description then constitutes ProVision's input, which starts inferring higher-level predicates (defined in the ontology) and produces a list of event occurrences detailing which events occur in each video.

2.1 Source Data

The video sequences which constitute the data for the implementation and evaluation of our formalism have been provided by DARPA as the 'development' video dataset. It contains 1302 video sequences in MPEG format, hereafter called *vignettes*, with a resolution of 1280x720 pixels and variable duration between 5 and 20 seconds. Portrayed subjects are mostly humans, vehicles (cars, bicycles and motorbikes) and other objects (boxes, balls and small items). The scene background is an urban outdoor environemnt, such as parks and streets. Each vignette file name contains one of the motion verbs listed in Table 1, and there are between 10 and 30 vignettes per verb.

As mentioned earlier, ProVision does not operate on vignettes directly, but on their *annotation*. An annotation is an XML file in Viper format [10,18,20] containing a structured description of the objects present in each vignette, namely object type ('person', 'vehicle' or 'other') and position of the bounding box covering their shape at each frame. At the moment, we can access two types of annotation files:

- *hand-annotated* data: this has been produced by several human annotators. Each annotator received a set of vignettes, manually specified the coordinates of each oject's bounding box at each frame and added temporally indexed occurrences of each event from Table 1 that they believed to be happening in the vignette.
- *tracked* data: this is generated by trackers and classifier algorithms, implemented by a Vision research group whom we collaborate with. They scan each vignette, detect moving objects and determine bounding boxes' coordinates and object types.

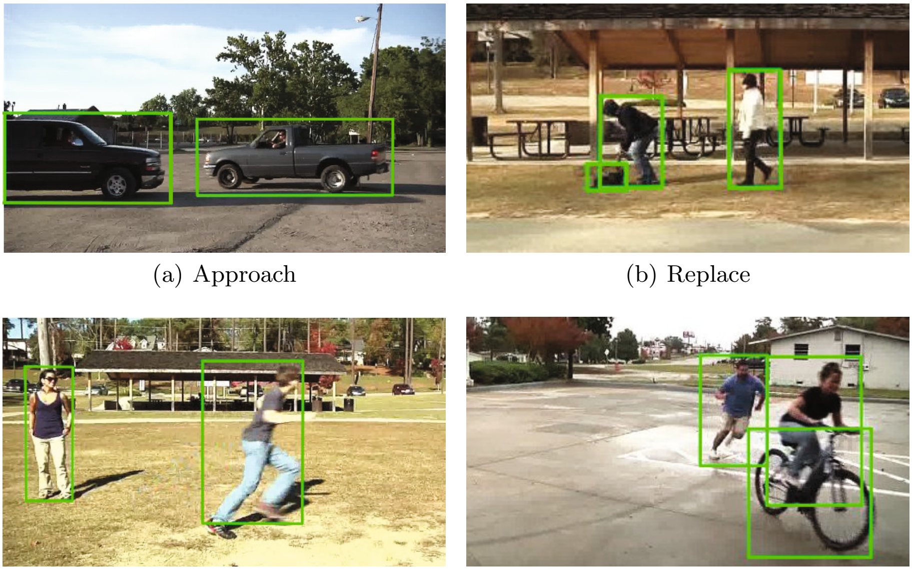
Some vignette and annotation samples are shown in Fig. 1.

Hand-annotated data is mostly noise- and error- free whilst tracked data is likely to contian spurious and/or noisy object tracks. The initial development stage of ProVision did not focus on managing such errors and noise in tracked data, hence we chose to test our ontology on hand-annotated data (see Sec. 5). Hand-annotated data also acts as *ground truth* for evaluation. Ideally, the occurrences returned by ProVision through logical inferences should match the occurrences that human annotators added to the annotation.

2.2 Verb Analysis

Table. 1 includes concepts with varying complexity and difficulty, from simple actions such as Move and Touch, to complex actions such as Exchange and Replace. A brief, informal overview identified some of the semantic properties of interest:

- *Speed of actions.* Some verbs, for example Flee, Run and Snatch, require formalising the qualitative notion of a *fast-paced* action.



(a) Approach

(b) Replace

(c) Flee

(d) Chase

Fig. 1. Vignette samples and tracked objects

- *Trajectories.* There can be ambiguity whether a particular movement trajectory, or part of, can be classified as Fall or Fly. The former suggests a dominant vertical component, the latter a dominant horizontal component (see Fig. 2 for an example).
- *Triggers.* An occurrence of Flee may imply the presence of a trigger, which could either be an object performing a corresponding Follow or Chase, or an object representing a 'danger'.
- *Contact.* Verbs Hit and Collide are very similar in meaning. It appears that distinguishing between the two involves examining how two objects come into contact: the former seems more intentional and through usual or predictable contact parts, whilst the latter suggests a more coincidental occurrence through less usual contact parts.
- *Granularity.* For example, an occurrence of Kick can be subsumed by Hit if some particular level of detail is not available or relevant.
- *Saliency.* Clusters of verbs like Approach, Move and Go, or Have and Hold are similar in meaning, but one of them may be more salient in describing a particular situation.
- *Submeanings.* Some verbs may have different separate sub-meanings: for example Exchange may refer to two agents exchanging position, two agents exchanging one or more objects, or one agent removing an object and replacing it with another.

The above list is by no means exhaustive; in fact, building a formal ontology of motion verbs would require extensive and systematic analysis of the

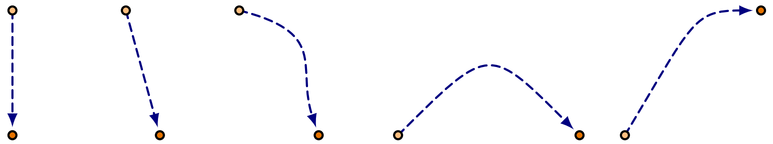


Fig. 2. Fly and Fall motion examples

semantic properties of each verb, also drawing from previous studies in the field of linguistics [24,19]. This analysis would allow us to fully characterise and model each verb within the ontology through the definition of its semantic properties. However, for each verb, most of the semantic properties that characterise its occurrence refer to very specific and fine-grained properties of objects and/or context involved during the occurrence. The problem we face in our domain is that such fine-grained and highly detailed information is not available in the vignette annotations, that constitute the only ground facts on which higher-level semantic properties can be logically inferred.

For example, full characterisation of the verb Approach would be likely to involve reasoning on the relative orientation of two objects approaching each other. However, in practice, the relative orientation of an object can only be guessed by examining the changes in the position of its bounding box over a time interval. Similarly, differentiating the characterisation of Hit and Kick would involve understanding which body parts are involved when two people get into contact, which is a too fine-grained property for its inference from the bounding boxes position to be completely reliable.

This observation directed us to characterise motion verbs through the definition of mid-level properties effectively inferrable from the data available to us, rather than achieving their semantically exhaustive characterisation. In other words, our ontology is more aimed at distinguishing between each verb rather than at fully characterising them.

Some properties between objects that will be relevant to our characterisation will be likely to suggest the integration of other formalisms within our ontology, such as Region Connection Calculus [21] for the analysis of two objects' relative positioning, and Qualitative Trajectory Calculus [25] for the analsysis of objects' trajectories.

3 Issues in Event Classification

There are several issues around the task of event detection and classification. For our ontological approach, we will mainly concentrate on uncertainty, vagueness, context and saliency.

3.1 Uncertainty

As described in the previous section, the input for our detection system is constituted by vignette annotations. Vignette annotations carry a varying degree of uncertainty, particularly high in tracked data, as it is very unlikely that such representation mirrors the amount and complexity of the information that a human eye may gather from watching the vignette. The main issues in regards to our application are:

- *Errors and noise.* An object within a vignette may have an inaccurate representation in the annotation. For example, a frequent issue is the shrinking or disappearance of the bounding box around an object becoming partially or totally occluded by another object. Spurious tracks are an undesirable common feature of tracked data.
- *Missing objects.* There are a number of annotations in which relevant active objects are missing from the annotation. This is problematic when having to infer a predicate involving such objects.
- *Background.* Scene background and most objects not at the scene focus are not reported in the annotation. This can ease processing logical inferences as input data is simpler, but can also toughen interpretation of vague concepts, as such background information may provide evidence to resolve ambiguities.
- *Granularity.* Annotation data is coarse as, for each object and vignette frame, only its bounding box position is available. This is not sufficient for verb characterisations that require detailed knowledge about an object's properties, for example the position of the limbs for events such as Hold, PickUp or Kick.
- *Three-dimensional representation.* Every position is specified by a rectangle, of which only its top-left coordinates, width and height are known. The absence of a coordinate along the z-axis is problematic for all those vignettes in which objects move away or towards the camera.

3.2 Vagueness

A major obstacle to interpreting real world events in terms of natural language vocabulary is vagueness. Vagueness is essentially a linguistic phenomenon which manifests itself when one attempts to formally define certain words and concepts from natural language. It is different from uncertainty: the issues outlined in the previous section arise from limited, insufficient or imprecise knowledge, whereas vagueness has to do with a lack of clear and precise criteria for the applicability of concepts. A comprehensive overview of the origins, nature and characteristics of vagueness is provided in [1,2,14,16,26].

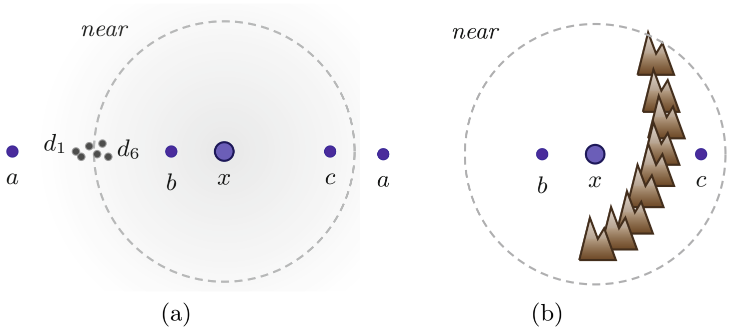
Vagueness may arise from several classes of natural language terms. The most relevant for our domain are: *spatial prepositions* (e.g. near, far, beside, close. . .); *adjectives* (e.g. tall, short, big, small, fast, slow. . .); *verbs* themselves (e.g. approach, chase, exchange. . .) and *nouns* (e.g. group, hill, river. . .).

A common flavour of vagueness is *Sorites Vagueness* [1], witnessed everytime the applicability boundary of a concept is blurred, and depends on one or more *observable properties* showing continuous variations among the sample on which the concept is applicable. Setting an *applicability threshold* on the values of said observables would establish a crisp applicability boundary, thus attaining a complete disambiguation. The problem lies on the fact that, typically, identifying the relevant observable properties and the most appropriate thresholds is not straightforward.

A classic example is the paradox of the *heap of sand* [26], where we are faced with determining what constitutes a heap. Common sense would suggest that, by removing a grain from a heap of sand, one would still be left with a heap. The reiteration of this process though would classify a single grain of sand as a heap, fact that is against common sense. To set things clear, one may then choose to disambiguate the term 'heap' by setting a threshold defining 'heap' as 'a group of $n \geq 10^4$ grains of sand'. Unfortunately, the result is that, for example, a group of $10^4 - 1$ grains of sand would then be unfairly classified as 'not a heap'.

This problem affects the formalisation of our ontology, especially when qualitative concepts are involved. For example, we may want to infer 'a is arriving at b' if a is moving towards b and a is *near* b. This generates the sub-problem of having to define the qualitative predicate *near*. Fig. 3 shows an example where, given object x, we want to determine for which $t \in \{a, b, c, d_1, \ldots, d_6\}$ the predicate $\mathsf{near}(t, x)$ holds. The shaded area around x represents the 'blur' in the validity of near. A precise version of the vague concept near can be attained by choosing a threshold equal to a fixed distance from x, corresponding to the dashed circle in Fig. 3(a). This would cause the undesired effect where $\mathsf{near}(d_6, x)$ would hold whilst $\mathsf{near}(d_1, x)$ would not, despite d_1 and d_6 (and all the other points in between) being in close proximity.

In Fig. 3(b) we show a different example where a, b, x and c represent towns on a map, with a long chain of mountains between x and c. A precisification of near as in the example above would cause $\mathsf{near}(c, x)$ to hold, but not $\mathsf{near}(a, x)$. Somehow, this is counterintuitive because, despite c's shorter linear distance from x, knowledge about the geographical feature would suggest that, on a practical level, a is nearer to x than c.



(a) (b)

Fig. 3. Defining near

3.3 Saliency

Some observed situations can be described by multiple concepts. Often, these concepts present similar and possibly ambiguous meanings, showing variations in strength or granularity (for example pairs of verbs such as Hit and Kick or Follow and Chase). Also, some events may be overshadowed by more relevant ones happening at the same time. For example, a vignette may show two vehicles approaching in the background, and a person hitting another person in the foreground. In this situation, most would agree that the occurrence of Hit is more salient than the occurrence of Approach.

Humans can naturally judge saliency of event occurrences and this is particularly prominent in our hand-annotated data. The data in fact shows marked under-reporting of low-salience events, such as the example above. This proves to be a problem particularly for the evaluation of our system's performance; an event occurrence correctly inferred by ProVision may still be classified as a false positive if such an event is under-reported in the hand-annotated data due to saliency issues (see Sec. 6).

4 Ontology

Our ontology builds upon *Event Calculus* [17,22] and *Versatile Event Logic* (VEL) [5], formalisms designed to reason about actions and events within logic. Given an ordered set of time points $\mathcal{T} = (T, <)$, the calculus expresses that fluent f holds at a particular time point $t \in \mathcal{T}$ with the construct $\mathsf{HoldsAt}(f, t)$.

4.1 Logical Formalism

The vocabulary of our logical language can be specified by the tuple:

$$\mathcal{V} = \langle \mathcal{T}, \mathcal{I}, \mathcal{O}, \mathcal{O}_t, \mathcal{B}, \mathcal{F}, \mathcal{E}, \Sigma \rangle$$

where:

- $\mathcal{T}$ is the set of ordered time points (e.g. $\mathcal{T} = \{t_1,\, t_2 \ldots\}$);
- $\mathcal{I}$ is the set of time intervals (e.g. $i = [t_1, t_2]$);
- $\mathcal{O}$ is the set of objects (event participants);
- $\mathcal{O}_t$ is the set of object types (e.g. $\mathcal{O}_t = \{\text{person}, \text{vehicle}, \text{bicycle} \ldots\}$);
- $\mathcal{B}$ is a set of rectangular Bounding Boxes encircling objects (a generic $b \in \mathcal{B}$ is of the form $b = \langle x, y, w, h\rangle$, where (x, y) is the coordinate of the top-left corner of the rectangle and w and h are, respectively, its width and height);
- $\mathcal{F}$ is a set of fluents (see Sec. 4.2);
- $\mathcal{E}$ is the set of event-types;
- Σ is the set of event-tokens.

We also introduce expressions to manipulate the entities listed above:

- $b(i, t)$ if and only if t is the initial time point of interval i;
- $e(i, t)$ if and only if t is the terminating time point of interval i;
- $i' \subset i$ if and only if interval i' is a proper subset of i;
- $dur(i, \delta)$ if and only if δ is the duration of interval i;
- $t_1 \leq t_2$ if and only if $t_1 = t_2$ or $t_1 < t_2$, according to $\mathcal{T}$ ordering function;
- $type(o, o_t)$ if and only if $o_t \in \mathcal{O}_t$ is the type of $o \in \mathcal{O}$;
- $box_x(b, x)$, $box_y(b, y)$, $box_w(b, w)$ and $box_h(b, h)$ are true if and only if x, y, w and h are respectively the x-coordinate, y-coordinate, width and height of bounding box $b \in \mathcal{B}$;
- $\mathsf{HoldsAt}(bbox(o, b), t)$ is true if and only if $b \in \mathcal{B}$ is the bounding box representing the position of object $o \in \mathcal{O}$ at time point t;
- $EventType(\sigma, e)$ is true if and only if event token $\sigma \in \Sigma$ is of type $e \in \mathcal{E}$.
- $\mathsf{HoldsAt}(pos(o, o_x, o_y), t)$ is true if and only if point (o_x, o_y) represents the position of object $o \in \mathcal{O}$ at time point t (definable in several ways, most commonly (o_x, o_y) is defined as the centroid of the object's bounding box);
- $edist(o_x, o_y, p_x, p_y)$ is a function calculating the euclidean distance between points (o_x, o_y) and (p_x, p_y);
- $\mathsf{HoldsAt}(dist(o_1, o_2, d), t)$ is true if and only if d is the distance between o_1 and o_2 at time point t (definable through pos and an appropriate distance measure such as $edist$).

4.2 Fluents, Processes, Event-Types and Event-Tokens

Vocabulary $\mathcal{V}$ allows for two types of time-dependent formal expressions: propositional expressions whose validity can be stated over time (*fluents*) and expressions referring to temporal entities that occur over some interval (*events*).

A fluent's truth-value may be established at single time points. Fluents describe either a state that may hold or not hold, or a process that may be active or inactive at each time point. Given fluent f and notation $\mathsf{HoldsAt}(f, t)$, it is possible to define $\mathsf{HoldsOver}(f, i)$ to express the validity of f over the interval $i = [t_1, t_2] \in \mathcal{I}$:

$$\mathsf{HoldsOver}(f, [t_1, t_2]) \equiv \forall\, t\big[(t_1 \leq t \leq t_2) \rightarrow \mathsf{HoldsAt}(f, t)\big] \tag{1}$$

If $\mathsf{HoldsOver}(f, [t_1, t_2])$ is true for some t_1, t_2, from the definition above it follows that $\mathsf{HoldsOver}(f, [t_i, t_j])$ is also true, for every $[t_i, t_j] \subseteq [t_1, t_2]$. A predicate holding only on the largest interval is $\mathsf{HoldsOn}(f, i)$, which is true if and only if i is the greatest continuous temporal interval over which f is true, i.e. there does not exist $i' \supset i$ such that $\mathsf{HoldsOn}(f, i')$:

$$\begin{aligned}\mathsf{HoldsOn}(f, [t_1, t_2]) \equiv{}& \mathsf{HoldsOver}(f, [t_1, t_2]) \;\wedge \\ &\wedge \exists\, t_1'\Big[t_1' < t_1 \wedge \forall\, t'\big[t_1' < t' < t_1 \rightarrow \neg\mathsf{HoldsAt}(f, t')\big]\Big] \wedge \\ &\wedge \exists\, t_2'\Big[t_2 < t_2' \wedge \forall\, t'\big[t_2 < t' < t_2' \rightarrow \neg\mathsf{HoldsAt}(f, t')\big]\Big]\end{aligned} \tag{2}$$

An event represents a complex action and we distinguish between *event-types* and *event-tokens* [5]. An event-type $e \in \mathcal{E}$ is associated with a set of episodes of a particular event, for example: 'John approaches Mary', formalised as $\mathsf{Approach}(\text{John}, \text{Mary})$. An event-token $\sigma \in \Sigma$ constitutes an occurrence of a particular event-type over a temporal interval. To express the occurrence of event type $e \in \mathcal{E}$ over time interval $i \in \mathcal{I}$ we introduce the construct $\mathsf{Occurs}(e, i)$. The definition of an event occurrence often involves specifying a particular sequence of fluents or sub-events that has to hold for the event to occur.

For clarity of notation in the formulae to follow, event predicates are capitalised as in $\mathsf{Approach}$, whilst fluents are lowercase as in $\mathsf{approaching}$.

4.3 Precisifications

In order to provide for the disambiguation of vague concepts, we enrich our language with some ideas from Supervaluation Semantics [13,15], though a few other approaches to vagueness in logics can be found in the literature, such as the Egg-Yolk model [7] or the more popular Fuzzy Logic [27].

The theoretical account of supervaluationism states that a formula may admit multiple models, each obtainable via an assignment of referents to terms and truth-values to predicates. Such an assignment is called a *precisification*, and allows to obtain a precise interpretation of a vague term. The advantage of supervaluationism over multi-valued logics, such as Fuzzy Logic, is the preservation of classical logic inference rules.

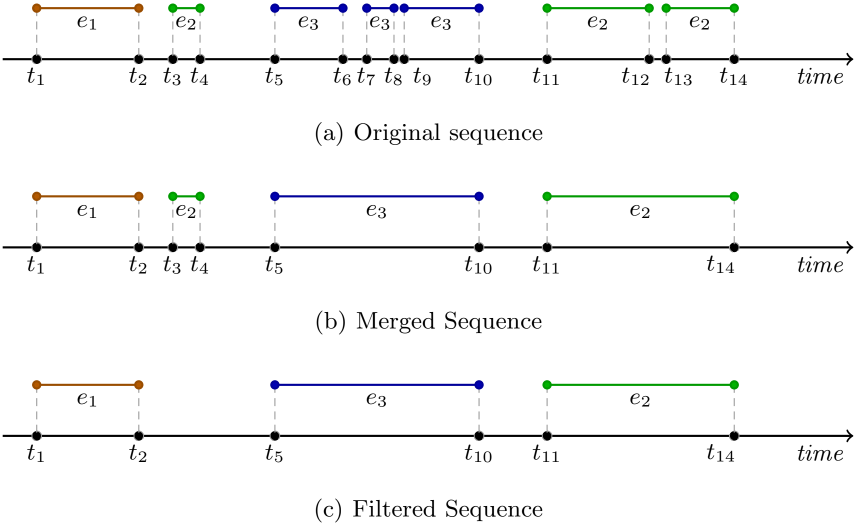
Standpoint semantics [3,4,6] is an elaboration of supervaluation semantics where the precisification is explicitly embedded in the language syntax. In Section 3 we argued that most vague concepts can be disambiguated by identifying an *applicability threshold* for some observable property relevant to the concept. For example: $near(a, b)$ could be made precise by specifying a threshold on the distance between a and b. In standpoint semantics, formal definitions of vague terms are parameterised with these thresholds thus becoming fully precise, as the following sample definition of *near* shows:

$$\mathsf{HoldsAt}(near[\delta](a, b), t) \equiv \mathsf{HoldsAt}(dist(a, b, x), t) \wedge x < \delta \qquad (3)$$

Following this idea, some of the definitions in the following sections have been parameterised with thresholds specifying the applicability of predicates, with a particular choice of thresholds forming a *precisification*. This approach proves particularly useful when testing the system, since the values inside a precisification will have to be fine-tuned to yield maximum accuracy. We also believe that this method allows for future reasoning on thresholds themselves and automatic optimisation of precisifications, even though this extension lies outside the scope of this paper.

4.4 Occurrence Smoothing

Testing the implementation of this ontology on video annotations is likely to produce isolated or interrupted inferences of event occurrences, due to the experimental nature of the data. In this section we introduce temporal-indexing


e_1
e_2
e_3
e_3
e_3
e_2
e_2
t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9 t_{10} t_{11} t_{12} t_{13} t_{14}
time
e_1
e_2
e_3
e_2
t_1 t_2 t_3 t_4 t_5 t_{10} t_{11} t_{14}
time
e_1
e_3
e_2
t_1 t_2 t_5 t_{10} t_{11} t_{14}
time

(a) Original sequence

(b) Merged Sequence

(c) Filtered Sequence

Fig. 4. Merge and filter feature

constructs that extend the truth value of a particular predicate over small temporal gaps of a long occurrence span, and falsify isolated occurrences, likely to be spurious. The idea is illustrated in Fig. 4. These constructs are parameterised with thresholds δ_m and δ_f that fine-tune the merging and filtering of events.

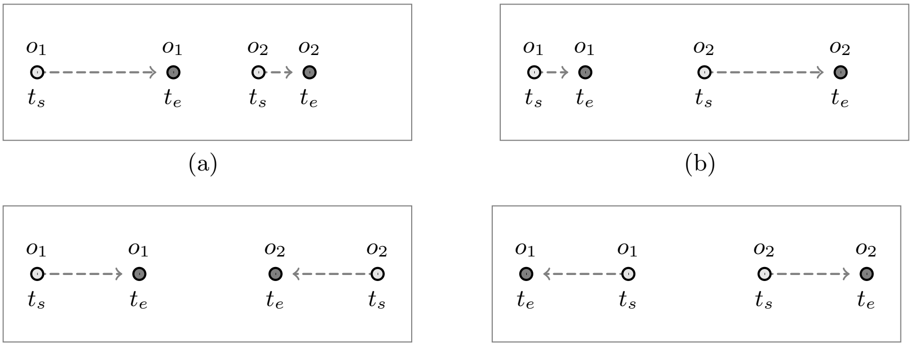
The first stage, corresponding to the transition between Fig. 4(a) and 4(b), is given by the new constructs HoldsAtM and OccursM. The former establishes that fluent f holds at time points part of an interval $[t_1, t_2]$ smaller than threshold δ_m where fluent f holds at both t_1 and t_2. The latter joins separate occurrences of a single event-type separated by an interval smaller than δ_m:

$$\begin{aligned}\mathsf{HoldsAtM}[\delta_m](f,t) \equiv \exists\, t_1, t_2[t_1 \leq t \leq t_2\, \wedge \\ \wedge\, \mathsf{HoldsAt}(f,t_1) \wedge \mathsf{HoldsAt}(f,t_2) \wedge \mathrm{dur}([t_1,t_2]),\delta) \wedge \delta < \delta_m]\end{aligned} \tag{4}$$

$$\begin{aligned}\mathsf{OccursM}(e,[t_s,t_e]) \equiv \mathsf{Occurs}(e,[t_s,t_e]) \vee \exists\, t_1,t_2,\delta\big[(t_s < t_1 < t_2 < t_e) \,\wedge \\ \wedge\, \mathsf{Occurs}(e,[t_s,t_1]) \wedge \mathsf{Occurs}(e,[t_2,t_e]) \wedge dur([t_1,t_2],\delta) \wedge \delta < \delta_m\big]\end{aligned} \tag{5}$$

The second stage, corresponding to the transition between Fig. 4(b) and 4(c), considers the occurrences resulting from the 'merge' stage, as defined above, and filters isolated occurrences of very little duration. This is achieved by the two predicates HoldsAtF and OccursF that hold only if the fluent (resp. event-type) holds (resp. occurs) on an interval with duration greater than δ_f:

$$\begin{aligned}\mathsf{HoldsAtF}[\delta_f](f,t) \equiv \exists\, t_1,t_2,\delta\ [t_1 \leq t \leq t_2 \wedge \mathrm{dur}([t_1,t_2],\delta) \wedge \delta > \delta_f\, \wedge \\ \wedge\, \forall\, t'[(t_1 \leq t' \leq t_2) \rightarrow \mathsf{HoldsAtM}(f,t')]]\end{aligned} \tag{6}$$



(a)

(b)

(c)

(d)

Fig. 5. Positions of objects o_1, o_2 over interval $[t_s, t_e]$

$$\mathsf{OccursF}[\delta_f](e, [t_s, t_e]) \equiv \mathsf{OccursM}(e, [t_s, t_e]) \wedge \exists\, \delta\big[dur([t_s, t_e], \delta) \wedge \delta > \delta_f\big] \quad (7)$$

The constructs HoldsOverM, HoldsOnM, HoldsOverF and HoldsOnF can be easily defined by adjusting the definitions in (1) and (2).

4.5 The Verb Approach

Let us consider two objects $o_1, o_2 \in \mathcal{O}$ moving in space over the interval $[t_s, t_e] \in \mathcal{I}$. Fig. 5 shows some possible positions for o_1 and o_2 at t_s and t_e. An average observer would most likely assert the following:

- In Fig. 5(a), o_1 is approaching o_2, but o_2 is not approaching o_1 as o_2 is moving in the opposite direction;
- In Fig. 5(b), despite o_1 moving towards o_2, o_1 is not approaching o_2 as their distance does not decrease;
- In Fig. 5(c), o_1 is approaching o_2 *and* o_2 is approaching o_1 as they are both actively moving towards each other, and their distance decreases as well;
- In Fig. 5(d), clearly neither o_1 is approaching o_2 nor o_2 is approaching o_1 as they are heading towards opposite directions.

This is a simplification of the relative movement for two generic objects in space and does not take into account futher semantic properties that may be relevant. This is in line with the methodology outlined in Sec. 2.2.

We define the occurrence of event-type Approach and fluent approaching through the fluents getCloser and moveTowards:

$$\mathsf{Occurs}(\mathsf{Approach}(o_1, o_2), [t_s, t_e]) \equiv \mathsf{HoldsOn}(\mathsf{approaching}(o_1, o_2), [t_s, t_e]) \quad (8)$$

$$\begin{aligned} &\mathsf{HoldsAt}(\mathsf{approaching}(o_1, o_2), t) \equiv \\ &\qquad \mathsf{HoldsAt}(\mathsf{moveTowards}(o_1, o_2), t) \wedge \mathsf{HoldsAt}(\mathsf{getCloser}(o_1, o_2), t) \end{aligned} \quad (9)$$

The fluent getCloser holds at time point t if and only if the distance between two objects $o_1, o_2 \in \mathcal{O}$ over an interval surrounding t monotonically decreases:

$$\begin{aligned}&\mathsf{HoldsAt}(\mathsf{getCloser}(o_1, o_2), t) \equiv \\ &\quad \exists\, t_s, t_e, \big[(t_s < t < t_e) \wedge \forall\, t', t''\big[(t_s \leq t' < t'' \leq t_e) \rightarrow \exists\, d', d'' \\ &\quad \big[\mathsf{HoldsAt}(\mathit{dist}(o_1, o_2, d'), t') \wedge \mathsf{HoldsAt}(\mathit{dist}(o_1, o_2, d''), t'') \wedge d'' < d'\big]\,\big]\,\big]\end{aligned} \tag{10}$$

The fluent moveTowards holds at time point t if and only if the distance between o_1 and the start point of o_2 monotonically decreases over an interval surrounding t:

$$\begin{aligned}&\mathsf{HoldsAt}(\mathsf{moveTowards}(o_1, o_2), t) \equiv \exists\, t_s, t_e, x_2, y_2\big[(t_s < t < t_e) \wedge \\ &\quad \wedge \mathsf{HoldsAt}(\mathrm{pos}(o_2, x_2, y_2), t_s) \wedge \forall\, t', t''\big[(t_s \leq t' < t'' \leq t_e) \rightarrow \exists\, d', d'', \\ &\quad x_1', y_1', x_1'', y_1''\big[\mathsf{HoldsAt}(\mathrm{pos}(o_1, x_1', y_1'), t') \wedge \mathsf{HoldsAt}(\mathrm{pos}(o_1, x_1'', y_1''), t'') \wedge \\ &\quad \wedge \mathit{edist}(x_1', y_1', x_2, y_2) = d' \wedge \mathit{edist}(x_1'', y_1'', x_2, y_2) = d'' \wedge d'' < d'\big]\,\big]\big]\end{aligned} \tag{11}$$

The definitions above can be parameterised as mentioned in Sec. 4.3. We introduce thresholds T_w, representing the *detection window* over which the validity of a particular predicate is tested, and T_s, representing the minimum speed at which objects must be moving (particularly useful to prune erroneously inferred occurences due to minimal object movement likely to be caused by noisy data):

$$\begin{aligned}&\mathsf{HoldsAt}(\mathsf{getCloser}[T_w, T_s](o_1, o_2), t) \equiv \exists t_s, t_e, d_s, d_e, \delta\ \big[(t_s < t < t_e) \wedge \\ &\quad \wedge \mathit{dur}([t_s, t], T_w) \wedge \mathit{dur}([t, t_e], T_w) \wedge \mathsf{HoldsAt}(\mathit{dist}(o_1, o_2, d_s), t_s) \wedge \\ &\quad \wedge \mathsf{HoldsAt}(\mathit{dist}(o_1, o_2, d_e), t_e) \wedge \mathit{dur}([t_s, t_e], \delta) \wedge \frac{d_s - d_e}{\delta} > T_s\big]\end{aligned} \tag{12}$$

$$\begin{aligned}&\mathsf{HoldsAt}(\mathsf{moveTowards}[T_w, T_s](o_1, o_2), t) \equiv \exists\, t_s, t_e, x_{1e}, y_{1e}, x_{2s}, y_{2s}, d_s, d_e, \delta \\ &\quad \big[(t_s < t < t_e) \wedge \mathit{dur}([t_s, t], T_w) \wedge \mathit{dur}([t, t_e], T_w) \wedge \mathit{dur}([t_s, t_e], \delta) \\ &\quad \mathsf{HoldsAt}(\mathit{pos}(o_1, x_{1e}, y_{1e}), t_e) \wedge \mathsf{HoldsAt}(\mathit{pos}(o_2, x_{2s}, y_{2s}), t_s) \wedge \\ &\quad \mathsf{HoldsAt}(\mathit{dist}(o_1, o_2, d_s), t_s) \wedge \mathrm{edist}(x_{1e}, y_{1e}, x_{2s}, y_{2s}) = d_e \wedge \frac{d_s - d_e}{\delta} > T_s\big]\end{aligned} \tag{13}$$

$$\begin{aligned}&\mathsf{HoldsAt}(\mathsf{approaching}[win, s](o_1, o_2), t) \equiv \\ &\quad \mathsf{HoldsAt}(\mathsf{moveTowards}[win, s](o_1, o_2), t) \wedge \\ &\quad \wedge \mathsf{HoldsAt}(\mathsf{getCloser}[win, s](o_1, o_2), t)\end{aligned} \tag{14}$$

4.6 The Verb Hold

The meaning of Hold in our context is that a person is carrying or supporting an object with his/her hands, and the position of the person is mostly stationary. In our domain, given objects $o_1, o_2 \in \mathcal{O}$, we can intuitively define that $\mathsf{Hold}(\mathsf{o_1}, \mathsf{o_2})$ holds at time point t if o_1 is of type 'Person', o_2 is of type 'Other' and object o_2 is within o_1's range and in contact with o_1's hands.

However, given the nature of the available data (see Sec. 2.1), the position of o_1's hands cannot be extracted easily, hence we have to resort to a simplification of the above intuitive definition. The following definitions are parameterised according to the thresholds T_h, B_h, L_w, R_w and Tol which have values between 0 and 1, and are used to infer o_2's positioning relative to o_1. We define event-type Hold and fluent hold through the fluent $\mathsf{holdingPosition}$:

$$\begin{aligned}&\mathsf{Occurs}(\mathsf{Hold}[T_h, B_h, L_w, R_w, Tol](o_1, o_2), [t_s, t_e]) \equiv \\ &\qquad \mathsf{HoldsOn}(\mathsf{hold}[T_h, B_h, L_w, R_w, Tol](o_1, o_2), [t_s, t_e])\end{aligned} \tag{15}$$

$$\begin{aligned}&\mathsf{HoldsAt}(\mathsf{hold}[T_h, B_h, L_w, R_w, Tol](o_1, o_2), t) \equiv \\ &\qquad type(o_1, \text{person}) \wedge type(o_2, \text{other}) \wedge \\ &\qquad \wedge \mathsf{HoldsAt}(\mathsf{holdingPosition}[T_h, B_h, L_w, R_w, Tol](o_1, o_2), t)\end{aligned} \tag{16}$$

The fluent $\mathsf{holdingPosition}(\mathsf{o_1}, \mathsf{o_2})$ holds at time point t if and only if the bounding boxes of o_1 and o_2 are positioned a way suggesting that o_2 is in contact and within reach of o_1. Thresholds T_h and B_h constrain uppermost and lowermost position of o_2's box with respect to the height of o_1, likewise L_w and R_w constrain the leftmost and rightmost position of o_2 with respect to the width of o_1. Threshold Tol stretches the above contraints according to the size of o_2.

$$\begin{aligned}&\mathsf{HoldsAt}(\mathsf{holdingPosition}[T_h, B_h, L_w, R_w, Tol](o_1, o_2), t) \equiv \\ &\qquad \mathsf{HoldsAt}(bbox(o_1, b_1), t) \wedge \mathsf{HoldsAt}(bbox(o_2, b_2), t) \wedge \\ &\qquad \wedge box_x(b_1, x_1) \wedge box_y(b_1, y_1) \wedge box_w(b_1, w_1) \wedge box_h(b_1, h_1), \wedge \\ &\qquad \wedge box_x(b_2, x_2) \wedge box_y(b_2, y_2) \wedge box_w(b_2, w_2) \wedge box_h(b_2, h_2), \wedge \\ &\qquad \wedge [(y_2 + h_2) - (y_1 + B_h h_1)] < Tol \cdot h_2 \wedge \\ &\qquad \wedge [(y_1 + T_h h_1) - y_2] < Tol \cdot h_2 \wedge \\ &\qquad \wedge [(x_1 + L_w w_1) - x_2] < Tol \cdot w_2 \wedge \\ &\qquad \wedge [(x_2 + w_2) - (x_1 + R_w w_1)] < Tol \cdot w_2\end{aligned} \tag{17}$$

5 Experimental Results

The event detection system ProVision is a Prolog implementation of the ontology outlined in the previous section. ProVision produces event inferences by grounding initial base predicates with the information extracted from the annotation files. Its performance is measured through the comparison of the inferred event occurrences with the event occurrences in the hand-annotated data (see Sec. 2).

After running the event inference on a particular vignette, each frame f is placed in one of the following sets:

- TP (True Positives): if f is within the span of an inferred event occurrence *also* in the ground truth;
- FP (False Positives): if f is within the span of an inferred event occurrence *not* in the ground truth;
- TN (True Negative): if no inferred event occurrences nor occurrences in the ground truth involve f;
- FN (False Negative): if f is within the span of a ground truth occurrence but ProVision prduced no inferred occurrence involving f.

At the end of the statistic calculation, each set is such that $|TP| + |FP| + |TN| + |FN| = T$ where T is the total number of frames in the vignette. The measures of Precision, Recall, Fvalue, MCC (Matthews Correlation Coefficient) and occurrence rates $TP^{\%}$, $FP^{\%}$, $TN^{\%}$, $FN^{\%}$ are calculated:

$$\text{Prec} = \frac{|TP|}{|TP| + |FP|} \qquad \text{Rec} = \frac{|TP|}{|TP| + |FN|} \qquad \text{Fvalue} = 2 \cdot \frac{\text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}}$$

$$\text{MCC} = \frac{(|TP| \cdot |TN|) - (|FP| \cdot |FN|)}{\sqrt{(|TP| + |FP|)(|TP| + |FN|)(|TN| + |FP|)(|TN| + |FN|)}}$$

$$TP^{\%} = \frac{|TP_i|}{|TP_i| + |FN_i|} \qquad FP^{\%} = \frac{|FP_i|}{|TN_i| + |FP_i|}$$

$$TN^{\%} = \frac{|TN_i|}{|TN_i| + |FP_i|} \qquad FN^{\%} = \frac{|FN_i|}{|TP_i| + |FN_i|} \tag{18}$$

The values for Prec and Rec are set to 1 if the denominator is 0. Values for Prec and Rec range between 0 and 1, whilst values for MCC range between −1 and 1.

An overall system accuracy figure for the detection of verb v over a set of n vignettes is obtained by summing the values TP, FP, TN and FN obtained for each vignette, thus calculating the number of true/false positives/negatives for the recognition of verb v across the set. The statistic measures above are then computed on these total values.

5.1 Sample Statistics and Baseline Accuracy

Tracked data tends to abund with error and noise; the detection tests discussed in this section have been carried out on hand-annotated data, as ProVision is not yet fully capable of managing tracked data effectively.

Tests on verbs Approach and Hold have been run on two sets of vignettes:

- *Whole set.* This set contains all 1302 vignettes in the development dataset.
- *Restricted set.* This set contains only the vignettes whose hand-annotated annotation file reports an occurrence of the event being tested.

Sample statistics detailing the frequency of occurrence of the verbs in question are shown in Table 2 where, for each verb and set, the total number of frames

Table 2. Sample statistics

Verb	Set	Vignettes	Frames	Positives	Negatives	PosRate	NegRate
Approach	Whole	1302	595,110	5,254	589,856	0.88 %	99.12 %
Approach	Restricted	70	33,340	5,254	28,086	15.76 %	84.24 %
Hold	Whole	1302	595,110	26,034	569,076	4.37 %	95.63 %
Hold	Restricted	98	42,682	26,034	16,648	61.00 %	39.00 %

across the vignettes, the number of Positive and Negative frames (i.e. the frames within or outside the span of a hand-annotated occurrence of the verb) and their occurrence rates PosRate and NegRate are reported.

Baseline detection accuracy statistics reported in Table 3 have been calculated by implementing three very simple baseline detection algorithms, :

- *All* algorithm. Approach(o_1, o_2) holds at every frame with two distinct objects o_1 and o_2. Hold(o_1, o_2) holds at every frame with two objects o_1 and o_2.
- *Some* algorithm. Approach(o_1, o_2) holds at every frame with two distinct objects o_1 and o_2 and where o_1 is moving. Hold(o_1, o_2) holds at every frame with two distinct objects o_1 and o_2 and where o_1 is of type 'Person' and o_2 of type 'Other'.
- *None* algorithm. Approach(o_1, o_2) and Hold(o_1, o_2) never hold at any frame.

5.2 Detection Results

The detection tests on verbs Approach and Hold have been run several times in order to test precisifications yielding the best results. A precisification P specifies the following thresholds:

$$P = [T_w, T_s, T_h, B_h, L_w, R_w, \mathit{Tol}, \delta_m, \delta_f]$$

where thresholds T_w and T_s specify detection window and minimum speed for detecting Approach (see Sec. 4.5), thresholds T_h, B_h, L_w, R_w and Tol specify object positioning constraints for detecting Hold (see Sec. 4.6) and thresholds δ_m and δ_f specify how to perform occurrence smoothing (see Sec. 4.4).

Experimental results are reported in Table 4, which reports Approach detection results for increasing values of threshold δ_f, and Table 5, which reports Hold detection results showing precisifications yielding maximum value for the underlined statistic.

ROC curve graphs showing the overall detection accuracy over different precisifications are shown in Fig. 6. Each dot on the graph represents a couple of $TP^{\%}$ and $FP^{\%}$ values associated with the event detection results for a specific choice of thresholds. In general, precisifications yielding high $TP^{\%}$ values have the undesirable effect of yielding high $FP^{\%}$ values too; figures showing point concentrations skewed towards topmost and leftmost areas of the graph denote algorithms with good overall performances.

Table 3. Baseline accuracy

Verb	Set	Bl	Prec	Rec	Fv	MCC	$TP^{\%}$	$FP^{\%}$	$TN^{\%}$	$FN^{\%}$
Approach	W	A	0.010	0.969	0.020	0.028	96.90	86.69	13.32	3.10
Approach	W	S	0.016	0.917	0.031	0.077	91.74	50.36	49.64	8.26
Approach	W	N	1.000	0.000	0.000	0.000	0.00	0.00	100.00	100.00
Approach	R	A	0.159	0.969	0.273	0.023	96.90	95.61	4.39	3.10
Approach	R	S	0.280	0.917	0.429	0.347	91.74	44.16	55.84	8.26
Approach	R	N	1.000	0.000	0.000	0.000	0.00	0.00	100.00	100.00
Hold	W	A	0.044	1.000	0.084	0.012	100.00	99.68	0.32	0.00
Hold	W	S	0.068	0.750	0.125	0.116	75.01	46.71	53.29	24.99
Hold	W	N	1.000	0.000	0.000	0.000	0.00	0.00	100.00	100.00
Hold	R	A	0.610	1.000	0.758	0.000	100.00	100.00	0.00	0.00
Hold	R	S	0.605	0.750	0.670	−0.019	75.01	76.65	23.35	24.99
Hold	R	N	1.000	0.000	0.000	0.000	0.00	0.00	100.00	100.00

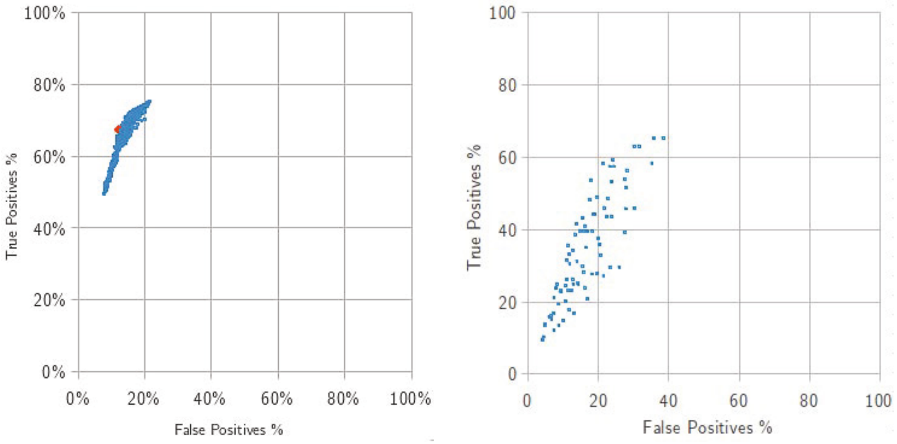
Bl: baseline algorithm (All, Some or None);
Set: vignette set (Whole or Restricted).

Table 4. Approach Detection Results

Set	T_w	T_s	δ_m	δ_f	Prec	Rec	Fv	Mcc	$TP^{\%}$	$FP^{\%}$	$TN^{\%}$	$FN^{\%}$
W	10	0.2	15	30	0.042	0.673	0.079	0.144	67.26	13.64	86.36	32.74
W	10	0.2	15	40	0.047	0.625	0.087	0.149	62.50	11.30	88.70	37.50
W	10	0.2	15	50	0.052	0.575	0.096	0.152	57.46	9.27	90.73	42.54
W	10	0.2	15	60	0.057	0.524	0.103	0.153	52.38	7.69	92.31	47.62
W	10	0.2	15	70	0.060	0.459	0.106	0.147	45.89	6.45	93.55	54.11
W	10	0.2	15	80	0.062	0.417	0.108	0.143	41.66	5.59	94.41	58.34
R	10	0.2	15	30	0.506	0.673	0.578	0.492	67.26	12.27	87.73	32.74
R	10	0.2	15	40	0.507	0.625	0.560	0.471	62.51	11.37	88.63	37.50
R	10	0.2	15	50	0.513	0.575	0.542	0.452	57.46	10.19	89.81	42.54
R	10	0.2	15	60	0.520	0.524	0.522	0.432	52.38	9.03	90.97	47.62
R	10	0.2	15	70	0.520	0.459	0.488	0.400	45.89	7.93	92.07	54.11
R	10	0.2	15	80	0.505	0.417	0.457	0.369	41.66	7.64	92.36	58.34

Table 5. Hold Detection Results

Set	T_h	B_h	L_w	Tol	Prec	Rec	Fv	Mcc	$TP^{\%}$	$FP^{\%}$	$TN^{\%}$	$FN^{\%}$
W	0.25	0.85	−0.1	0.2	$\underline{0.153}$	0.415	0.224	0.196	41.47	10.50	89.50	58.53
W	0.25	1.15	−0.5	0.3	0.096	$\underline{0.653}$	0.167	0.166	65.30	28.16	71.84	34.70
W	0.25	0.85	−0.1	0.3	0.145	0.537	$\underline{0.228}$	0.217	53.71	14.52	85.48	46.30
W	0.25	0.85	−0..25	0.3	0.142	0.582	0.228	$\underline{0.225}$	58.17	16.07	83.93	41.83
R	0.35	0.85	−0.1	0.3	$\underline{0.830}$	0.357	0.499	0.269	35.69	11.40	88.60	64.31
R	0.25	1.15	−0.5	0.3	0.727	$\underline{0.653}$	0.688	0.264	65.30	38.33	61.67	34.70
R	0.25	1.00	−0.5	0.3	0.740	0.652	$\underline{0.693}$	0.287	65.23	35.85	64.15	34.77
R	0.25	0.85	−0.25	0.3	0.811	0.582	0.678	$\underline{0.363}$	58.17	21.25	78.75	41.83



(a) Approach – Restricted set (b) Hold – Restricted set

Fig. 6. ROC curves for detection statistics

6 Discussion and Further Work

The results illustrated in Sec. 5.2 show that ProVision managed to detect 67.26% of true positive frames (against 12.27% of false positives) for the verb Approach and 65.30% of true positive frames (against 38.33% of false positives) for Hold across the restricted set of vignettes. These detection rates yield Precision and Recall figures of 0.506 and 0.673 for Approach and 0.727 and 0.653 for Hold.

When the whole set of vignettes is considered, true and false positives rates do not show significant variations, while Precision and MCC figures decrease sensibly. This is due to the distribution of event occurrences in hand-annotated data shown in Table 2. In fact, occurrences of Approach and Hold only involve 0.88% and 4.37% of frames respectively. Given this occurrence rate, even small $FP^{\%}$ values yield high a high number of false positives, hence the rapid deterioration of Precision, MCC and Fv values.

After careful examination of the vignettes, we believe that sample statistics for Approach are affected by under-reporting and/or inconsistent reporting of the event in hand-annotated data. We have observed several examples in which human annotators did not report occurrences of Approach in situations where more salient and semantically richer events dominate the scene foreground. An easy but time-consuming solution to this problem would be to proceed with vignette re-annotation, ensuring consistent and uniform reporting of events. A more interesting but decidedly challenging approach would involve incorporating a method to establish event saliency within the ontology. This capability would have ProVision discarding event occurrences deemed not salient, thus resembling the behaviour of a generic human annotator.

Not surprisingly, the verb Hold is not affected by under-reporting to the same extent of Approach. The high false positives detection rate is rather caused by the fact that an occurrence of Hold is inherently more difficult to detect when only the position of two objects' bounding boxes is given, hence the relatively high number of cases where two objects' relative positioning is mistaken for an occurrence of Hold.

Especially for Approach, most false negatives are associated to particularly difficult instances, or vignettes in which objects move along the z-axis (i.e. the direction towards the camera). We believe this issue is pervasive and of foremost importance, and are currently working towards a solution that infers the z-coordinate by observing changes in the height of the bounding boxes.

The strength of our approach is given by the capability of an ontology to integrate a detailed semantic characterisation of concepts, and to allow for immediate augmentation of detection capabilities by enriching the ontology with additional definitions. The issues affecting ProVision, which represents the ontology application, are mainly due to the difficulty of establishing whether particular concepts hold when examining real data. In fact, an ontology may specify many detailed semantic characteristics of a concept; however, manifestations of many of these characteristics may be extremely challenging to detect on real coarse-grained data, as our research demonstrates.

References

1. Bennett, B.: Modes of Concept Definition and Varieties of Vagueness. Applied Ontology 1(1), 17–26 (2005)
2. Bennett, B.: Spatial Vagueness. In: Jeansoulin, R., Papini, O., Prade, H., Schockaert, S. (eds.) Methods for Handling Imperfect Spatial Information. STUDFUZZ, vol. 256, pp. 15–47. Springer, Heidelberg (2010)
3. Bennett, B.: Possible Worlds and Possible Meanings: a Semantics for the Interpretation of Vague Languages. In: Commonsense 2011: Tenth International Symposium on Logical Formalizations of Commonsense Reasoning. AAAI Spring Symposium. AAAI, Stanford University (2011)
4. Bennett, B.: Standpoint Semantics: a framework for formalising the variable meaning of vague terms. In: Understanding Vagueness - Logical, Philosophical and Linguistic Perspectives. College Publications (in press, 2012)
5. Bennett, B., Galton, A.P.: A Unifying Semantics for Time and Events. Artificial Intelligence 153(1-2), 13–48 (2004), http://www.comp.leeds.ac.uk/qsr/pub/AIJ_Bennett_Galton_VEL.pdf
6. Bennett, B., Mallenby, D., Third, A.: An Ontology for Grounding Vague Geographic Terms. In: FOIS, pp. 280–293 (2008)
7. Cohn, A.G., Gotts, N.M.: Representing Spatial Vagueness: a Mereological Approach. In: Aiello, L.C., Doyle, J., Shapiro, S.C. (eds.) Proceedings of the 5th International Conference on Principles of Knowledge Representation and Reasoning (KR 1996), pp. 230–241 (1996)
8. DARPA: Defense Advanced Research Projects Agency: Mind's Eye Project Homepage, http://www.darpa.mil/Our_Work/I2O/Programs/Minds_Eye.aspx (last visited January 2012)
9. DARPA: Defense Advanced Research Projects Agency: DARPA Mind's Eye Program. Project specification (March 2010)

10. Doermann, D., Mihalcik, D.: Tools and Techniques for Video Performance Evaluation. In: Proceedings of the 15th International Conference on Pattern Recognition, vol. 4, pp. 167–170 (2000)
11. Dubba, K.S.R.: Learning Relational Event Models from Video. Ph.D. thesis (2012)
12. Dubba, K.S.R., Cohn, A.G., Hogg, D.C.: Event Model Learning from Complex Videos using ILP. In: Proceedings of ECAI 2010 - 19th European Conference on Artificial Intelligence, Lisbon, Portugal, August 16-20. Frontiers in Artificial Intelligence and Applications, vol. 215, pp. 93–98. IOS Press (2010)
13. Fine, K.: Vagueness, truth and logic. Synthese 30, 263–300 (1975)
14. Galton, A.: Spatial and Temporal Knowledge Representation. Earth Science Informatics 3(3), 169–187 (2009)
15. Keefe, R.: Vagueness: supervaluationism. Philosophy Compass 3(2), 315–324 (2008)
16. Keefe, R., Smith, P.: Vagueness: a reader. MIT Press (1997)
17. Kowalski, R.A., Sergot, M.J.: A Logic-based Calculus of Events. New Generation Computing 4(1), 67–95 (1986)
18. Language and Media Processing Laboratory: ViPER: The Video Performance Evaluation Resource, http://viper-toolkit.sourceforge.net (last visited January 2012)
19. Levin, B.: English Verb Classes and Alternations - A Preliminary Investigation. The University of Chicago Press (1993)
20. Mariano, V.Y., Min, J., Park, J.-H., Kasturi, R., Mihalcik, D., Li, H., Doermann, D., Drayer, T.: Performance Evaluation of Object Detection Algorithms. In: Proceedings of the 16th International Conference on Pattern Recognition, vol. 3, pp. 965–969 (2002)
21. Randell, D.A., Cui, Z., Cohn, A.G.: A Spatial Logic Based on Regions and Connection. In: Nebel, B., Rich, C., Swartout, W.R. (eds.) Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning (KR 1992), October 25-29, pp. 165–176. Morgan Kaufmann, Cambridge (1992)
22. Shanahan, M.: The Event Calculus Explained. In: Veloso, M.M., Wooldridge, M.J. (eds.) Artificial Intelligence Today. LNCS (LNAI), vol. 1600, pp. 409–430. Springer, Heidelberg (1999)
23. Sridhar, M., Cohn, A.G., Hogg, D.C.: Unsupervised Learning of Event Classes from Video. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, pp. 1631–1638 (2010)
24. Vendler, Z.: Verbs and Times. The Philosophical Review 66(2), 143–160 (1957)
25. van de Weghe, N., Cohn, A.G., Tre, G.D., Maeyer, P.D.: A Qualitative Trajectory Calculus as a Basis for Representing Moving Objects in Geographical Information Systems. Control and Cybernetics 35(1), 97–119 (2006)
26. Williamson, T.: Vagueness. The Problems of Philosophy, 1998 edn. Routledge (1994)
27. Zadeh, L.A.: Fuzzy Logic and Approximate Reasoning. Synthese 30(3-4), 407–428 (1975)

Relevance in Spatial Navigation and Communication

Thora Tenbrink

SFB/TR 8 Spatial Cognition, University of Bremen, Germany
tenbrink@uni-bremen.de

Abstract. Humans are incapable of reproducing exact "copies" of reality when conceptualizing and communicating about space. Instead, those aspects of a spatial environment are represented that are relevant for a particular purpose. This paper addresses how cognition and communication of space are affected by relevance. It starts out with a review of a broad range of evidence from theories and empirical findings on relevance in perception and communication, before turning more specifically to the role of relevance in navigation and description strategies.

Keywords: Spatial concepts, top-down processes, route descriptions, relevance theory, perception.

1 Introduction

Recently the locksmith, asked to exchange our garage door lock, wondered if our garage had a number. I had no idea. Getting there, he immediately spotted the small black digit above the door. Should I be embarrassed about my ignorance? I don't think so. My selective attention may be extreme, yet there is a common and systematic principle behind the disparity in visual perception between the locksmith and myself. For the locksmith, the number was relevant as he wished to be able to identify the correct door for purposes of documentation and communication. For me, it had not been relevant in a decade of using the garage for its storage function.

We may not be aware of it, but spatial environments are immensely complex. At a fine level of detail, it is possible in any situation to discern multiple types of perceivable (and often nameable) elements, such as different colors of pavement stones, sizes of streets or houses and their doors and windows, types of trees and flowers or grass, and so on indefinitely. Additionally there are multiple sources of dynamic distractions, such as cars and trains and pedestrians, dogs and birds, planes, ants, and mosquitoes, plus inanimate movable objects such as leaves, dirt and dust; furthermore a multitude of auditory and olfactory impressions. Given the immense complexity of perceptually available information, then, how do humans manage to feel at ease, and find their way around, in spatial environments?

Many theories and models have been put forward to explain human conceptions of space as well as navigation and wayfinding behavior. My claim in this paper is that one conceptual element is implicitly common to all of them, although it is rarely placed in the center of attention in scientific discourse. I will argue that one of the

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 358–377, 2012.

main factors driving human cognition, in space just as in any other domain, is *relevance* – i.e., the "relation to the matter at hand"[1]. Relevance helps explaining why we attend to some spatial elements at some times but to others at other times, why we are not constantly overwhelmed by the sheer overload of information surrounding us in our everyday lives, and also how we manage to produce a concise yet helpful route description to complete strangers on the street.

Relevance is a relative concept. Perceptual information is never relevant *per se*. It may be *salient* in contrast to other objects and types of information available; in fact, relevance and salience can be viewed as two major complementary guiding principles of human cognition (Tenbrink & Röhrbein, in prep.), related to the basic dichotomy of top-down vs. bottom-up cognitive processes (Anderson, 2000). While salience relies on distinctiveness relative to (perceptual or conceptual) context, relevance relies on *purpose*. Without a purpose in a particular activity context there can be no relevance. In order to grasp the influence of relevance on our spatial conceptions, it is therefore of substantial interest to get a clear idea of the purposes according to which we organize our daily lives (Hirtle, Timpf, & Tenbrink, 2011), and which structure our spatial environments (Couclelis, 2009). In this paper I will highlight the role of relevance for one small fraction of possible everyday purposes within the context of spatial navigation, namely situations in which we wish to find our way to another location. I will outline how the context and purpose of a route description affects the ways in which the route is described, drawing on evidence from theoretical work on route navigation as well as recent empirical findings on the specific linguistic choices that speakers take from the generic network of options available to them (cf. Tenbrink & Freksa, 2009). To set the stage, the next section will sketch the recognized role of relevance in cognitive science with respect to communication and perception.

2 The Notion of Relevance in Cognitive Science Theory

Following long-standing traditions in philosophy and psychology, relevance is generally recognized (or at least implicitly understood) in cognitive science as a guiding principle for human cognition (e.g., Mandler, 1985; Fauconnier & Turner, 2002; Ericsson et al., 2006). For example, for skill acquisition it is essential to encode relevant information in such a way that new related situations can be handled appropriately by efficient association processes (Ericsson & Kintsch, 1995); experts are able to retrieve a large body of relevant information whenever needed without overloading the capacities of short term memory. Humans consciously focus on, and prefer, relevant information for a purpose at hand; practice activities that are seen as highly relevant are enjoyable and involve a high amount of concentration (Starkes et al., 1996). Emotions such as anxiety are guided by personal relevance (MacLeod and Rutherford, 1992), just as well as attention processes (Mandler, 1985). In the following, I outline the role of relevance for communication and perception before turning to wayfinding more specifically.

[1] `www.merriam-webster.com/dictionary/relevance`, retrieved May 27, 2012.

2.1 Relevance in Communication

Since Grice (1975) postulated his famous conversational maxims, relevance has been widely acknowledged as a central principle underlying the interpretation of language use. The most developed approach spelling out what relevance means for communication was put forward by Sperber & Wilson (1986). In a nutshell, their *Relevance Theory* is aptly summarized in the following two quotes taken from a handbook chapter (Wilson & Sperber, 2002:607f.):

"The central claim of relevance theory is that the expectations of relevance raised by an utterance are precise enough, and predictable enough, to guide the hearer towards the speaker's meaning."

"Intuitively, an input (a sight, a sound, an utterance, a memory) is relevant to an individual when it connects with background information he has available to yield conclusions that matter to him: say, by answering a question he had in mind, improving his knowledge on a certain topic, settling a doubt, confirming a suspicion, or correcting a mistaken impression. In relevance-theoretic terms, an input is relevant to an individual when its processing in a context of available assumptions yields a positive cognitive effect. A positive cognitive effect is a worthwhile difference to the individual's representation of the world – a true conclusion, for example."

In other words, how an utterance is interpreted depends decisively on the relevance expectations of the listener, which in turn depend on the listener's background knowledge and needs, leading to changes toward an improved (cognitive) state of information. Following this account, all communication is related to relevance in this sense. Language production is constrained by the speaker's assumptions of relevance for the listener; likewise, a meaningful interpretation of utterances is impossible without a discourse context that guides the listener's expectations of relevance. Listeners expect speakers to produce utterances that are somehow relevant, and this expectation constrains utterance interpretation. Relevance expectations may, under certain circumstances, be as trivial as the filling-in of awkward silences in a small talk dinner party situation. Depending on the current discourse context, utterance 1 below can be interpreted in various ways so as to communicate any of the discourse inferences listed as (a) through (d), and many others.

1. This forest is really beautiful.
   (a) How about going for a walk there?
   (b) This part of the painting is particularly beautiful.
   (c) Everything else is detestable around here.
   (d) I would like to change the topic.

Another implication for communicative interpretation is that relevance emerges through contrast. Self-evident information is typically not perceived as relevant, as it does not yield any kind of "positive cognitive effect" in Wilson & Sperber's (2002) terms; it does not convey any communicative contribution (unless it conveys something else, e.g., on an emotional or interpersonal level). On a factual level, utterances become relevant by contrast to that which could have been the case, i.e., *the difference at stake* (Nemo, 1999). Thus, utterance 2 below *"represents both the* *fact that he is alive and the fact that this might (strongly) not have been the case. Consequently, the utterance can represent only a moment when something has happened (an accident, a heart attack, an assassination attempt, etc...)"* (Nemo, 1999).

2. Bill Clinton is alive.

Language has a number of ways to make differences at stake salient to the listener (Talmy, 2007), supporting their expectations of relevance. Language provides a systemic network of options for communicating meaning at any given time (Halliday, 1985; Tenbrink & Freksa, 2009; Bateman et al., 2010). To emphasize the most relevant part, typically the focus is on the new information (Halliday, 1985), setting it apart from the rest of the message. Thus, utterance 3 can establish a relevant contrast to a number of alternatives as outlined in (a) through (d):

3. The ball is under the table.
   (e) The *ball* rather than the *book* is under the table.
   (f) The ball (still) *is* under the table, as it should be.
   (g) The ball is *under,* not *on* the table.
   (h) The ball is under the *table*, not under the *chair*.

Insights such as these have been examined for various purposes, including grammatical accounts (Halliday, 1985), computational approaches to discourse (Grosz & Sidner, 1986), and accounts of the interplay of language and vision (Henderson & Ferreira, 2004; Holsanova, 2008). For instance, perceptual aspects and functional relevance guide the speakers' attention towards particular features of a situation, leading to the preference for certain spatial expressions rather than others (Regier, Carlson, & Corrigan, 2005). Complementarily, the same preposition may be understood to refer to different spatial aspects according to what is perceived as relevant (or perceptually salient) in a situation (Herskovits, 1986). Therefore, the inferences and conclusions that a spatial description lead to will depend on the current purposes and the associated information requirements. The next section will examine the role of relevance for visual perception more directly.

2.2 Relevance in Perception

Like the interpretation of language, human perception is guided by relevance in a given situation context (Fecteau & Munoz, 2006; Henderson et al., 2009). The cognitive processes involved are variously described in terms of selective looking (Neisser & Becklen, 1975), attention focus (Ward et al., 1996; Navalpakkam & Itti, 2005), contextual priming (Torralba, 2003), and top-down mechanisms guiding perception. Top-down mechanisms are seen as opposed to bottom-up mechanisms that involve processing and comprehending the available sensory input, driven by relative salience (Mannan, Ruddock, & Wooding, 1997; Reinagel & Zador, 1999; Vecera & Behrmann, 2001; Henderson, 2003; Rebhan et al., 2008). These two kinds of mechanisms work in tandem to reconcile the multiplicity of perceptually available information with the expectations, needs, and purposes of the perceiver.

Bottom-up mechanisms are stimulus-based; in visual perception, properties of the perceived scene can guide the viewer's eye movements towards locations that are in some sense distinct from their environment. Similar processes have been proposed for auditory perception as well (Kayser et al., 2005).

Top-down mechanisms relate to the scenario context or task purpose, often (to some extent) consciously, as when drivers purposefully check the rearview mirror at the margin of their visual field periodically. However, the principle does not hinge on conscious awareness. Judging spatial relationships subtly guides attention towards particular parts of a depicted scene (Franconeri et al., in press). Recognizing a scene as a whole can invoke generic schema knowledge that supports the identification of individual objects (Henderson, 2003). The mechanism helps to fill in perceptual gaps in meaningful ways, for example when identifying partly occluded or blurred objects within a given visual context (Oliva & Torralba, 2007; Lee & Ban, 2008). In spite of such effects, the main locus of conceived relevance remains in the perceiver. This stands in contrast to linguistic communication as outlined above, where language interpretation (by the perceiver) is guided by the expectation of relevance contributed by the *speaker*. Scenes, in contrast, do not voluntarily communicate relevant content for the benefit of the perceiver.

While the term *relevance* itself is not often invoked in the literature on visual perception, the concept nevertheless corresponds to the basic idea underlying these various approaches and terminologies. The scenario context guides the perceiver towards certain aspects of the situation that are more important than others, given the perceiver's situation-specific relevance expectations. This implies that other aspects are easily "overlooked", escaping the perceiver's attention – like the digit above the garage door. Incidentally, magicians use this principle to trick their audience into believing in just those fantastic appearances that would otherwise not be deemed believable. Scientifically, the phenomenon of *change blindness* fascinates researchers (Rensink, O'Regan, & Clark, 1997); even fairly salient changes can be ignored under certain circumstances, such as the replacement of a person during a route direction dialogue (Simons & Levin, 1998). Furthermore, the principle explains why people asked to describe a perceived scene or event often systematically share particular patterns of description (Lynch, 1960; Zacks et al., 2007; Le Yaouanc, Saux, & Claramunt, 2010): they share attentional focus and knowledge about functional structures, and are therefore likely to describe just those aspects that are perceived as relevant for current task purposes.

Since relevance also guides visual attention when interpreting geographic maps and displays of street networks (Fabrikant & Goldsberry, 2005), it has been proposed as a principle that *should* be used in the design of visual representations, so as to avoid cognitive load on the part of the viewer (Kosslyn, 2006). This corresponds to the notion of *data-ink ratio* proposed by Tufte (2001). Minimizing information that is essentially irrelevant to a task, for instance by simplifying design aspects (or selecting them systematically, see Bertel, Freksa, & Vrachliotis, 2004), can enhance learning processes substantially (Sweller, 2005). Accordingly, Hegarty (2011:464) notes that restricting the visually presented information in a display to the relevant aspects *"relieves the user of the need to maintain a detailed representation of this information* *in working memory, whereas presenting too much information in the display leads to visual clutter or distraction by irrelevant information."* This has interesting implications for the evaluation of spatial descriptions or maps in relation to human spatial cognition. The topic of wayfinding is addressed in the next section.

3 Relevance and Wayfinding

Given the overwhelming multiplicity of perceptual information available in real world environments, how do we single out what we need for finding our way? Again, one decisive factor is *relevance.* While this factor is sometimes mentioned in passing (as in some of the quotes cited below), it generally rather seems to be implicitly taken for granted in the wayfinding literature. As Wilson & Sperber (2002) point out, *"human cognition tends to be geared to the maximisation of relevance"* – it is in our nature to spot any elements that can be supportive in achieving our current goals. Typically, the cognitive processes pertinent for goals such as reaching spatial locations are examined in terms of *wayfinding strategies* (Golledge, 1997; Dabbs et al., 1998; Hochmair & Frank, 2002; Wiener et al., 2004; Hölscher et al., 2009). In order to find an efficient solution to a given navigation task, humans often focus on particular aspects of the environment, such as the angle of the initial direction or the familiarity of a region, rather than computing optimal paths from all of the available spatial information. In the exploration of new environments, humans aim to maximize information gain (Zetzsche, Wolter, & Schill, 2008).

As Allen (1999) spells out in much detail, there are huge individual differences in the ability to employ suitable cognitive strategies for challenging wayfinding tasks. This is due to the fact that different skills are relevant for different tasks, such as visualization (mental imagery), mental rotation, orientation, memory, and attention. Generally, repetition and experience help to activate *"expectations that enhance the salience of certain types of information"* (Allen 1999:69). In other words, people become more sensitive to relevant spatial elements in the environment with practice. Moreover, humans differ in their ability to develop an adequate *cognitive map* from the spatial information accumulated over time. The term *cognitive map*, which goes back to Tolman (1948), refers to a knowledge structure that represents information about spatial relationships in such a way as to allow for inferences and generalizations, particularly for wayfinding purposes (Golledge, 1987). In order for spatial elements to be cognitively represented, it is important that they are *relevant* for behavioral purposes (Sholl, 1996:157): *"a cognitive map or survey representation of a spatial layout codes (…) relations (…) among behaviorally relevant landmarks".* This implies flexibility for diverse behavioral purposes; *"the point of a cognitive map is to represent a great deal of information in a flexible format with an economy of effort"* (Allen, 1999:72). Enhanced flexibility allows wayfinders to use the information stored in their cognitive map for different task purposes (Chown et al., 1995), implying different degrees of relevance of various types of information according to task.

With respect to depictions of spatial environments in the form of maps, Schmid (2008) proposes to adapt the information provided in small displays to the previous knowledge of the user; in this way, irrelevant information can be considerably reduced. Furthermore, maps should ideally represent the *type* of information that is needed by humans for various wayfinding tasks, similar to the idealized features of cognitive maps. Stea, Blaut, and Stephens (1996:355) point out that such a focus on cognitively relevant information may to some extent involve distortion of the actual features and relationships in the real world: *"To communicate, maps maximise relevant information, or presentation of spatial information in the most relevant form. To do this, mapmakers 'distort'."* Indeed, distortions of spatial relationships have long been known to be integral features of cognitive maps (e.g., Tversky, 1992), indicating that metrically accurate representations may not always be optimally supportive or cognitively adequate.

Since graphic representations such as maps are (typically) static, their flexibility for different task purposes is necessarily restricted, forcing the user to attend to potentially irrelevant information (Meilinger & Knauff, 2008). Language is more adaptive in this regard. As Couclelis (1996:135) puts it, route directions *"can be assumed to be based on a model, in the direction-giver's mind, of relevant aspects of the environment. Central to that mental model is a cognitive route-planning task, based on some combination of re-experiencing, remembering, and inferencing. The information necessary and sufficient to answer the query must be extracted from that model"*. Thus, not only do route directions depend on the describer's cognitive representation, but they are also geared towards the expectations and necessities for the precise wayfinding goal at hand. The next section will address the impact of these issues on particular features of route descriptions according to task context.

4 Relevance in the Communication of Routes

There is a large body of literature on features of linguistic route descriptions, examined for various purposes (e.g., Klein, 1979; Taylor & Tversky, 1992; Couclelis, 1996; Denis, 1997; Lovelace, Hegarty, & Montello, 1999; Tversky & Lee, 1999; Allen, 2000; Gryl, Moulin, & Kettani, 2002; Klippel, 2003; Daniel & Denis, 2004; Tenbrink & Winter, 2009; Hölscher, Tenbrink, & Wiener, 2011). The bulk of this work centers on a single standard scenario, namely that of the stranger on the street asking how to get to another location in the same city. Although deviations from this scenario are sometimes addressed and easily conceivable, e.g., asking about two alternative paths, more than one goal, long-distance or within-building navigation, or asking a friend rather than a stranger, the standard scenario appears to be the unquestioned default interpretation for the notion of route directions. From this body of work, it is possible to derive a number of generalized conclusions about the nature of such descriptions. Route descriptions include a number of *standard conceptual elements* that appear to be generally relevant for the purpose of route finding, but may *vary* in some ways according to features of the situation that lead to different degrees of relevance. They are linguistically *underspecified* and representationally incomplete, due to the perceived irrelevance of some aspects. Finally, they reflect the perceived relevance of certain *non-spatial factors*. The following subsections will address these features in more detail.

4.1 Standard Description Elements

Central elements of route descriptions that are frequently mentioned in the literature (cited above) include the following (independent of the scenario):

- starting point and destination, including deictic elements like *from here*, street names, buildings, and other location references, such as *to the central station*;
- intermediate decision points, typically references to intersections; these may have a particular (nameable) structure, such as *at the T-crossing*;
- route segments, such as paths, streets, and hallways; e.g., *along Main Street.* Paths are segmented via decision points and can be chunked, combining several decision points;
- actions and movement directions, e.g., *walk, head, go* for straight paths, and *follow* for curved paths;
- reorientations (at decision points) with schematized directions and angles, e.g., *turn right*. Typically directions are expressed using projective terms such as *left* and *right*; depending on culture, compass directions such as *north* and *south* may also come into play;
- landmarks, located either at decision points such as *turn left at the church*, or along the route for confirmation, as in *you will pass a grocery store*;
- regions and areas, such as *downtown* or *the park*;
- distances, typically qualitative rather than quantitative and spatial as well as temporal, e.g., *after a little while, a short road, about 200 meters.*

Similar basic elements can be found in sketch maps drawn by humans asked to depict a route (Tversky & Lee, 1999), and they also resemble the information provided in aspect maps (Barkowsky & Freksa, 1997). It can be concluded that these elements constitute the basic conceptual building blocks used by humans to formulate route descriptions. These are the elements that humans find relevant for the particular cognitive task of wayfinding. In fact, we are used to these basic elements to such an extent that we hardly notice how much is actually left out, considering the perceptual and informational richness of real spatial environments, or how else a route description could be formulated in theory. Thus, the following answer to a question like "How do I get to the train station?" would not be considered acceptable under normal circumstances:

4. You first rotate your body, then walk 23 steps, then move for five minutes in the direction in which the ducks usually fly. You listen to the sound of the trains, wait for the correct moment and then go 42 steps on the staircase downwards.

Information typically left out includes precise quantitative measures of angles and distances, specific features of buildings and regions, dynamic elements such as pedestrians, traffic, and animals or moving objects, weather information, quality and material of streets, paths, and neighboring fences, and so on. In part, such information is not feasible for route directions since it is subject to frequent change. As Talmy (2000:184) pointed out, in spatial descriptions *"the Figure is a moving or conceptually movable entity whose site, path, or orientation is conceived as a variable the particular value of which is the relevant issue. The Ground is a reference entity, one that has a stationary setting relative to a reference frame, with respect to which the Figure's site, path, or orientation is characterized."* Route directions are no exception. The wayfinder is a movable entity, whose path is described by reference to static entities. Dynamic entities are rarely relevant for route descriptions because they are not reliable over time. However, exceptions are conceivable: if a route giver perceives a car that is about to turn into just the street that the wayfinder is looking for, this may be used as a reference.

Furthermore, route elements can only be mentioned if they are available to the describer. Precise quantitative measures typically do not enter route descriptions because they are not perceptually available to humans. Changes in the distributional pattern of spatial elements can sometimes be traced back to differences in spatial expertise (Tenbrink, Bergmann, & Konieczny, 2011). Some types of information, though available in theory, rarely seem to be sufficiently relevant for any wayfinding task. For instance, the number of windows in houses adjacent to the route may be conceptually simple to assess, just as the material of street surfaces or types of trees along the street. Such information only comes into play exceptionally, for instance if these features are the only ones (or particularly *salient* ones) available to discriminate spatial locations from others, or if the route description is formulated with a particular task purpose in mind that transcends the spatial need of finding the way to a target.

A comparison with automatically generated route directions, such as those provided by web services, revealed that the basic cognitive route elements, which seem so self-evident to humans, are only to a limited extent reflected by such systems (Tenbrink & Winter, 2009). System-generated descriptions necessarily rely on the information available in databases, which primarily provide quantitative data such as exact distances, but hardly any references to landmarks or regions (Hansen, Richter, & Klippel, 2006). Accordingly, it remains a major challenge for route information systems to provide cognitively adequate route descriptions (Cuayáhuitl et al., 2010; Mast, Jian, & Zhekova, 2012). Complementarily, humans expect that a system as the receiver of a route description has specific requirements, which usually leads to simpler and less varied spatial instructions in comparison to human addressees (Tenbrink et al., 2010). In particular, while humans addressing other humans variously refer to goal and landmark locations, turn directions, and segments of the route, humans addressing a system stick to an extremely simple route instruction scheme that is based almost entirely on movement and turn directions such as *right* and *straight on.* Furthermore, while human-human dialogues are characterized by negotiation and switches between route and survey perspectives, humans addressing a system use the route perspective much more consistently.

4.2 Patterns of Variability

Although linguistic route directions formulated by humans for human wayfinders systematically rely on the above-mentioned repertory of basic route elements, these route descriptions are obviously not invariable. Different task requirements lead to different expectations on the part of the wayfinder and so variability in route descriptions can often be traced back to relevance for a given task. Most crucially, relevance affects the level of granularity with which particular elements are described. Addressing information needs in relation to the spatial situation as well as order effects, Tenbrink & Winter (2009) found that route directions changed substantially in length and degree of detail. For instance, in their data, the same situation was described in the two ways shown in examples 5 and 6, reflecting striking differences in granularity.

5. Dem Straßenverlauf ca. 200m folgen, auf der linken Seite befinden sich eine Turnhalle und die Haupt/Realschule, gegenüber ist das Gymnasium Ganderkesee mit einem Parkplatz, den Parkplatz bis zum Ende hochgehen.
   [Follow the road for approx. 200m, on the left hand side there is a gym and the junior high school, and on the opposite side there is the Gymnasium Ganderkesee *(a secondary school)* with a parking place, walk up the parking place until its end.]
6. Geradeaus, am Ende rechts.
   [Straight on, at the end to the right.]

Systematic granularity differences were also identified by Hölscher et al. (2011), who compared route descriptions formulated for an imagined stranger in town with plans for one's own future route. Although the relative distribution of route elements (e.g., amount of landmarks mentioned relative to references to action) remained stable, strangers received additional details that were deemed relevant to find their way in the unknown environment. At the same time, the route choices also differed; strangers received simpler routes containing fewer turn changes, indicating that route givers took memorizability into account as a relevant factor.

A reanalysis of the data collected in the study by Hölscher et al. (2011) furthermore revealed interesting differences with respect to the verbalization of the same intersections (Tenbrink, Hölscher, & Wiener, 2008), depending on their function and relevance within the route context (cf. Klippel, 2003). If an intersection serves as a decision point with an associated turning action, it is more important to provide details than if it is merely a structural feature along a straight path, because a wayfinder who needs to change direction should identify the intersection unambiguously. Example 7 illustrates a description with a direction change, and example 8 shows a description without a direction change at that same intersection.

7. Wenn Sie von hier aus immer gerade laufen, kommen Sie irgendwann zur Kaiser-Joseph Straße. Eine Straße voll mit Geschäften, wo auch Straßenbahnen fahren. Biegen Sie jetzt an der Straße rechts ab, und laufen Sie weiter.
   [If you walk always straight on from here, you will at some time get to the Kaiser-Joseph street. A street full of shops, where there are also trams. Now turn right at that street, and keep walking.]

8. Folge der Straße. Passiere dabei 3 Fußgängerampeln, davon eine mit Straßenbahnschienen.
   [Follow the street. Pass three pedestrian lights along the way, one of which has tram tracks.]

These examples illustrate a number of patterns. In our data, references to intersections with decision points (such as example 7) contained direct mention of the intersection reliably more often; in example 8, the intersection is only implied in terms of one of the three "pedestrian lights" that indicate the crossing of streets. They also more often contained references to buildings such as shops and generally a higher number of spatial details. Furthermore, the associated verbs differed systematically, as seen in the examples where the verbs *abbiegen [turn], kommen [come],* and *weiter laufen [keep walking]* are opposed to the verbs *folgen [follow]* and *passieren [pass].* Finally, the mention of new directions such as *right* systematically indicated change, whereas references to intersections without direction change often did not involve any mention of directions at all. Along these lines, the functional relevance of a particular intersection within a route description leads to systematic linguistic differences.

The relevance of elements of route descriptions is furthermore affected by the perspective available to the wayfinder (Hund, Haney, & Seanor, 2008). Information about left and right directions as well as landmarks were found to be more relevant for people driving through a town, while compass directions were more relevant for people looking at a map. Differences in the features of route descriptions such as these systematically affect their interpretation by the wayfinder. For example, survey and route descriptions vary with respect to the cognitive map that can be derived from a linguistic representation (Taylor & Tversky, 1992). The following subsection highlights further implications and challenges involved in interpreting route descriptions.

4.3 Gaps in Spatial Meanings

While route descriptions exhibit systematic patterns with respect to granularity, even the most detailed linguistic description will never be a complete representation of spatial reality. On the one hand, as we have seen, reality is too complex and contains changing and irrelevant details. On the other hand, language as such is never fully complete with respect to the meanings expressed. Carston (2002:19f.) captured this idea in terms of her *underdeterminacy hypothesis*:

"Linguistic meaning underdetermines what is said. (...) What is meant by this is that the linguistic semantics of the utterance, that is, the meaning encoded in the linguistic expressions used, the relatively stable meanings in a linguistic system, meanings which are widely shared across a community of users of the system, underdetermines the proposition expressed (what is said). The hearer has to undertake processes of pragmatic inference in order to work out not only what the speaker is implicating but also what proposition she is directly expressing."

In other words, there are two levels of inferences that the receiver of a linguistic message needs to address. On the one hand, an utterance may convey pragmatic meanings that can only be inferred from context, as in cases of irony, sarcasm, humor, indirect requests such as *"It is cold in here"* intended to mean: *"Shut the window"*, and other types of conversational implicatures (Grice, 1975). On the other hand, more pertinent to route descriptions, there is the general fact of linguistic underspecification that, as Carston (2002) claims, necessitates context-based inference processes already when interpreting the directly expressed meaning. To take a simple example, adjectives like *small* are inherently relational, indicating a relation to something else that needs to be inferred from the context (Freksa, 1980). A small elephant is much larger than even a large mouse, because elephants (and mice) have certain standard sizes in comparison to which a particular exemplar might be small (or large). This is just one of the many ways in which language use invokes nonlinguistic knowledge, tightly interwoven with context. Carston (2002) claims that linguistic underdeterminacy is an essential feature of language, i.e., there are no "eternal sentences" at all; context always plays a decisive role for interpreting language in use.

Again, the conceptual element that resolves these issues is relevance. Following Sperber & Wilson (1986, see subsection 2.1 above), Carston (2002:83) describes how relevance makes communication possible despite the ubiquitous underdeterminacy of language in use: *"The pragmatic inferential capacity, whose specific domain is utterances and other communicative acts, employs a particular interpretive strategy, distinct from that of the more general capacity of mental state attribution, and warranted by the presumption of optimal relevance that is automatically conveyed by such stimuli."* According to a recent argument, ambiguity and underspecification actually *enhance* communication rather than hampering it, since efficient linguistic units can be re-used and interpreted effortlessly according to context (Piantadosi, Tily, & Gibson, 2012). It seems that the impact of relevance on our interpretive capacities cannot be overestimated in this regard.

Route descriptions are no exception; underspecification is an essential feature of any spatial message. Consider the following example:

9. Geradeaus durch die Colombistraße, bis zum Ende. Straße überqueren. Dann links lang laufen, geradeaus bis zur großen Straße. Dort über die Ampel.
   [Straight on through the Colombi street, until its end. Cross the street. Then walk along left, straight on until the large street. There across the traffic light.]

Utterance 9, chosen randomly from our corpus (Hölscher et al., 2011), is completely typical in its mention of spatial elements and representation format, and thus may well be perceived as unambiguous and sufficiently specified. However, at a closer look, at least the following questions may sensibly be asked:

- *geradeaus [straight on]*: in which direction is the wayfinder supposed to walk? Is it the direction they are currently heading – or the route giver? Is it the direction in which they were walking before? At least, there should be an accompanying gesture (or meaningful glance) that clarifies this. Bauer et al. (2009) identified this kind of underspecification as a serious issue that needed to be specifically operationalized in human-robot interaction.
- Where exactly is the *Colombi street*? Pragmatic inference may suggest that this is the very road where the speakers are located – but this is not part of the utterance.

- *links lang laufen [walk along left]*: Does this mean *turn left*, or rather *walk along on the left hand side of the street?* If a turning action is intended, does this mean walking into a different street? In which angle; would this mean 90° exactly? Klippel et al. (2004) addressed the latter issue by categorizing the underlying route direction concepts based on empirical evidence.
- Is there only one *large street*? How large is it? Since *large* is a relational adjective, a substantial amount of world knowledge is needed to interpret this reference.
- Which *traffic light* is being referred to? Is there only one traffic light – in which area? In which direction should the wayfinder cross this particular location?

Some of these issues may seem trivial, due to the fact that the route description adheres to normal conventions that lead to standard inferences, such as interpreting *large* as *larger than the previously walked street* or *larger than average*, *left* as a particular (extended) area on the lefthand region of the wayfinder, and so on. Moreover, the describer obviously expects the wayfinder to recognize the relevant spatial locations upon encountering them in the real world, which warrants the use of the definite article in references such as *the large street* and *the traffic light.* Along these lines, the underspecification can be resolved by the context of the route description. The description will be interpreted as conveying precisely those pieces of information that are relevant to make the correct navigation decisions at these locations. The extent to which this is sufficient for wayfinding will be the main factor determining the quality of the route description as perceived by the wayfinder.

4.4 Relevance of Non-spatial Aspects

Route descriptions relate to context not only in a spatial sense, but are also situated in a discourse context that can systematically affect which aspects of the spatial environment are perceived as relevant for a communicative situation (Porzel, 2010). For example, how a spatial question is answered by people on the street is affected by factors such as travel modality (car, bike, walking), weather, clock time (for instance when asking about the cinema or a castle that may be open for visitors only at particular times), and more; such factors may not necessarily be consciously accessible to speakers who answer the given questions intuitively. Similarly, in the above-mentioned study, Hölscher et al. (2011) found that route choices were systematically affected by the situation. For the benefit of strangers, people aimed to describe routes that were simple and easy to find, and that contained few direction changes but many salient landmarks that helped to find the way. For their own future routes, they rather chose routes that were attractive and not too busy, along with the aim of finding a fast, short, and direct route to the given goal. These aspects were mentioned in a post-hoc questionnaire by the participants as explicit route choice criteria, and confirmed by the actual route choices. Although not all of these factors are spatial, they nevertheless affected the spatial descriptions, without necessarily conveying this kind of information explicitly.

Generally, spatial descriptions can be seen as part of a more general activity (Hirtle, Timpf, & Tenbrink, 2011). A person asking about the way to a hospital may do so as part of a leisurely stroll through nature, or with the plan of visiting a friend who has gone through a routine examination, or in order to bring an injured person there as quickly as possible. Route givers will typically grasp this higher-order activity intuitively to some extent, and tailor their descriptions in a way appropriate to the demands of this activity. This affects the level of granularity of description just as well as the type and amount of additional information offered along with the spatial description, such as *It is a very pleasant walk up there*, or *If you stop directly in front of the emergency entrance people will come out to assist you.*

The extent to which spatial descriptions can be tailored to the underlying discourse purposes naturally depends on the information available to the speaker. In an intriguing large-scale study about the notion of *place*, Winter et al. (2011) simply asked people in a web-based study to "Tell us where you are". Since no further contextual information was provided, responses varied considerably, for instance with respect to deixis (*I'm at home* vs. a postal address), specificity (generic name vs. precise location); affective evaluation (as in *The view towards the beach has a colorful sunset most of the time*) and other descriptive information (*My house is at the top of a hill*). Some people provided a location description or a procedural route description that enabled actually finding the location (e.g., *I am in Hawthorn suburb near to Auburn train station, and at level two of an apartment in Queens Ave*); some actually appeared to invite the addressee to join them (as in *Take Arthur's Seat Road to get up there, it's crazy tight but lots of fun*). Thus, the responses reflected the different ways in which the study participants perceived the request; given no context, they were free to decide how to answer the question, opening up the space of possibilities in this regard. Together, their answers convey the various aspects of *place* that may become relevant to speakers in diverse discourse contexts.

5 Conclusion

Relevance affects how we perceive the world, and how we communicate about it. This paper has addressed this idea in general terms based on a review of findings on cognition and communication, and furthermore spelled out in some detail how relevance affects wayfinding and communication about routes. Some general patterns emerge. Relevance affects *what* is perceived, conceptualized, and communicated, as well as *how* this happens. With respect to the former, relevance leads to the selection of those situational aspects that are felt to be supportive for a given goal, guiding the focus of perceptual attention just as well as the choice of routes and the reference to environmental features. With respect to the latter, relevance affects the level of granularity and the perspective taken on a scene, both perceptually and linguistically. Details of a situation that are more fine-grained than a certain basic level of conceptualization (Rosch et al., 1976; Morris & Murphy, 1990) are attended to and communicated only if they are perceived as relevant. These processes are at work alongside with processes guided by salience; outstanding features of a situation tend to draw attention even without direct relevance for a goal, and may enter communication simply because of their conspicuity.

Humans are not necessarily aware of these processes. In fact, it is conceivable that conceptualizations and communicative choices are conceived of as objective and even complete, for instance in a route description that contains reference to every spatial segment and decision point along the route. Such a conception is a direct consequence of the immediate and unconscious integration of relevance. It is only in unclear or ambiguous situations that we become aware of the complexity of aspects that are actually available to be processed and communicated. Complementarily, the variability of aspects that may be perceived as relevant becomes evident through the analysis of a wide range of data, such as route descriptions formulated without a clear context. Such data allow for the (abductive) inference of the speakers' underlying communicative goals. If a place description contains route instructions and a positive evaluation of the place to be described, this can be read as an invitation to actually get there. Route descriptions pointing to the quickest path aim to support efficient navigation. References to interesting landmarks and beautiful scenery may be directed at tourists looking for inspiration and recreation. Such inferences are possible precisely because this kind of information would otherwise be irrelevant.

As a consequence, any analysis of conceptual representations, such as linguistic data representing route navigation concepts, must necessarily account for the discourse context as perceived by the speakers. It is important to note who is speaking to whom, for which reason and with which discourse goal in mind, whether efficient navigation is at stake or rather personal preference, small talk, spatial knowledge, functional goals, affect and evaluation, and so on. Some aspects may never be accessible to the researcher, such as a speaker's underlying (possibly unconscious) goals in conveying spatial information (e.g., to provide answers of the kind that the researcher wishes to collect). Raising awareness to those aspects that are transparently relevant in a situation may support clarity in this regard.

Apart from conscientiously controlling contextual factors that potentially affect relevance as perceived by experimental participants, future work should vary the relevance of situational factors systematically so as to establish more precisely, and more generally, the effects of relevance on spatial concepts and descriptions. It is time to leave standard route direction scenarios so as to establish more broadly which spatial factors are affected by relevance in which way, and which cognitive subprocesses of wayfinding may vary according to relevance in relation to the context in which a navigation task takes place (cf. Wiener, Büchner, & Hölscher, 2009). It stands to reason that these conclusions would be valid beyond the restricted topic of navigation and route communication that has been the focus of this paper. Viewed this way, it seems that research has only just begun to capture the flexibility of human cognition in spatial environments.

Acknowledgements. Many thanks go to Florian Röhrbein for collaboration leading to central insights on issues of relevance and salience in visual perception, Jan Wiener and Christoph Hölscher for collaboration on wayfinding strategies, Robert Porzel for inspiration concerning context effects and Sabine Timpf for discussion of activity types, Stephan Winter for insight into the PLACE data, the project members of

I5-[DiaSpace] and I6-[NavTalk] for collaboration and feedback, Christian Freksa for many inspiring discussions on language and knowledge representation, and three anonymous reviewers for their insightful comments. Funding by the DFG is appreciated for projects I5-[DiaSpace] and I6-[NavTalk], SFB/TR 8 Spatial Cognition.

References

1. Allen, G.L.: Spatial abilities, cognitive maps, and wayfinding: Bases for individual differences in spatial cognition and behavior. In: Golledge, R. (ed.) Wayfinding Behavior: Cognitive Maps and other Spatial Processes, pp. 46–80. Johns Hopkins, Baltimore (1999)
2. Allen, G.L.: Principles and practices for communicating route knowledge. Applied Cognitive Psychology 14, 333–359 (2000)
3. Anderson, J.R.: Cognitive psychology and its implications, 5th edn. Worth Publishers, New York (2000)
4. Barkowsky, T., Freksa, C.: Cognitive Requirements on Making and Interpreting Maps. In: Hirtle, S., Frank, A. (eds.) Spatial Information Theory: A Theoretical Basis for GIS, pp. 347–361. Springer, Berlin (1997)
5. Bateman, J., Hois, J., Ross, R.J., Tenbrink, T.: A Linguistic Ontology of Space for Natural Language Processing. Artificial Intelligence 174, 1027–1071 (2010)
6. Bauer, A., Gonsior, B., Wollherr, D., Buss, M.: Heuristic Rules for Human-Robot Interaction Based on Principles from Linguistics - Asking for Directions. In: Proceedings of AISB, Edinburgh, April 6-9 (2009)
7. Bertel, S., Freksa, C., Vrachliotis, G.: Aspectualize and conquer in architectural design. In: Gero, J.S., Tversky, B., Knight, T. (eds.) Visual and Spatial Reasoning in Design III. Key Centre of Design Computing and Cognition, pp. 255–279. Sydney University (2004)
8. Carston, R.: Thoughts and utterances: The pragmatics of explicit communication. Blackwell, Oxford (2002)
9. Chown, E., Kaplan, S., Kortenkamp, D.: Prototypes, Location, and Associative Networks (PLAN): Towards a Unified Theory of Cognitive Mapping. Cognitive Science 19, l–51 (1995)
10. Couclelis, H.: Verbal directions for way-finding: space, cognition, and language. In: Portugali, J. (ed.) The Construction of Cognitive Maps, pp. 133–153. Kluwer, Dordrecht (1996)
11. Couclelis, H.: The Abduction of Geographic Information Science: Transporting Spatial Reasoning to the Realm of Purpose and Design. In: Hornsby, K.S., Claramunt, C., Denis, M., Ligozat, G. (eds.) COSIT 2009. LNCS, vol. 5756, pp. 342–356. Springer, Heidelberg (2009)
12. Cuayáhuitl, H., Dethlefs, N., Richter, K.-F., Tenbrink, T., Bateman, J.: A dialogue system for indoor wayfinding using text-based natural language. International Journal of Computational Linguistics and Applications 1(1-2), 285–304 (2010)
13. Dabbs Jr., J.M., Chang, E.-L., Strong, R.A., Milun, R.: Spatial Ability, Navigation Strategy, and Geographic Knowledge Among Men and Women. Evolution and Human Behavior 19(2), 89–98 (1998)
14. Daniel, M.-P., Denis, M.: The production of route directions: Investigating conditions that favor conciseness in spatial discourse. Applied Cognitive Psychology 18(1), 57–75 (2004)
15. Denis, M.: The description of routes: A cognitive approach to the production of spatial discourse. Cahiers de Psychologie Cognitive 16(4), 409–458 (1997)

16. Ericsson, K.A., Charness, N., Feltovich, P.J., Hoffman, R.R. (eds.): The Cambridge Handbook of Expertise and Expert Performance. Cambridge University Press, Cambridge (2006)
17. Ericsson, K.A., Kintsch, W.: Longterm working memory. Psychological Review 102, 211–245 (1995)
18. Fabrikant, S.I., Goldsberry, K.: Thematic relevance and perceptual salience of dynamic geovisualization displays. In: 22nd International Cartographic Conference, A Coruna, Spain, July 9-16 (2005)
19. Fauconnier, G., Turner, M.: The Way We Think: Conceptual Blending and the Mind's Hidden Complexities. Basic Books, New York (2002)
20. Fecteau, J.H., Munoz, D.P.: Salience, relevance, and firing: a priority map for target selection. Trends in Cognitive Sciences 10(8), 382–390 (2006)
21. Franconeri, S.L., Scimeca, J.M., Roth, J.C., Helseth, S.A., Kahn, L.: Flexible visual processing of spatial relationships. Cognition (in press)
22. Freksa, C.: Communication about visual patterns by means of fuzzy characterizations. In: XXIInd International Congress of Psychology, Leipzig (1980)
23. Golledge, R.G.: Environmental cognition. In: Stokols, D., Altman, I. (eds.) Handbook of Environmental Psychology, vol. I. Wiley, New York (1987)
24. Golledge, R.G.: Defining the criteria used in path selection. In: Ettema, D.F., Timmermans, H.J.P. (eds.) Activity-Based Approaches to Travel Analysis, pp. 151–169. Elsevier, New York (1997)
25. Grice, H.P.: Logic and conversation. In: Cole, P., Morgan, J. (eds.) Syntax and Semantics, vol. 3, pp. 41–58. Academic Press, New York (1975)
26. Gryl, A., Moulin, B., Kettani, D.: A conceptual model for representing verbal expressions used in route descriptions. In: Coventry, K.R., Olivier, P. (eds.) Spatial Language: Cognitive and Computational Perspectives, pp. 19–42. Kluwer, Dordrecht (2002)
27. Halliday, M.A.K.: An Introduction to Functional Grammar. Edward Arnold, London (1985)
28. Hansen, S., Richter, K.-F., Klippel, A.: Landmarks in OpenLS — A Data Structure for Cognitive Ergonomic Route Directions. In: Raubal, M., Miller, H.J., Frank, A.U., Goodchild, M.F. (eds.) GIScience 2006. LNCS, vol. 4197, pp. 128–144. Springer, Heidelberg (2006)
29. Hegarty, M.: The Cognitive Science of Visual-Spatial Displays: Implications for Design. Topics in Cognitive Science 3, 446–474 (2011)
30. Henderson, J.M.: Human gaze control during real-world scene perception. Trends in Cognitive Sciences 7(11), 498–504 (2003)
31. Henderson, J.M., Ferreira, F. (eds.): The interface of language, vision, and action: Eye movements and the visual world. Psychology Press, New York (2004)
32. Henderson, J.M., Malcolm, G.L., Schandl, C.: Searching in the dark: Cognitive relevance drives attention in real-world scenes. Psychonomic Bulletin & Review 16(5), 850–856 (2009)
33. Herskovits, A.: Language and Spatial Cognition. Cambridge University Press, Cambridge (1986)
34. Hirtle, S.C., Timpf, S., Tenbrink, T.: The Effect of Activity on Relevance and Granularity for Navigation. In: Egenhofer, M., Giudice, N., Moratz, R., Worboys, M. (eds.) COSIT 2011. LNCS, vol. 6899, pp. 73–89. Springer, Heidelberg (2011)
35. Hochmair, H.H., Frank, A.U.: Influence of estimation errors on wayfinding-decisions in unknown street networks – analyzing the least-angle strategy. Spatial Cognition and Computation 2(4), 283–313 (2002)

36. Holsanova, J.: Discourse, vision, and cognition. Benjamins, Amsterdam (2008)
37. Hölscher, C., Büchner, S.J., Meilinger, T., Strube, G.: Adaptivity of wayfinding strategies in a multi-building ensemble: The effects of spatial structure, task requirements, and metric information. Journal of Environmental Psychology 29(2), 208–219 (2009)
38. Hölscher, C., Tenbrink, T., Wiener, J.: Would you follow your own route description? Cognition 121, 228–247 (2011)
39. Hund, A.M., Haney, K.H., Seanor, B.D.: The role of recipient perspective in giving and following wayfinding directions. Applied Cognitive Psychology 22(7), 896–916 (2008)
40. Kayser, C., Petkov, C.I., Lippert, M., Logothetis, N.K.: Mechanisms for allocating auditory attention: An auditory saliency map. Current Biology 15(21), 1943–1947 (2005)
41. Klein, W.: Wegauskünfte [Route descriptions]. Zeitschrift für Literaturwissenschaft und Linguistik (LiLi) 33, 9–57 (1979)
42. Klippel, A.: Wayfinding choremes - Conceptualizing wayfinding and route direction elements. Dissertation. Monograph Series of the Transregional Collaborative Research Center SFB/TR 8 Spatial Cognition, vol. I (2003) ISBN 3-88722-590-2
43. Klippel, A., Dewey, C., Knauff, M., Richter, K.-F., Montello, D.R., Freksa, C., Loeliger, E.-A.: Direction Concepts in Wayfinding Assistance Systems. In: Baus, J., Kray, C., Porzel, R. (eds.) Artificial Intelligence in Mobile Systems (AIMS 2004), pp. 1–8 (2004)
44. Kosslyn, S.: Graph design for the eye and mind. Oxford University Press, Oxford (2006)
45. Lee, M., Ban, S.-W.: Incremental Knowledge Representation Based on Visual Selective Attention. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part II. LNCS, vol. 4985, pp. 940–949. Springer, Heidelberg (2008)
46. Le Yaouanc, J.-M., Saux, E., Claramunt, C.: A semantic and language-based representation of an environmental scene. GeoInformatica 14(3), 333–352 (2010)
47. Lovelace, K.L., Hegarty, M., Montello, D.R.: Elements of Good Route Directions in Familiar and Unfamiliar Environments. In: Freksa, C., Mark, D.M. (eds.) COSIT 1999. LNCS, vol. 1661, pp. 65–82. Springer, Heidelberg (1999)
48. Lynch, K.: The Image of the City. The Technology Press and the Harvard University Press, Cambridge, MA (1960)
49. MacLeod, C., Rutherford, E.M.: Anxiety and the selective processing of emotional information: Mediating roles of awareness, trait and state variables, and personal relevance of stimuli. Behaviour Research and Therapy 30(5), 479–491 (1992)
50. Mandler, G.: Cognitive Psychology: An Essay in Cognitive Science. Routledge, London (1985)
51. Mannan, S., Ruddock, K., Wooding, D.: Fixation patterns made during brief examination of two-dimensional images. Perception 26(8), 1059–1072 (1997)
52. Mast, V., Jian, C., Zhekova, D.: Elaborate Descriptive Information in Indoor Route Instructions. In: CogSci 2012: 34th Annual Conference of the Cognitive Science Society, Sapporo, Japan, August 1-4 (2012)
53. Meilinger, T., Knauff, M.: Ask for your way or use a map: a field experiment on spatial orientation and wayfinding in an urban environment. Spatial Science 53(2), 13–24 (2008)
54. Morris, M.W., Murphy, G.L.: Converging operations on a basic level in event taxonomies. Memory & Cognition 18, 407–418 (1990)
55. Navalpakkam, V., Itti, L.: Modeling the influence of task on attention. Vision Research 45(2), 205–231 (2005)
56. Neisser, U., Becklen, R.: Selective looking: Attending to visually specified events. Cognitive Psychology 7, 480–494 (1975)

57. Nemo, F.: The pragmatics of signs, the semantics of relevance, and the semantic/pragmatic interface. In: Turner, K. (ed.) The Semantics-Pragmatics Interface from Different Points of View. CRiSPI Series, pp. 343–417. Elsevier Science, Amsterdam (1999)
58. Oliva, A., Torralba, A.: The role of context in object recognition. Trends in Cognitive Science 11(12), 520–527 (2007)
59. Piantadosi, S.T., Tily, H., Gibson, E.: The communicative function of ambiguity in language. Cognition 122, 280–291 (2012)
60. Porzel, R.: Contextual Computing: Models and Applications. Springer, Heidelberg (2010)
61. Rebhan, S., Röhrbein, F., Eggert, J.P., Körner, E.: Attention Modulation Using Short- and Long-Term Knowledge. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 151–160. Springer, Heidelberg (2008)
62. Regier, T., Carlson, L., Corrigan, B.: Attention in spatial language: Bridging geometry and function. In: Carlson, L., van der Zee, E. (eds.) Functional Features in Language and Space: Insights from Perception, Categorization and Development, pp. 191–204. Oxford University Press, Oxford (2005)
63. Reinagel, P., Zador, A.M.: Natural scene statistics at the centre of gaze. Network 10, 341–350 (1999)
64. Rensink, R.A., O'Regan, J.K., Clark, J.J.: To see or not to see: The need for attention to perceive changes in scenes. Psychological Science 8, 368–373 (1997)
65. Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., Boyes-Braem, P.: Basic objects in natural categories. Cognitive Psychology 8, 382–439 (1976)
66. Schmid, F.: Knowledge based wayfinding maps for small display cartography. Journal of Location Based Services 2(1), 57–83 (2008)
67. Sholl, J.M.: From visual information to cognitive maps. In: Portugali, J. (ed.) The Construction of Cognitive Maps, pp. 157–186. Kluwer, Dordrecht (1996)
68. Simons, D.J., Levin, D.T.: Failure to detect changes to people during a real-world interaction. Psychonomic Bulletin and Review 5(4), 644–649 (1998)
69. Sperber, D., Wilson, D.: Relevance: Communication and Cognition. Harvard University Press, Cambridge (1986)
70. Starkes, J.L., Deakin, J.M., Allard, F., Hodges, N., Hayes, A.: Deliberate practice in sports: What is it anyway? In: Ericsson, K.A. (ed.) The Road to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sports, and Games, pp. 81–106. Erlbaum, Mahwah (1996)
71. Stea, D., Blaut, J.M., Stephens, J.: Mapping as a cultural universal. In: Portugali, J. (ed.) The Construction of Cognitive Maps, pp. 345–360. Kluwer, Dordrecht (1996)
72. Sweller, J.: Implications of cognitive load theory for multimedia learning. In: Mayer, R.E. (ed.) The Cambridge Handbook of Multimedia Learning, pp. 19–30. Cambridge University Press, New York (2005)
73. Talmy, L.: Attention phenomena. In: Geeraerts, D., Cuyckens, H. (eds.) Handbook of Cognitive Linguistics, ch. 11, pp. 264–293. Oxford University Press (2007)
74. Taylor, H.A., Tversky, B.: Spatial mental models derived from survey and route descriptions. Journal of Memory and Language 31, 261–292 (1992)
75. Tenbrink, T., Bergmann, E., Konieczny, L.: Wayfinding and description strategies in an unfamiliar complex building. In: Carlson, L., Hölscher, C., Shipley, T.F. (eds.) Proceedings of the 33rd Annual Conference of the Cognitive Science Society, pp. 1262–1267. Cognitive Science Society, Austin (2011)
76. Tenbrink, T., Freksa, C.: Contrast sets in spatial and temporal language. In: ICSC International Conference on Spatial Cognition 2009, Rome; Cognitive Processing 10 (suppl. 2), S322–S324 (2009)

77. Tenbrink, T., Hölscher, C., Wiener, J.: Route instructions at choice points. In: Spatial Language in Context: Computational and Theoretical Approaches to Situation Specific Meaning. Workshop at the International Conference Spatial Cognition 2008, Freiburg, Germany, September 15-19 (2008)
78. Tenbrink, T., Röhrbein, F.: Relevance and Salience in Cognitive Processes: From Perception to Language (in prep.) (manuscript)
79. Tenbrink, T., Ross, R.J., Thomas, K.E., Dethlefs, N., Andonova, E.: Route instructions in map-based human-human and human-computer dialogue: a comparative analysis. Journal of Visual Languages and Computing 21(5), 292–309 (2010)
80. Tenbrink, T., Winter, S.: Variable Granularity in Route Directions. Spatial Cognition and Computation 9, 64–93 (2009)
81. Tolman, E.C.: Cognitive maps in rats and men. Psychological Review 55, 189–208 (1948)
82. Torralba, A.: Contextual priming for object detection. International Journal of Computer Vision 53(2), 169–191 (2003)
83. Tufte, E.: The Visual Display of Quantitative Information. Graphics Press (2001)
84. Tversky, B.: Distortions in cognitive maps. Geoforum 23(2), 131–138 (1992)
85. Tversky, B., Lee, P.U.: Pictorial and Verbal Tools for Conveying Routes. In: Freksa, C., Mark, D.M. (eds.) COSIT 1999. LNCS, vol. 1661, pp. 51–64. Springer, Heidelberg (1999)
86. Vecera, S.P., Behrmann, M.: Attention and unit formation: A biased competition account of object-based attention. In: Shipley, T.F., Kellman, P.J. (eds.) From Fragments to Objects: Segmentation and Grouping in Vision, pp. 145–180. Elsevier Science, Amsterdam (2001)
87. Ward, R., Duncan, J., Shapiro, K.: The slow time-course of visual attention. Cognitive Psychology 30, 79–109 (1996)
88. Wiener, J.M., Büchner, S.J., Hölscher, C.: Taxonomy of human wayfinding tasks: A knowledge-based approach. Spatial Cognition and Computation 9(2), 152–165 (2009)
89. Wiener, J.M., Schnee, A., Mallot, H.A.: Use and Interaction of Navigation Strategies in Regionalized Environments. Journal of Environmental Psychology 24(4), 475–493 (2004)
90. Wilson, D., Sperber, D.: Relevance Theory. In: Ward, G., Horn, L. (eds.) Handbook of Pragmatics, pp. 607–632. Blackwell, Oxford (2002)
91. Winter, S., Baldwin, T., Cavedon, L., Stirling, L., Kealy, A., Duckham, M., Rajabifard, A., Richter, K.-F.: Starting to Talk about Place. In: Surveying and Spatial Sciences Conference. Surveying and Spatial Sciences Institute, Wellington, NZ
92. Zacks, J.M., Speer, N.K., Swallow, K.M., Braver, T.S., Reynolds, J.R.: Event perception: A mind/brain perspective. Psychological Bulletin 133, 273–293 (2007)
93. Zetzsche, C., Wolter, J., Schill, K.: Sensorimotor representation and knowledge-based reasoning for spatial exploration and localisation. Cognitive Processing 9, 283–297 (2008)

Mental Travel Primes Place Orientation in Spatial Recall*

Kai Basten[1], Tobias Meilinger[2], and Hanspeter A. Mallot[1]

[1] University of Tübingen, Cognitive Neuroscience Lab, Department of Biology, Tübingen, Germany,
mail@kai-basten.de, hanspeter.mallot@uni-tuebingen.de,
[2] Max-Planck-Institute for Biological Cybernetics, Department of Human Perception, Cognition and Action, Tübingen, Germany
tobias.meilinger@tuebingen.mpg.de

Abstract. The interplay of spatial long-term and working memories and the role of oriented and orientation-independent representations is an important but poorly understood issue in spatial cognition. Using a novel priming paradigm, we demonstrate that spatial working memory codes of a given location depend on previous tasks such as mental travels and are thus situated in behavioural context. In two experiments, 136 passersby were asked to sketch an image of a highly familiar city square either without or with prior metal travel, i.e. an imaginated walk along a route crossing the square. With prior mental travel participants drew the sketch more often in the orientation of the imagined route and less often in the orientation found without prior mental travel. This indicates that participants adjusted or selected information from long-term memory according to the situational context. We suggest that orientation priming plays a role in path planning and may facilitate way-finding afterwards. Possible mechanisms of orientation priming are discussed with respect to theories of orientation dependence in spatial memory.

Keywords: spatial cognition, priming, frame of reference, working memory, place recognition.

1 Introduction

Finding one's way in large-scale spaces is a core cognitive function in humans and animals. In this task, spatial knowledge from long-term memory has to be activated and transferred to a working memory stage where planning, reasoning, and verbalisation takes place.

The relation of spatial long-term and working memories is often discussed in terms of the distinction of allocentric and egocentric representations. Allocentric spatial knowledge comprises the geometric shape of an environment [20] as

* Supported by the German Federal Ministry of Education and Research (BMBF) within the Tübingen Bernstein Center for Computational Neuroscience (Grant No 01GQ1002A).

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 378–385, 2012.

well as object-to-object spatial relations either without a preferred orientation [18] or within an environmental reference frame [14,13]. In contrast, egocentric systems code self-to-object relations as in perceptual representations or in view-dependent snapshots of the environment [18], [20]. Although the recall of spatial memories in general may involve allocentric codes, visual place recognition will mainly work with egocentric codes [5,20]. Indeed, the anticipation of future viewpoints can facilitate later recognition of the previously anticipated views indicating egocentric involvement [1]. Subsuming allocentric representations (e.g., the structure of an object or a room) within an egocentric working memory stage may combine both types of spatial representations [12].
The information processing required to generate egocentric working memories from long-term memory depends on the assumed type of long-term place representation: If places are represented in a view-independent way [3], orientation has to emerge at the time of retrieval (transformation). If places are represented as a collection of views [8], the retrieval process will have to select (or interpolate) the appropriate view (selection). For path planning, Wiener & Mallot [21] hypothesised an egocentric working memory stage generated from the reference memory for each combination of ego position and local target locations along the path. Anticipating novel views within working memory may be involved in such processing [7].

The purpose of the present study was to test whether the orientation of a working memory representation built from spatial long-term memory can be primed by prior working memory activity such as path planning and mental travel. For this we used a novel priming paradigm and assessed the orientation specificity of retrieved spatial long-term knowledge as required for the production of place sketches. Our hypothesis predicts that the view orientation of these place sketches is primed by the direction of prior mental travel and thus by prior egocentric working memory representations. In contrast to most studies accessing memory representations of newly learned scenes, we tested highly-familiar long-term memory contents (i.e., a central square in the participants' home town) which is likely to have been encountered in many different orientations. Place sketches were analysed for view orientation. Sketches without prior mental travel (Experiment 1) worked as a baseline for sketches with prior mental travel (Experiment 2).

2 Experiment 1

2.1 Methods

Passersby at a University cafeteria were asked to sketch the "Holzmarkt", a well-known square in the medieval city centre of Tübingen (see Figure 1) within an 8 × 8 cm box provided on a DIN A6 (10.4 × 14.8 cm) sheet of paper. The University cafeteria was located approximately 2.5 km northwest of the Holzmarkt. About 30% of the people addressed agreed to participate. If participants asked in which perspective they should draw the square they were told to choose the perspective which they felt was most appropriate. After drawing, participants were asked to write down on the same sheet of paper their age, gender, place of residence (i.e., city district), and years of residency in Tübingen. Participation took approximately five minutes and was rewarded with candy.

From the 56 sketches obtained, six were excluded for incomplete data. Data from 27 women and 23 men (average age 22 years, SD = 2.2 years) were analyzed; on average, they had lived in or near Tübingen for 3.2 years (SD = 4.5 years). Three independent raters categorized the orientation of the sketches into north-up, south-up, east-up or west-up. They gave identical judgments for 49 of the 50 sketches (98%) corresponding to a very good chance corrected interrater-reliability of kappa = .96. Only the remaining 49 sketches were analyzed further.

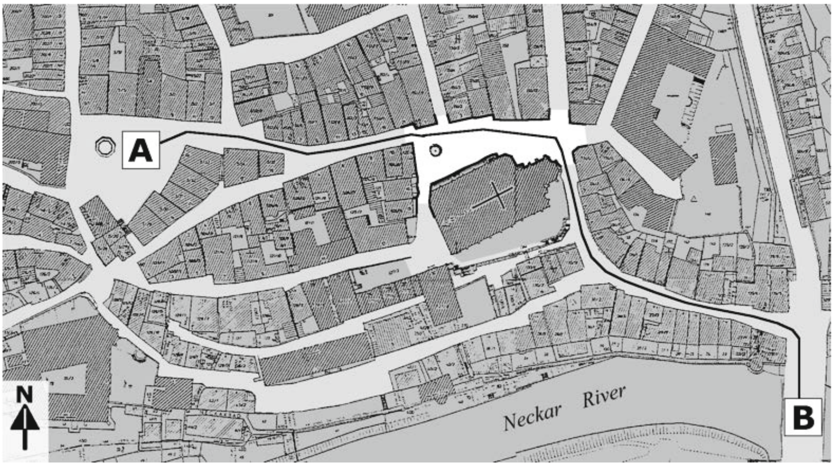


Fig. 1. City map of the Holzmarkt (white) and its surroundings including a prominent church bordering the square on the south side. Participants were asked to draw the Holzmarkt in both experiments. In Experiment 2 (orientation priming) participants were additionally asked to imagine walking a route (black line) either from A (Market) to B (Neckar Bridge, eastward route) or from B to A (westward route) before drawing. No map was shown to the participants.

2.2 Results

As shown in Figure 2 (left) participants sketched the square with a preferred orientation (χ^2 test against a uniform distribution: χ^2 (3, N = 49) = 78, p < .001). Eighty percent drew the sketches south-up in contrast to 25% expected by an equal distribution (one-tailed binomial test with $\pi = .25$ and N = 49: p < .001). Other orientations were drawn less often than 25% (north-up: 6.1%, p = .001; west-up: 10.2%, p = .009; east-up: 4.1%, p < .001). Individual differences (gender, age, time and place of residency) did not show any effects, neither here nor in the second experiment and thus are not further reported. Results are discussed in detail in conjunction with Experiment 2 below.

3 Experiment 2

3.1 Methods

General methods were the same as in Experiment 1, but before sketching, participants were asked to imagine walking a route across the Holzmarkt either in eastward or westward direction (Fig. 1). In the eastward condition, the written request *"Imagine to walk from the Market to the Neckar Bridge by crossing the Holzmarkt"* was handed to the participants. In the westward condition, Market and Neckar Bridge were exchanged. No participant of Experiment 2 participated also in Experiment 1. The experiment was run in multiple eastward/westward blocks, and participants were assigned to one of the two conditions in the sequence of recruitment.

All sketches of 81 participants could be analyzed (41 westwards, 40 eastwards, 42 women, 39 men). On average participants were 26 years old (SD = 9.5) and had lived in Tübingen for 6.7 years (SD = 9.1 years). All three raters agreed on orientation rating in all 81 sketches (interrater-reliability of kappa =1.0). All sketches were analysed further.

3.2 Results

The distributions of sketch orientations differed between the two priming conditions ($\chi^2(3, N = 81) = 41.5$, $p < .001$, Cramér's V = .72) and from Experiment

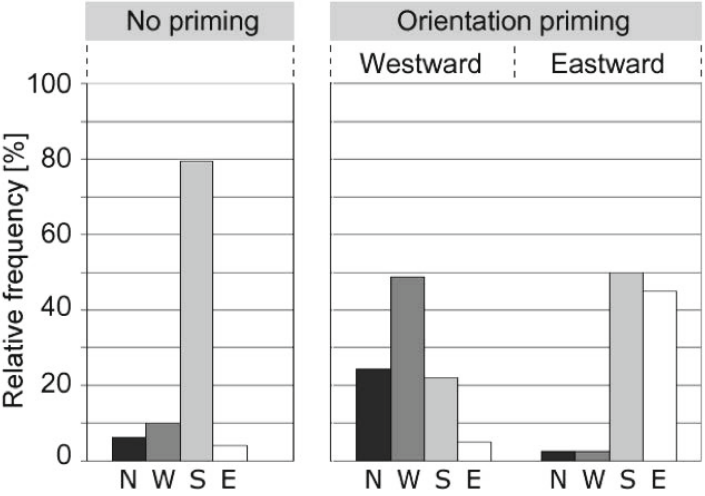


Fig. 2. Frequency of sketch orientations without (Experiment 1) or with orientation priming (Experiment 2). Sketch orientations: N = North-up, W = West-up, S = South-up, E = East-up.

1 conducted without priming (see Figure 2; route westwards: $\chi^2(3, N = 90) = 31.5$, $p < .001$, Cramér's V = .59; route eastwards: $\chi^2(3, N = 89) = 21.9$, $p < .001$, Cramér's V = .50).

Participants were more likely to orient their sketches according to the primed direction. In the westwards priming condition, participants more often drew a sketch west-up (49% tested with a one-tailed binomial test against a probability of 10.2% as observed in Experiment 1, N = 41: $p < .001$) and north-up (24% tested against 6.1%, N = 41, $p < .001$). Less often they drew a sketch south-up (22% vs. 79.6%, N = 41, $p < .001$) which was the orientation mainly drawn without priming. We observed no differences in the frequency of sketching east-up (4.9% vs. 4.1%, N = 41, p = .505). In the eastwards priming condition, participants drew a sketch east-up more often than in the no-priming condition (45% vs. 4.1%, N = 40, $p < .001$). The south-up orientation was chosen less frequently (50% vs. 79.6%, N = 40, $p < .001$), but still was the most frequent one in this condition. No differences were observed for the north-up (2.5% vs. 6.1%, N = 40, p = .290) and west-up orientations (2.5% vs. 10.2%, N = 40, p = .075).

4 Discussion

In Experiment 1, participants chose to draw sketches of the Holzmarkt square mostly with the south-up orientation. As all participants were highly familiar with the area (i.e., had experienced the square many times from multiple perspectives), the physical structure of the square itself likely determined the orientation of its recall [14]. Orientation preference may be derived from the geometric layout (east-west) as well as from the salient landmark (church) at the south side, or the south-up geographical slant of the square (see [16]). If the place representation relies on one intrinsic orientation, our data suggest that this is the southward orientation, despite the fact that the long axis of the square is east-west. Alternatively, place representations could contain views with multiple orientations as has been suggested for place recognition [19] and route following [9]. In this case, we need to assume that one view (i.e., southwards) is preferred. The salient ancient church bordering the square on the south side or a particular shop may influence the retrieved orientation as well. Such an effect may depend on participant's special interests, which were not evaluated.

In Experiment 2, when primed by imagining a route crossing the square, participants' sketches were more often oriented along the direction of the imagined route and less often in the orientation preferred without priming. Thus, mentally walking a route, as might also be done during route planning [3], primed the orientation in which a location was recalled. That is to say, orientation priming changed the accessibility of the orientation of a place memory.

Standard priming procedures affect the accessibility of a stimulus by presenting semantically related, often co-occurring, or perceptually similar stimuli [15], by presenting an object located on a route before versus after the target object [6], or by presenting objects located close by versus further away from the target or within the same versus different spatial regions [10].

The orientation priming procedure used in the present experiment differs from other forms of spatial priming in that the primed item is not a particular place but the orientation in which this place is sketched or imagined. Also, priming is triggered not by the spatial, perceptual, or conceptual proximity of a stimulus, but by the (assumed) orientation of a working memory representation used during prior mental travel.

Spatial updating performance tested within the same space used for learning can be influenced by verbal instructions that persuade participants to be tested in the same space or in a different space as in the learning session [17]. Further, learning a layout and then update this layout by walking around the layout showed an advantage for judgments of relative direction from the updated orientation which is not the case when no updating happened at all. Nevertheless, subsequent maps of the spatial layout were always drawn from the learned perspective. Hence, no orientation priming affect occur in the drawings. In contrast to our experiment this approach tested short-term knowledge that was perceived from one perspective only. The present work examines spatial knowledge of a highly familiar city space that was experienced many times from various perspectives.

The orientation priming effect reported here affects the process of retrieving egocentric working memories of places from long-term memory (LTM). This process depends on the structure of long-term memory which may be view-dependent, aligned to an intrinsic reference direction, or independent of orientation. If LTM place representations are sets of views taken from various viewpoints (view-dependent memory [5,20]), retrieval amounts to a selection process that picks one particular view to represent the place in working memory. In this case, orientation priming is the pre-activation of the view in LTM resulting in its subsequent selection. If, on the contrary, the environment is stored relative to an intrinsic reference direction and accessed more easily in that orientation [11,14], imagining it in a different perspective requires a transformation such as a mental rotation into that perspective. In orientation priming, the transformed perspective might persist in working memory for subsequent recall. In orientation-independent memory [3,18], the retrieval process may select a particular landmark object (rather than a specific view) and assumes the perspective under which this object appears when looking from the square. Orientation priming will then result from pre-activated object representations. In summary, different mechanisms will be responsible for orientation priming in the different long-term memory models: priming could result from the prior selection of views or landmark objects, or it might originate from the orientation of persisting working memory contents.

The distinction between the selection and transformation mechanisms for view-dependent long-term memory is akin to the distinction between representations for place recognition and representations for locating a goal relative to a given position, as suggested by Valiquette and McNamara [19]. It seems therefore possible that the two hypothesized mechanisms, selection and transformation, may even co-exist and support performance in different tasks.

Orientation priming affects the orientation of the recall of highly-familiar places. Our study substantiates that the retrieved orientation is not (only) dependent on the long-term memory content, but rather depends on the situational planning task. Further, our method of place sketches proves successful as appropriate tool to access the orientation of memory representations, which are well established in everyday long-term memory.

The present results were obtained within a mental travel task which primed recalled orientations. It will be an interesting opportunity for future research to test whether such an effect is also found in tasks involving physically walking along routes and whether this orientation priming will interfere with subsequent scene recognition tasks testing different perspectives (N, E, S, W).

Orientation priming is also in line with embodied cognition approaches which propose that representations, and in particular short-term representations, are based on sensorimotor and thus orientation-dependent representations [2,22]. Neuronal correlates supporting such view-dependent representation of scenes can be found in the parahippocampal place area, which is activated during mental imagery of places and in mental navigation [4]. Recently, also a computational model of anticipating and storing views was proposed which is consistent with both neuronal processing as well as the present results [7].

One final question concerns the function of orientation priming. The purpose of processing information about a specific route will generally be to follow this route afterwards. Recognition of locations along the route should be facilitated, if the representation is aligned with the upcoming perspective [1]. In this sense offline planning of routes might facilitate later online-cognition while walking the route. Orientation priming might thus help to effectively prepare for anticipated situations in way-finding tasks.

Acknowledgements. We would like to thank Sandra Holzer and participants of a spatial cognition course for help with data collection and analysis. Helpful comments on earlier versions of the text have been provided by Isabelle and Heinrich Bülthoff.

References

1. Amorim, M.-A.: ‘What is my Avatar Seeing?’: The Coordination of ‘Out of Body’ and ‘Embodied’ Perspectives for Scene Recognition Across Views. Visual Cognition 10(2), 157–199 (2003)
2. Barsalou, L.W.: Grounded Cognition. Annual Review of Psychology 59, 617–645 (2008)
3. Byrne, P., Becker, S., Burgess, N.: Remembering the Past and Imagining the Future: A Neural Model of Spatial Memory and Imagery. Psychological Review 114, 340–375 (2007)
4. Epstein, R.A.: Parahippocampal and Retrosplenial Contributions to Human Spatial Navigation. Trends in Cognitive Sciences 12, 388–396 (2008)
5. Gillner, S., Weiss, A., Mallot, H.A.: Visual Place Recognition and Homing in the Absence of Feature-Based Landmark Information. Cognition 109, 105–122 (2008)

6. Janzen, G.: Memory for Object Location and Route Direction in Virtual Large-Scale Space. The Quarterly Journal of Experimental Psychology 59, 493–508 (2006)
7. Lee, M., Duch, W., Sato, N.: Spatial Imagery of Novel Places Based on Visual Scene Transformation. Cognitive Systems Research 14, 26–36 (2012)
8. Mallot, H.A., Basten, K.: Embodied Spatial Cognition: Biological and Artificial Systems. Image and Vision Computing 27, 1658–1670 (2009)
9. Mallot, H.A., Gillner, S.D.: Route Navigation without Place Recognition. What is Recognized in Recognition-Triggered Responses? Perception 29, 43–55 (2000)
10. McNamara, T.P.: Mental Representations of Spatial Relations. Cognitive Psychology 18, 87–121 (1986)
11. McNamara, T.P., Sluzenski, J., Rump, B.: Human Spatial Memory and Navigation. In: Roediger III, H.L. (ed.) Cognitive Psychology of Memory. Learning and Memory: A Comprehensive, vol. 2, pp. 157–178. Elsevier, Oxford (2008)
12. Meilinger, T., Vosgerau, G.: Putting Egocentric and Allocentric into Perspective. In: Hölscher, C., Shipley, T.F., Olivetti Belardinelli, M., Bateman, J.A., Newcombe, N.S. (eds.) Spatial Cognition VII. LNCS (LNAI), vol. 6222, pp. 207–221. Springer, Heidelberg (2010)
13. Mou, W., Fan, Y., McNamara, T.P., Owen, C.B.: Intrinsic Frames of Reference and Egocentric Viewpoints in Scene Recognition. Cognition 106, 750–769 (2008)
14. Mou, W., McNamara, T.P.: Intrinsic Frames of Reference in Spatial Memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 28, 162–170 (2002)
15. Neelly, J.H.: Semantic Priming Effects in Visual Word Recognition: A Selective Review of Current Findings and Theories. In: Besner, D., Humphreys, G.W. (eds.) Basic Processing in Reading: Visual Word Recognition, pp. 264–336. Erlbaum, Hillsdale (1991)
16. Restat, J.D., Steck, S.D., Mochnatzki, H.F., Mallot, H.A.: Geographical Slant Facilitates Navigation and Orientation in Virtual Environments. Perception 33(6), 667–687 (2004)
17. Shelton, A.L., Marchette, S.A.: Where do you think you are? Effects of conceptual current position on spatial memory performance. Journal of Experimental Psychology: Learning, Memory, and Cognition 36, 686–698 (2010)
18. Sholl, M.J.: The Role of a Self-Reference System in Spatial Navigation. In: Montello, D.R. (ed.) COSIT 2001. LNCS, vol. 2205, pp. 217–232. Springer, Heidelberg (2001)
19. Valliquette, C., McNamara, T.P.: Different Mental Representations for Place Recognition and Goal Localisation. Psychonomic Bulletin & Review 14(4), 676–680 (2007)
20. Wang, F.R., Spelke, E.S.: Human Spatial Representation: Insights from Animals. Trends in Cognitive Sciences 6, 376–382 (2002)
21. Wiener, J., Mallot, H.A.: 'Fine-to-Coarse' Route Planning and Navigation in Regionalized Environments. Spatial Cognition and Computation 3, 331–358 (2003)
22. Wilson, M.: Six Views of Embodied Cognition. Psychonomic Bulletin & Review 9(4), 625–636 (2002)

Linking Cognitive and Computational Saliences in Route Information

Makoto Takemiya[1], Kai-Florian Richter[2], and Toru Ishikawa[3]

[1] Graduate School of Interdisciplinary Information Studies, The University of Tokyo
qq107405@iii.u-tokyo.ac.jp

[2] Department of Infrastructure Engineering, The University of Melbourne, Australia
krichter@unimelb.edu.au

[3] Center for Spatial Information Science, The University of Tokyo
ishikawa@csis.u-tokyo.ac.jp

Abstract. Finding a destination in a new spatial environment can be a daunting task. To aid navigation, many people take advantage of route directions, either provided by other people or by electronic navigation services. However, their effectiveness may be hampered if they are overly complex. While most people are generally good at focusing on important information, this is a challenge for navigation services. Thus, being able to automatically determine important points along a route that need to be included in route directions would provide a further step towards cognitively ergonomic navigation services. In the present study, methods for calculating the salience—or importance—of decision points are correlated with the frequency of decision points appearing in route directions. Results show that metrics based on the probability of a decision point being traversed and information-theoretic quantities of decision points correlate significantly with incidence in route directions, indicating that it is possible to identify crucial decision points in advance. This has implications for the design of navigation services that are able to adapt their assistance in real time.

Keywords: navigation, route directions, individual differences, salience.

1 Introduction

Imagine your paper is accepted to a conference and you have to go to a city in Germany for the first time. Fortunately, you have a friend who is familiar with the area, so you can get directions on how to get from your hotel to the conference venue. Unfortunately, your friend is a poor judge of which details are important enough to be included in route directions and they are overly complex and put a large burden on your cognitive faculties.

This story illustrates the importance of what to leave out when giving route directions. Implicit in knowing what to leave out (and in some cases what to include) is the realization that some points in spatial environments are more important than others for specific tasks, such as wayfinding. The present work focuses on facilitating the creation of cognitively ergonomic route directions by

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 386–404, 2012.

exploiting information about the importance of intersections in an environment. Focusing on the important points along a route, while at the same time ignoring those that are not, is critical to provide instructions that are concise, and easy to understand and to remember. While some humans struggle with this, distinguishing between important and unimportant points is a major challenge for existing navigation services.

The important points along a route are also crucial for presenting overview information on a route [32]. Turn-by-turn instructions are usually sufficient to reach a destination, however they limit the survey knowledge that people acquire and make it hard for users to cope with failures of their navigation services [30,34]. This may be one reason why contemporary GPS devices reduce wayfinding efficacy [16], compared with traditional maps, since it can be harder to grasp the overall wayfinding context.

Recent research aims for the automatic identification of the important points along a route. Takemiya and Ishikawa [39] were able to classify the performance of wayfinders using only existential information of decision points traversed in spatial environments, thus showing that an important relationship exists between the structure of environments and wayfinding performance. At the same time, work on the algorithmic generation of cognitively ergonomic route directions [7,20], which is based on empirical results in spatial cognition (e.g., [9,25]), highlights that specific elements of a route are more important than others when communicating information about how to get from an origin to a destination.

This paper links both the automatic identification of important decision points and the generation of cognitively ergonomic route directions. In particular, it explores the hypothesis that high-salience decision points identified by Takemiya and Ishikawa are important both for the classification of wayfinders and in cognitively ergonomic route directions that describe the routes taken by wayfinders. In other words, one outcome of this paper will be an answer to the question as to whether people's wayfinding behavior is an indicator of what they will deem relevant in communicating route information. Another, related, result will be whether those points identified in the classification are actually important for successful wayfinding (if a correlation with the route directions exists) or whether their salience may rather be an artifact of either the routes or classification schemes used (if not). This analysis will provide an important insight into how decision points that are important to wayfinding are related to route directions.

In the following sections we first outline previous work studying decision points, human wayfinding, and route directions (Section 2). We then explain how to determine the importance of decision points (Section 3). Section 4 discusses the implications that a relationship between cognitively ergonomic route directions and real-time classification of wayfinders would have. In Section 5, we present a setup and results of a study exploring this relationship. These results and their implications for the design of navigation services are discussed in Section 6.

2 The Role of Decision Points in Navigation Assistance

This section discusses the important role that decision points play in route following (Section 2.1) and then introduces relevant previous work that exploits decision points in supporting wayfinders (Sections 2.2, 2.3).

2.1 Decision Points in Route Following

Navigation in environments that are structured by path networks (e.g., urban environments, suburbs, or parks) can be conceptualized as movement along path segments (the streets) until a point is reached where several path segments meet (an intersection) [9,35]. At these points, the wayfinder has to make a decision on how to continue; accordingly, they are termed *decision points*.

Decision points are highly relevant to route following and for providing information about how to follow a specific route (so-called *route directions*) [1,8,25]. Here, wayfinders may go wrong and, thus, miss their destination, even when all they have to do along the path segments is continuing to move. Therefore, people often highlight information at decision points when providing route directions [1,28]; good route directions have been shown to focus on anchoring turning actions at decision points [9,25]. Also, decision points have been shown to receive special attention in the mental processing that happens while route following [18].

Arguably, not all decision points will be equally important when following a route. At some intersections, wayfinders continue in their current direction of movement. Such intersections tend to not receive much attention and they are only implicitly present in instructions [20,28]. For example, in the instruction 'turn left at the third intersection,' there are two intersections involved that are not mentioned at all, but the implicit assumption is that wayfinders will know to continue to go straight at those intersections. Further, the importance of decision points may depend on transport modality or the level of granularity in which instructions are provided [40,41], or the (perceived) difficulty of successfully navigating an intersection [14]. In this paper, we focus on the importance of decision points for successfully reaching a destination in pedestrian wayfinding.

2.2 Cognitively Ergonomic Route Directions

Cognitively ergonomic route directions aim for a lower cognitive load and enhanced location awareness at the same time [21]. They employ principles of human direction-giving, without adopting their deficits (such as slips and mix-ups of intersections). Cognitively ergonomic route directions utilize two important principles in instruction giving: 1) references to landmarks and 2) combining multiple consecutive decision points into a single instruction, termed *spatial chunking* [22]. Landmarks, if available, are important for organizing spatial knowledge, and are frequently referred to in human route directions [9,13]. Landmarks may signal crucial actions, locate other landmarks in relation to the referenced landmark, or confirm that the correct path is being followed [25,28]. Combining landmarks with spatial chunking results in a powerful mechanism to reduce the cognitive load by subsuming potentially large parts of a route (i.e., several consecutive decision points), with a single instruction, such as 'turn right at the church.'

Existing research has addressed the automatic generation of route directions that account for human principles of direction giving. Some of this work only covers part of the generation process, such as the identification [31] or integration [3] of landmarks. Others focus on generating instructions that mimic the way humans present such information [7,43], or that adapt to human conceptualization of wayfinding situations [23].

The approach of *context-specific route instructions* [35] generates route instructions for a given route that are easy to conceptualize and remember. Context-specific route directions account for environmental characteristics and a route's properties, by adapting the route directions to the surrounding environment. A computational process, called GUARD (Generation of Unambiguous, Adapted Route Directions), has been developed by Richter [33] for generating context-specific route instructions. GUARD unambiguously describes a specific route to the destination. Figure 1 provides an overview of the generation process.

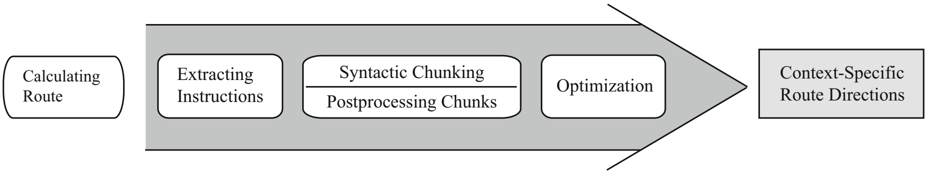


Fig. 1. Overview of GUARD, the generation process for context-specific route directions (adapted from [33])

The generation of context-specific route instructions is a three-step process. In the first step, for every decision point along the route from the starting point to the destination, all instructions that unambiguously describe the action to be taken are generated, resulting in a set of possible instructions for each decision point. GUARD accounts for different types of landmarks in generating instructions whose role in the route instructions depends on their locations relative to the route [10]. Next, GUARD performs *spatial chunking* (cf. [20,33] for more details). Finally, the actual context-specific route directions are generated. From all possible instructions, those that best describe the route are selected. As this is realized as an *optimization* process, "best" depends on the chosen optimization criterion. Optimization results in a sequence of chunks that cover the route from origin to destination. Due to the aggregation of instructions performed in chunking, some decision points will be represented only implicitly (e.g., points where the wayfinder continues straight and are thus not mentioned in the directions), thus reducing the communicated information. Therefore, the points represented explicitly in the sequence of chunks are the most salient points with respect to following a route.

2.3 Classifying Wayfinders

Research has shown that individual differences exist between different people's spatial abilities [4,11,17]. Thus, a desirable feature of navigation services would be functionality to adapt to these differences, which in turn requires ways to automatically classify differences between wayfinders.

Takemiya and Ishikawa [39] presented a practical way to classify the performance of wayfinders in real-time. As shown in Figure 2, the conditional probability-based classification of wayfinders reliably discriminated the performances of good and poor wayfinders, as defined by the criteria in [39]. To accomplish this classification, only information about the existence of decision points in wayfinders' route traversals was used. However, as can be seen in Figure 2, the efficacy of classification greatly improved after some specific decision points were used for classification, compared to other decision points that had much smaller effects on classification efficacy.

From this observation, Takemiya and Ishikawa developed methodology for calculating the salience, or importance, of decision points [38], with respect to the efficacy of classifying wayfinding performances.

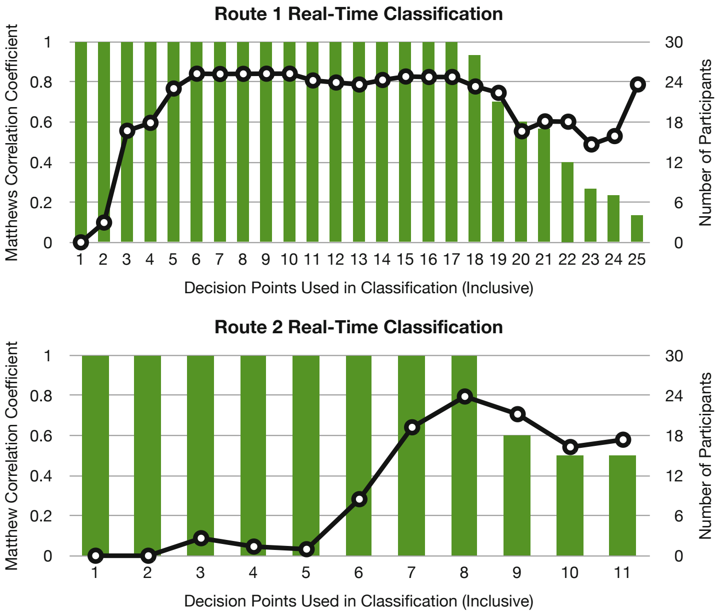


Fig. 2. Real-time route classification results, where the x-axis numbers the nth decision point along a route and only decision points 1 to n are used to classify wayfinders. Bars show the number of participants classified at each iteration. The line graph shows the Matthews correlation coefficient (maximum value of 1), which summarizes the confusion matrix of the classifier [27]. Adapted from [38].

3 Calculating Decision-Point Salience

This work considers two different kinds of decision-point salience: computational and cognitive saliences. Computational salience refers to the importance of decision points for discriminating between people exhibiting different levels of wayfinding performances (e.g., good or poor), whereas cognitive salience is the importance of decision points to wayfinders traversing an environment. The importance of computational salience to wayfinding was first suggested by Takemiya and Ishikawa [38], as a possible way to model how individual differences exhibit themselves in the distribution of where humans are likely to travel in an environment.

The cognitive salience of decision points is similar to the points identified in cognitively ergonomic route directions, since these are the points important to humans navigating an environment. It is important to note that decision points important to humans are not necessarily useful for classifying wayfinders. Cognitive salience in wayfinding has been previously considered in relation to the visual, cognitive, and structural qualities of landmarks [26,37]. Approaches have also been developed to calculate the cognitive salience of landmarks (e.g., [24,31]), and the prominence of individual streets in a street network (e.g., [5,42]).

To study the factors involved in calculating computational and cognitive saliences, the following methods for calculating the computational salience of decision-points were considered:

Conditional Probability Scores. The absolute value of the conditional probability of a decision point being in one class or another (either in the good or the poor routes).

Probability. The probability of a decision point being included in traversals in the generated route data.

PageRank. An algorithm for computing the standing probability of an ergodic Markov chain. This algorithm was developed for ranking web pages in search results for the Google search engine, but has also proven useful in other domains containing data structured in graphs. One notable use for processing spatial information was ranking popular intersections in a spatial environment [19]. Please see [2,29] for implementation and mathematical details of the algorithm.

Entropy and Information Gain. Shannon's information entropy [36] is a measure of the statistical heterogeneity of a set of data. Information gain is the amount by which entropy is decreased (homogeneity increased) by changing or partitioning a set of data. With respect to decision points, entropy is the heterogeneity of performance classes (i.e., good or poor) of wayfinders that traversed a point. The information gain of a decision point is thus calculated by taking the set of all route traversals and dividing them into two groups: those that contain the decision point under consideration, and those that do not. The difference between the sum of the entropies for the two separate sets of decision points and for all the points is the information gain, since the entropy of the independent sets of traversals is different than the entropy of the combined set of all traversals.

Graph Entropy. We define the graph entropy of a decision point to be equal to the difference between the entropy of all the routes and the entropy of all the routes *sans* the routes that contain the current point. In other words, we first calculate the entropy of all routes and then, for each decision point, we remove the routes that contain the current point from the set of all routes and recalculate the entropy using the same equation; the difference between entropies is the graph entropy in our definition.

Connectivity and Integration Metrics. Connectivity, local integration, and global integration are common metrics used in the space syntax literature [12]. These metrics are described in the following equation, adapted from Jiang [19] (similar to Jiang, we used a k_i of 2 for local integration):

$$\sum_{s=1}^{k} s \times N_s = \begin{cases} \textit{connectivity} & \Longleftrightarrow\ s = 1 \\ \textit{local integration} & \Longleftrightarrow\ 2 \leq s \leq k_i \\ \textit{global integration} & \Longleftrightarrow\ s = k \end{cases} \tag{1}$$

where N corresponds to the number of decision points with a shortest distance of s steps away from the current point.

Outflux Scores. The outflux scores are meta algorithms that were created by Takemiya and Ishikawa [38]. Using the salience measures, regions of similar scores emerge. Decision points bordering two of these regions may be important since they offer an opportunity to get from one region to another. This opportunity is captured in the outflux scores, which measure the difference between the score of the current point and neighboring points.

As each of the above algorithms and metrics can elucidate the importance of a decision point from a different perspective, it is important to examine whether any statistical relationships exist between decision points used in route directions and the salience of decision points calculated from each of the above methods.

4 Exploring the Relationship between Route Directions and Decision-Point Salience

The present work aims at exploring the relationship between the importance of decision points to route directions and to classifying wayfinders.

As discussed in Section 2.1, when following a route, some decision points are more important than others. These decision points divide a route through an environment into manageable parts. The splits typically occur at points where wayfinders must change orientation, often co-located with salient landmarks. Consequently, these points become functionally important to people navigating an environment; they will be prominent in their spatial memory of the environment [6,18]. Cognitively ergonomic route directions capture this prominence. Spatial chunking results in a division of route directions into manageable chunks; the end points of these chunks correspond to the important (or cognitively salient) decision points. This way, cognitively ergonomic route directions link behavior and communication. The decision points where important behavioral actions have to be performed are emphasized in the cognitively ergonomic route directions.

On the other hand, the real-time classification of wayfinders discussed in Section 2.3 is based on behavioral data (i.e., the decision points traversed). In fact, as pointed out in Section 3, a distinction needs to be made between computational and cognitive saliences. That is, while some decision points (and some metrics) may be useful for identifying decision points important (computationally salient) for classifying wayfinders, these decision points may not necessarily be prominent to the wayfinders themselves.

If a correlation exists, however, between those decision points with a high computational salience score and those decision points identified as prominent in cognitively ergonomic route directions, then there is a direct link between computational and cognitive saliences, as well as a link between behavior and communication in the computation of decision points' salience scores. More specifically, a correlation between computational salience scores and cognitively ergonomic route directions has at least the following four implications; it:

1. establishes significantly-correlated salience scores to be relevant for humans, i.e., they not only reflect computational importance (as discriminators in classification), but also cognitive importance (identifying points crucial for wayfinding);
2. provides a link between behavior and communication (this has immediate application to the real-time classification developed in [39] in navigation services; see discussion in Section 6.3);
3. enables the prediction (pre-computation) of decision points salient to humans in an environment without the need of actual behavioral data;
4. may provide a crucial step in developing algorithms for the automatic generation of overview information on a route to take [32].

To establish such a correlation, computational tests were performed, as reported in the next section.

5 Experimental Design and Results

This section will explain the data (Section 5.1) and experimental setup (Section 5.2) used to explore the relationship between route directions and the salience of decision points, and then present the results of the experiments (Section 5.3). In a nutshell, the study takes both computer-generated and human-subject routes through an environment, and generates context-specific route directions and computes salience scores of decision points for these routes, respectively. Then, the correlation between salience scores and the frequency of occurrence of decision points in the chunks for cognitively ergonomic route directions is calculated.

5.1 The Route Data

The experiment used the same environment as in previous studies by Takemiya and Ishikawa [39,38], namely an area in Nara, Japan. Two routes with two different pairs of start and goal locations were specified in these studies; the same pairs are used in this study again. In the present study, both computer-generated and human-subject routes were used.

For the computer-generated routes, we used the approach from [39], where a modified A* routing algorithm was used to create semi-random paths from the start to goal locations. A* is a heuristic search algorithm that uses an estimated cost to the goal (in our case the calculated Euclidean distance) to decide which path to take at each node. In order to generate routes that better explored the spatial environments, we modified A* to make a path leading from a node 'tabu' with a 10% random probability. Setting a path 'tabu' artificially blocks this path for the algorithm, making it unusable for path search and thus introduces variability into the generated routes. This algorithm was used to generate good and poor routes; for the latter, the heuristic function was inverted (i.e., maximizing distance to the goal). For each pair of start and goal locations, 750 routes were generated. After generation, good and poor routes were clustered to remove misclassified routes (e.g., some of the generated poor routes may in fact be good due to the randomness in the generation process). Overall, 924 good and 576 poor routes were generated; they are shown in Figure 3.

For the human-subject routes, we used routes traveled by the human participants in the experiment detailed in [39] (Figure 4). Overall, 60 routes (30 traversals each for routes 1 and 2) were collected in this experiment from people finding their way using paper maps, and subsequently classified as either good or poor. As reported in [38], since the probability of decision points being traversed correlated highly across the generated and empirically observed route traversals (0.93 and 0.83, respectively, for routes 1 and 2), it is expected that results should be similar for both the computer-generated and human-subject data. While humans do not always follow exactly the same routes they would describe to others [15], the routes collected from the human participants provide a good sample of which routes people deem reasonable to traverse the environment. Thus, they are used as a further source for exploring the relationship between route directions and decision-point salience. The computer-generated and human-subject routes were used to generate cognitively ergonomic route directions.

5.2 The Experimental Setup

A decision point is taken to be salient in route directions, if it appears as the last decision point in a chunk of context-specific route directions (Section 2.2). Therefore, for each set of both generated and human-subject routes (good and poor routes between the start and goal locations for routes 1 and 2 in the environment used in the study), context-specific route directions were calculated, and the last decision point of each chunk in these directions was recorded. To normalize results, the number of occurrences was divided by the number of routes in

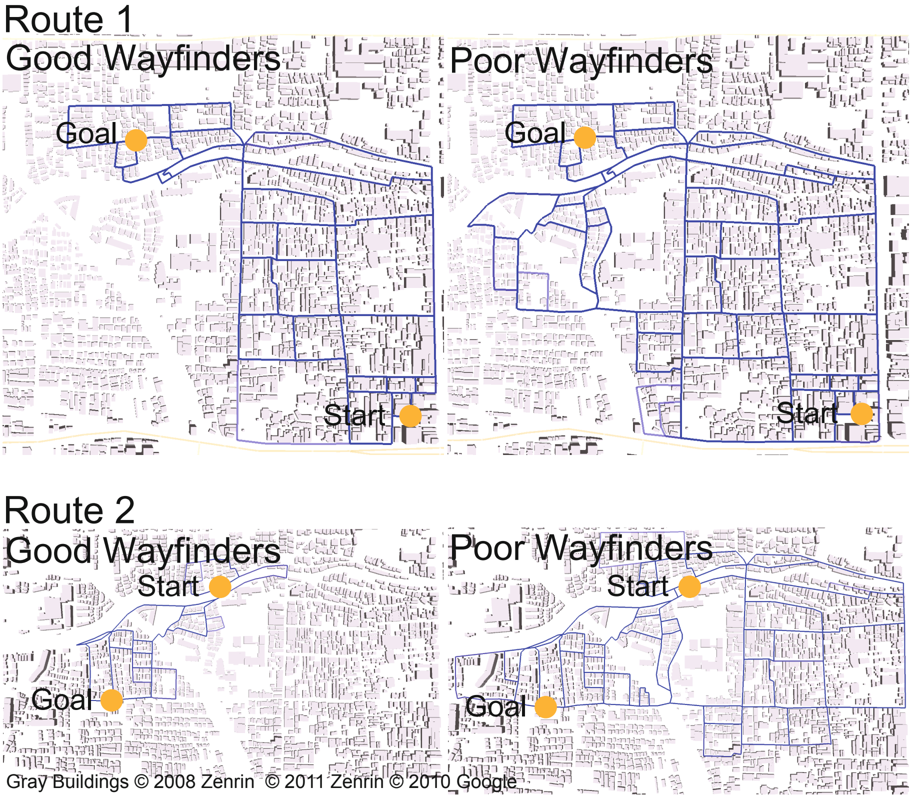


Fig. 3. Generated (a) good and (b) poor traversals for routes 1 and 2. Thicker (darker) lines indicate more traversals using an edge.

the respective set to get the frequency score. A score of 1 means that a decision point is salient (occurs at the end of a route-direction chunk) for every route of the set, a score of 0 that it is for none of the routes.

Computational salience scores of decision points were calculated using the methods listed in Section 3 (cf. also [38]). The same modified A* algorithm for generating random routes was used; to counter randomness in the generation process, 100 iterations of calculating salience scores were averaged. Again, this calculation was performed for each set of routes.

For both the computer-generated and human-subject routes, the Pearson correlation between the frequency of decision points in the cognitively ergonomic route directions and the salience scores was calculated. This was done for good, poor, and overall (combined good and poor) routes, for both start/goal location pairs for routes 1 and 2 in the empirical study. Additionally, calculations were also done for all good (good routes combined from routes 1 and 2 of the empirical study) and poor routes.

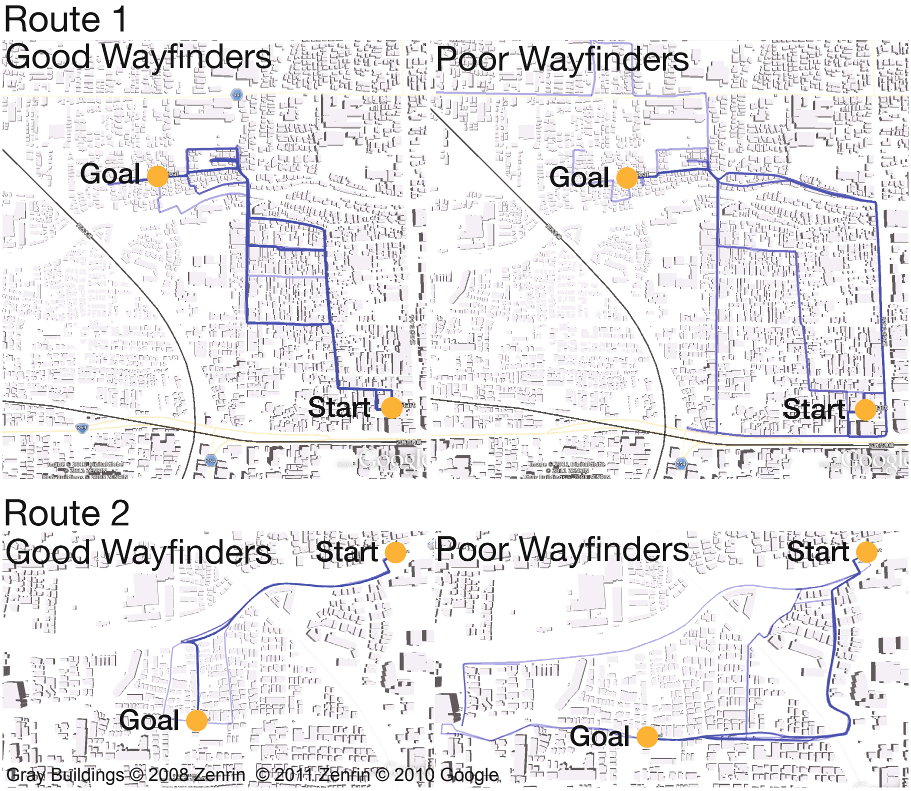


Fig. 4. Good and poor wayfinder traversals for routes 1 and 2, from the study conducted in [39]

For all analyses, we also correlated the frequency of decision points in route directions with a set of random numbers downloaded from `http://www.random.org/`. This was done to confirm whether the use of decision points in route directions is random or if there is some underlying, perhaps structural, reason for frequently using some decision points more than others.

5.3 Results

Figures 5 and 6 show the Pearson correlation values between decision-point frequency in route directions and calculated decision-point salience, for generated and empirical route data, respectively. Non-white cells are statistically significant, $p < .05$. For the generated route traversals, there was a consistent relationship between frequency of occurrence in route directions and conditional probability, probability, information gain, and entropy, across both routes 1 and 2. The strongest correlation across all the data was with probability. For human-subject routes, only probability was significantly correlated with frequency of occurrence in route directions, for both routes 1 and 2. Graph entropy was significantly negatively correlated for route 1 for both generated and human-subject routes, although this negative correlation did not manifest itself for route 2. This confirms differences between routes 1 and 2 that were shown in [38].

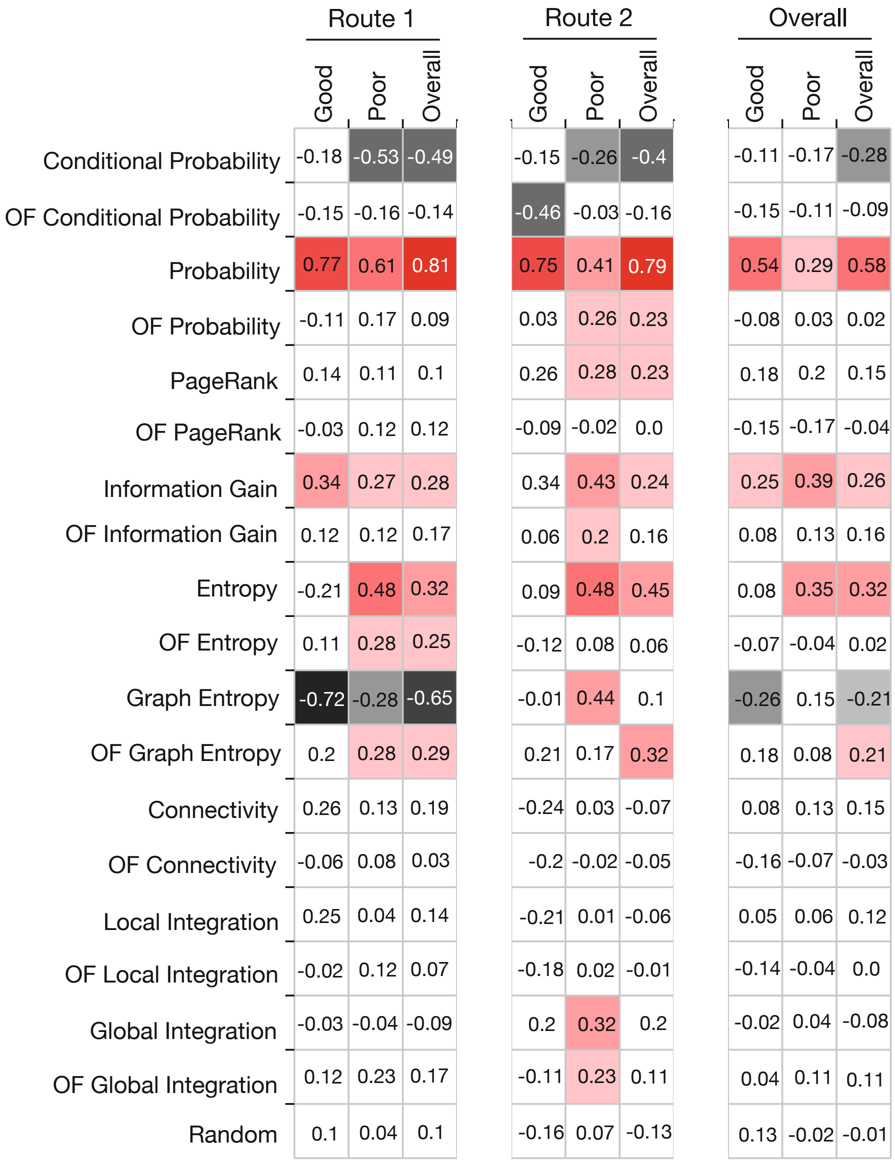
6 Discussion

Overall, the results of this study illustrate that there is a relationship between the computational salience of decision points and their frequency of occurrence in route directions. Thus, the four implications stated in Section 4 become valid. Significantly correlated computational salience scores for decision points reflect relevant cognitive aspects of wayfinding for humans; they also establish a link between behavior and communication (cf. [1,8], but see also [15]). Salient decision points mark crucial spots for successfully following a route; the developed salience measures allow their prediction without the need of behavioral data. The results and their implications are further discussed in the following.

6.1 Importance of Environmental Structure

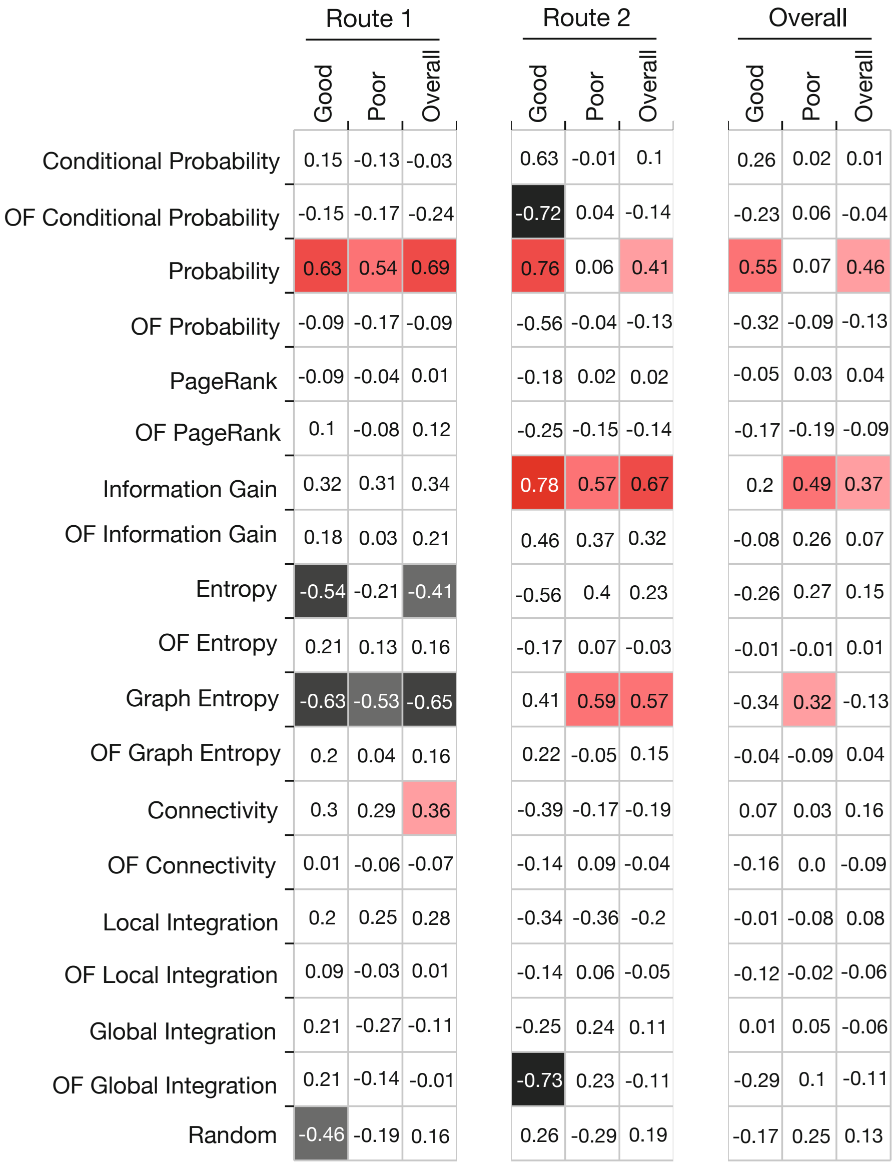
While the correlations with graph entropy were consistent between computer-generated and human-subject route data, information gain and entropy scores were different for routes 1 and 2. For route 1, the human-subject data also had a significant negative correlation with random values (for good routes only). The differences between human-subject and computer-generated data for route 1 are likely caused by the structure of the environment. As shown in Figure 4, the environment for route 1 is grid-like, so wayfinders were able to take many paths through the center of the grid that were all very similar in distance. At each point in the grid, it was essentially up to each individual wayfinder as to whether or not they would turn, or go straight and then turn at the next intersection. Thus, the turning pattern is probably close to random, since there are many different approximately-equally good paths a wayfinder may take. However, the good wayfinders in the human-subject study traversed only a small number of points in the environment, while the computer-generated routes explored many different possibilities. Thus, this correlation with random values only exhibited itself in the human-subject data. This type of pattern was not seen in route 2, since the connectivity of decision points limited the available paths, and hence there were no significant correlations with random values.

The major benefits of context-specific route directions is their ability to chunk several instructions into a single, concise instruction. This chunking in turn becomes more powerful with the presence of landmarks. Therefore, we initially added potential landmark objects to the geographical data set underlying our analysis (no landmark data was available from the original studies [38,39]). We added objects based on our local knowledge and plausible assumptions as to what might constitute a landmark in general (e.g., temples, shrines and unique buildings). However, our results show almost no difference in the patterns of



	Route 1			Route 2			Overall		
	Good	Poor	Overall	Good	Poor	Overall	Good	Poor	Overall
Conditional Probability	-0.18	-0.53	-0.49	-0.15	-0.26	-0.4	-0.11	-0.17	-0.28
OF Conditional Probability	-0.15	-0.16	-0.14	-0.46	-0.03	-0.16	-0.15	-0.11	-0.09
Probability	0.77	0.61	0.81	0.75	0.41	0.79	0.54	0.29	0.58
OF Probability	-0.11	0.17	0.09	0.03	0.26	0.23	-0.08	0.03	0.02
PageRank	0.14	0.11	0.1	0.26	0.28	0.23	0.18	0.2	0.15
OF PageRank	-0.03	0.12	0.12	-0.09	-0.02	0.0	-0.15	-0.17	-0.04
Information Gain	0.34	0.27	0.28	0.34	0.43	0.24	0.25	0.39	0.26
OF Information Gain	0.12	0.12	0.17	0.06	0.2	0.16	0.08	0.13	0.16
Entropy	-0.21	0.48	0.32	0.09	0.48	0.45	0.08	0.35	0.32
OF Entropy	0.11	0.28	0.25	-0.12	0.08	0.06	-0.07	-0.04	0.02
Graph Entropy	-0.72	-0.28	-0.65	-0.01	0.44	0.1	-0.26	0.15	-0.21
OF Graph Entropy	0.2	0.28	0.29	0.21	0.17	0.32	0.18	0.08	0.21
Connectivity	0.26	0.13	0.19	-0.24	0.03	-0.07	0.08	0.13	0.15
OF Connectivity	-0.06	0.08	0.03	-0.2	-0.02	-0.05	-0.16	-0.07	-0.03
Local Integration	0.25	0.04	0.14	-0.21	0.01	-0.06	0.05	0.06	0.12
OF Local Integration	-0.02	0.12	0.07	-0.18	0.02	-0.01	-0.14	-0.04	0.0
Global Integration	-0.03	-0.04	-0.09	0.2	0.32	0.2	-0.02	0.04	-0.08
OF Global Integration	0.12	0.23	0.17	-0.11	0.23	0.11	0.04	0.11	0.11
Random	0.1	0.04	0.1	-0.16	0.07	-0.13	0.13	-0.02	-0.01

Fig. 5. Correlations between frequency of decision-point occurrence in route directions using computer-generated routes, and the calculated salience scores. Colored cells are significant, $p < .05$. Route 1 to the left, route 2 in the middle, combined routes 1 and 2 to the right. Due to different numbers of points in the good, poor, and overall classes for routes 1 and 2, values for significant correlations are different for each column of results. Outflux is abbreviated as "OF."



	Route 1			Route 2			Overall		
	Good	Poor	Overall	Good	Poor	Overall	Good	Poor	Overall
Conditional Probability	0.15	-0.13	-0.03	0.63	-0.01	0.1	0.26	0.02	0.01
OF Conditional Probability	-0.15	-0.17	-0.24	-0.72	0.04	-0.14	-0.23	0.06	-0.04
Probability	0.63	0.54	0.69	0.76	0.06	0.41	0.55	0.07	0.46
OF Probability	-0.09	-0.17	-0.09	-0.56	-0.04	-0.13	-0.32	-0.09	-0.13
PageRank	-0.09	-0.04	0.01	-0.18	0.02	0.02	-0.05	0.03	0.04
OF PageRank	0.1	-0.08	0.12	-0.25	-0.15	-0.14	-0.17	-0.19	-0.09
Information Gain	0.32	0.31	0.34	0.78	0.57	0.67	0.2	0.49	0.37
OF Information Gain	0.18	0.03	0.21	0.46	0.37	0.32	-0.08	0.26	0.07
Entropy	-0.54	-0.21	-0.41	-0.56	0.4	0.23	-0.26	0.27	0.15
OF Entropy	0.21	0.13	0.16	-0.17	0.07	-0.03	-0.01	-0.01	0.01
Graph Entropy	-0.63	-0.53	-0.65	0.41	0.59	0.57	-0.34	0.32	-0.13
OF Graph Entropy	0.2	0.04	0.16	0.22	-0.05	0.15	-0.04	-0.09	0.04
Connectivity	0.3	0.29	0.36	-0.39	-0.17	-0.19	0.07	0.03	0.16
OF Connectivity	0.01	-0.06	-0.07	-0.14	0.09	-0.04	-0.16	0.0	-0.09
Local Integration	0.2	0.25	0.28	-0.34	-0.36	-0.2	-0.01	-0.08	0.08
OF Local Integration	0.09	-0.03	0.01	-0.14	0.06	-0.05	-0.12	-0.02	-0.06
Global Integration	0.21	-0.27	-0.11	-0.25	0.24	0.11	0.01	0.05	-0.06
OF Global Integration	0.21	-0.14	-0.01	-0.73	0.23	-0.11	-0.29	0.1	-0.11
Random	-0.46	-0.19	0.16	0.26	-0.29	0.19	-0.17	0.25	0.13

Fig. 6. Correlations between frequency of decision-point occurrence in route directions using human-subject route traversals, and calculated decision-point salience. Colored cells are significant, $p < .05$. Route 1 to the left, route 2 in the middle, combined routes 1 and 2 to the right.

significant correlations for cognitively ergonomic route directions using landmarks and those that did not. Consequently, results based on 'landmark data' are not reported here. The lack of any effect of incorporating landmarks may be attributed to the chosen environmental setup. Routes in the environment were relatively short, often exhibiting several turns. Typically, each turn marks a salient action and, consequently, a decision point in the route directions. If routes are short, chunks can be created based on a principle of counting intermediate decision points ('turn right at the third intersection'), compared to long routes where landmarks are required to mark these turning points ('turn right at the church' vs. 'turn right at the 21st intersection,' an instruction that is hard to execute due to the likelihood of miscounting or missing intersections; cf. [23,20]). Another factor could be that the landmarks chosen based on our knowledge of the area could be different from what *in situ* wayfinders navigating a new area would consider as landmarks.

Thus, while the use of landmarks may make the generated route directions easier to understand, by the nature of the experimental design, here they are not crucial in combining instructions. This study looks only at the frequency of decision points in route directions, not the ease of understanding. However, the results may also suggest a more general effect that cognitive salience is more a function of the structure of space (i.e., structural salience) than of visually or semantically salient features, such as landmarks. Future work will study this effect further.

6.2 Cognitive versus Computational Salience

With respect to computational salience (i.e., the importance of decision points for classifying wayfinders), Takemiya and Ishikawa [38] recommended the use of outflux local integration, which measures the local graph-theoretic connectedness of a decision point to nearby points; for computing cognitive salience, outflux probability was recommended.

In the present study, consistent across all results is the significance of the correlation between the frequency of decision points in route directions and the probability that decision points occurred in route traversals (except for some of the poor wayfinding ability results using human-subject data). This indicates that some decision points are crucial for successfully navigating an environment, and that these will also be pointed out in instructions on how to navigate the environment: cognitively ergonomic route directions highlight those points along a route that are important for successfully finding the way to a destination. Since outflux probability, a derivative of probability, was recommended in [38] for cognitive salience, our results support the importance of probability for cognitive salience. This is further emphasized by the significantly correlated outflux probability for the generated 'overall' routes for route 2. Since this relationship is only seen in this setting, it is likely that the structure of the environment contributed to this relationship. Future work will explore the relationship between environmental structure, probability of point traversal, and cognitive salience.

The results also indicate that some of the measures used for calculating salience scores (see Section 3) are only relevant for classifying wayfinders, but have no relationship to route directions and, thus, to communication. There are no significant correlations with local measures of graph-theoretic connectivity (i.e., connectivity and local integration), which were previously identified to be suitable for calculating computational salience. This is somewhat intuitive. Since those decision points with high frequency in route directions are important for successfully navigating an environment, they will be traversed by a variety of wayfinders (good and poor), which makes them less useful for discriminating classes of wayfinding performance.

6.3 Incorporating Decision-Point Salience into Route Directions

The results provide a quantitative basis for determining important points for wayfinding. Overall, when the abilities of wayfinders are unknown, probability, information gain, entropy, graph entropy, and outflux graph entropy will all work well for determining points that are crucial for successfully navigating an environment. Generating cognitively ergonomic route directions that incorporate instructions to pass these points will reduce cognitive load and likely will decrease wayfinding errors.

Moreover, generating and presenting high-level information on a route (the overview information discussed in [32]) may help to prepare wayfinders for what is to come and, this way, may reduce the negative effects that have been identified for current navigation services (cf. [16,30,34]). Focusing on the salient decision points in the overview information will highlight the crucial spots along a route and will ensure that they are already known to a wayfinder prior to route-following.

Furthermore, notable differences between good and poor routes exist for salience scores based on information-theoretic methods. For routes 1 and 2, the significant correlations with entropy and graph entropy differed between good and poor wayfinders. This may be exploited to identify these good and poor wayfinders, which can be done online as demonstrated by [39]. This way, route directions can be adapted (online as well) to an individual's wayfinding performance by, for example, providing more details to poorer wayfinders while presenting good wayfinders with more concise instructions that may focus on the identified crucial decision points.

Future work will examine specific ways of incorporating salience information into route directions and the generation of overview information on a route, and test this empirically.

7 Conclusions

Decision points are the key locations where humans navigating an environment must decide which of several available paths to take to their destination. The present work explores the relationship between the importance of decision points to both cognitively ergonomic route directions and to wayfinding behavior. A variety of algorithms were used to calculate the salience–or importance–of decision points; these were correlated with the frequency of decision points' occurrence in cognitively ergonomic route directions.

Our results show that salience scores based on the probability of a decision point being traversed, as well as information-theoretic quantities, are significantly correlated with frequency of decision points in cognitively ergonomic route directions. This outcome suggests the efficacy of using salience scores to improve the automatic generation of route directions. Finally, incorporating our method of calculating computational and cognitive saliences into existing frameworks for using cognitive salience in route directions is an obvious step for future work.

References

1. Allen, G.L.: Principles and practices for communicating route knowledge. Applied Cognitive Psychology 14(4), 333–359 (2000)
2. Bryan, K., Leise, T.: The $25,000,000,000 eigenvector: The linear algebra behind Google. SIAM Rev. 48(3), 569–581 (2006)
3. Caduff, D., Timpf, S.: The landmark spider: Representing landmark knowledge for wayfinding tasks. In: Barkowsky, T., Freksa, C., Hegarty, M., Lowe, R. (eds.) Reasoning with Mental and External Diagrams: Computational Modeling and Spatial Assistance - Papers from the 2005 AAAI Spring Symposium, Menlo Park, CA, pp. 30–35 (2005)
4. Carlson, L.A., Hölscher, C., Shipley, T.F., Dalton, R.C.: Getting lost in buildings. Current Directions in Psychological Science 19(5), 284–289 (2010)
5. Claramunt, C., Winter, S.: Structural salience of elements of the city. Environment and Planning B 34(6), 1030–1050 (2007)
6. Couclelis, H., Golledge, R.G., Gale, N., Tobler, W.: Exploring the anchor-point hypothesis of spatial cognition. Journal of Environmental Psychology 7, 99–122 (1987)
7. Dale, R., Geldof, S., Prost, J.P.: Using natural language generation in automatic route description. Journal of Research and Practice in Information Technology 37(1), 89–105 (2005)
8. Daniel, M.P., Denis, M.: Spatial descriptions as navigational aids: A cognitive analysis of route directions. Kognitionswissenschaft 7, 45–52 (1998)
9. Denis, M.: The description of routes: A cognitive approach to the production of spatial discourse. Cahiers Psychologie Cognitive 16(4), 409–458 (1997)
10. Hansen, S., Richter, K.-F., Klippel, A.: Landmarks in OpenLS — A Data Structure for Cognitive Ergonomic Route Directions. In: Raubal, M., Miller, H.J., Frank, A.U., Goodchild, M.F. (eds.) GIScience 2006. LNCS, vol. 4197, pp. 128–144. Springer, Heidelberg (2006)
11. Hegarty, M., Montello, D., Richardson, A., Ishikawa, T., Lovelace, K.: Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. Intelligence 34(2), 151–176 (2006)
12. Hillier, B.: Space is the Machine: A Configurational Theory of Architecture. Cambridge University Press, Cambridge (1996)
13. Hirtle, S.C., Jonides, J.: Evidence of hierarchies in cognitive maps. Memory & Cognition 13(3), 208–217 (1985)

14. Hirtle, S.C., Richter, K.F., Srivinas, S., Firth, R.: This is the tricky part: When directions become difficult. Journal of Spatial Information Science (1), 53–73 (2010)
15. Hölscher, C., Tenbrink, T., Wiener, J.: Would you follow your own route description? cognitive strategies in urban route planning. Cognition 121(2), 228–247 (2011)
16. Ishikawa, T., Fujiwara, H., Imai, O., Okabe, A.: Wayfinding with a GPS-based mobile navigation system: A comparison with maps and direct experience. Journal of Environmental Psychology 28(1), 74–82 (2008)
17. Ishikawa, T., Montello, D.: Spatial knowledge acquisition from direct experience in the environment: Individual differences in the development of metric knowledge and the integration of separately learned places. Cognitive Psychology 52(2), 93–129 (2006)
18. Janzen, G., van Turennout, M.: Selective neural representation of objects relevant for navigation. Nature Neuroscience 7(6), 673–677 (2004)
19. Jiang, B.: Ranking spaces for predicting human movement in an urban environment. International Journal of Geographical Information Science 23(7), 823–837 (2009)
20. Klippel, A., Hansen, S., Richter, K.F., Winter, S.: Urban granularities - a data structure for cognitively ergonomic route directions. GeoInformatica 13(2), 223–247 (2009)
21. Klippel, A., Richter, K.F., Hansen, S.: Cognitively ergonomic route directions. In: Karimi, H. (ed.) Handbook of Research on Geoinformatics, ch. XXIX, pp. 230–237. IGI: Information Science Reference, Hershey (2009)
22. Klippel, A., Tappe, H., Habel, C.: Pictorial Representations of Routes: Chunking Route Segments during Comprehension. In: Freksa, C., Brauer, W., Habel, C., Wender, K.F. (eds.) Spatial Cognition III. LNCS (LNAI), vol. 2685, pp. 11–33. Springer, Heidelberg (2003)
23. Klippel, A., Tappe, H., Kulik, L., Lee, P.U.: Wayfinding choremes — a language for modeling conceptual route knowledge. Journal of Visual Languages and Computing 16(4), 311–329 (2005)
24. Klippel, A., Winter, S.: Structural Salience of Landmarks for Route Directions. In: Cohn, A.G., Mark, D.M. (eds.) COSIT 2005. LNCS, vol. 3693, pp. 347–362. Springer, Heidelberg (2005)
25. Lovelace, K.L., Hegarty, M., Montello, D.R.: Elements of Good Route Directions in Familiar and Unfamiliar Environments. In: Freksa, C., Mark, D.M. (eds.) COSIT 1999. LNCS, vol. 1661, pp. 65–82. Springer, Heidelberg (1999)
26. Maaß, W.: How Spatial Information Connects Visual Perception and Natural Language Generation in Dynamic Environments: Towards a Computational Model. In: Kuhn, W., Frank, A.U. (eds.) COSIT 1995. LNCS, vol. 988, pp. 223–240. Springer, Heidelberg (1995)
27. Matthews, B.W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure 405(2), 442–451 (1975)
28. Michon, P.-E., Denis, M.: When and Why Are Visual Landmarks Used in Giving Directions? In: Montello, D.R. (ed.) COSIT 2001. LNCS, vol. 2205, pp. 292–305. Springer, Heidelberg (2001)
29. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab (November 1999)

30. Parush, A., Ahuvia, S., Erev, I.: Degradation in Spatial Knowledge Acquisition When Using Automatic Navigation Systems. In: Winter, S., Duckham, M., Kulik, L., Kuipers, B. (eds.) COSIT 2007. LNCS, vol. 4736, pp. 238–254. Springer, Heidelberg (2007)
31. Raubal, M., Winter, S.: Enriching Wayfinding Instructions with Local Landmarks. In: Egenhofer, M., Mark, D.M. (eds.) GIScience 2002. LNCS, vol. 2478, pp. 243–259. Springer, Heidelberg (2002)
32. Richter, K.F.: From turn-by-turn directions to overview information on the way to take. In: Gartner, G., Cartwright, W., Peterson, M.P. (eds.) Location Based Services and TeleCartography. Lecture Notes in Geoinformation and Cartography, pp. 205–216. Springer, Heidelberg (2007)
33. Richter, K.F.: Context-Specific Route Directions - Generation of Cognitively Motivated Wayfinding Instructions, DisKI, vol. 314. IOS Press, Amsterdam (2008) also published as SFB/TR 8 Monographs vol. 3
34. Richter, K.F., Dara-Abrams, D., Raubal, M.: Navigating and learning with location based services: A user-centric design. In: Gartner, G., Li, Y. (eds.) Proceedings of the 7th International Symposium on LBS and Telecartography, pp. 261–276 (2010)
35. Richter, K.-F., Klippel, A.: A Model for Context-Specific Route Directions. In: Freksa, C., Knauff, M., Krieg-Brückner, B., Nebel, B., Barkowsky, T. (eds.) Spatial Cognition IV. LNCS (LNAI), vol. 3343, pp. 58–78. Springer, Heidelberg (2005)
36. Shannon, C.: A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423 (1948)
37. Sorrows, M.E., Hirtle, S.C.: The Nature of Landmarks for Real and Electronic Spaces. In: Freksa, C., Mark, D.M. (eds.) COSIT 1999. LNCS, vol. 1661, pp. 37–50. Springer, Heidelberg (1999)
38. Takemiya, M., Ishikawa, T.: Determining decision-point salience for real-time wayfinding support. Journal of Spatial Information Science (in press, 2012)
39. Takemiya, M., Ishikawa, T.: I Can Tell by the Way You Use Your Walk: Real-Time Classification of Wayfinding Performance. In: Egenhofer, M., Giudice, N., Moratz, R., Worboys, M. (eds.) COSIT 2011. LNCS, vol. 6899, pp. 90–109. Springer, Heidelberg (2011)
40. Tenbrink, T., Winter, S.: Variable granularity in route directions. Spatial Cognition & Computation: An Interdisciplinary Journal 9(1), 64–93 (2009)
41. Timpf, S., Volta, G.S., Pollock, D.W., Frank, A.U., Egenhofer, M.J.: A Conceptual Model of Wayfinding using Multiple Levels of Abstraction. In: Frank, A.U., Formentini, U., Campari, I. (eds.) GIS 1992. LNCS, vol. 639, pp. 348–367. Springer, Heidelberg (1992)
42. Tomko, M., Winter, S., Claramunt, C.: Experiential hierarchies of streets. Computers, Environment and Urban Systems 32(1), 41–52 (2008)
43. Tversky, B., Lee, P.U.: Pictorial and Verbal Tools for Conveying Routes. In: Freksa, C., Mark, D.M. (eds.) COSIT 1999. LNCS, vol. 1661, pp. 51–64. Springer, Heidelberg (1999)

Representing Space: Exploring the Relationship between Gesturing and Geoscience Understanding in Children

Bryan J. Matlen[1], Kinnari Atit[2], Tilbe Göksun[3],
Martina A. Rau[1], and Maria Ptouchkina[4]

[1] Carnegie Mellon University
Pittsburgh, PA 15212 USA
{bmatlen,mrau}@cmu.edu
[2] Temple University
Philadelphia, PA 19122 USA
kinnari.atit@temple.edu
[3] University of Pennsylvania,
Philadelphia, PA 19104 USA
tilbe@mail.med.upenn.edu
[4] Capgemini,
Chicago, IL 60606 USA
maria.ptouchkina@capgemini.com

Abstract. Learning in science requires the ability to think spatially and gesturing has been shown to ground students' understanding of spatial relationships. However, despite theoretical reasons to hypothesize a relation between the use of gesture and science understanding, few studies provide strong empirical evidence of a link between these factors. In the present study, we explored whether spontaneous use of gesture is associated with children's understanding of spatially intensive geoscience concepts. Eight- to sixteen-year-old children ($N = 27$, $M = 11.79$ yrs) were provided instruction about the causal mechanisms of mountain and volcano formation and were then interviewed for their understanding of these mechanisms. Analyses of children's responses to the interview questions revealed significant positive correlations between children's knowledge of geoscience and the spontaneous production of iconic, content-relevant gestures. These findings help to empirically establish a long hypothesized link between gesture and science understanding, and suggest that gesturing may facilitate understanding of difficult spatial science concepts.

Keywords: Gesture, Spatial Reasoning, Geoscience Education, Children.

1 Introduction

Scientists often gesture when they reason about and explain science concepts (Goodwin, 2007; Kastens, Liben, & Agrawal, 2006; Resnick, Atit, Goksun, & Shipley, 2011). This phenomenon is not surprising, given that gesturing can facilitate spatial reasoning (Alibali, 2005; Goldin-Meadow, 2000) and spatial reasoning is an

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 405–415, 2012.

important aspect of learning and communicating scientific concepts. For instance, recent studies have documented empirical links between spatial reasoning abilities and understanding in scientific disciplines (Kozhevnikov, Motes, & Hegarty, 2007; Coleman & Gotch, 1998; Hegarty, Crookes, Dara-Abrams, & Shipley, 2008; Orion, Ben-Chaim, & Kali, 1997). Furthermore, real-world scientists commonly utilize spatial representational tools – such as models (Nersessian, 2009), diagrams (Novick, 2006), and sketching (Ainsworth, Prain, & Tytler, 2011) – along with gestures (Goodwin, 2007; Kastens, Liben, & Agrawal, 2006; Resnick, Atit, Göksun, & Shipley, 2011) to reason about scientific concepts.

Though scientists often utilize representational tools such as gesture, still relatively little is known about the relationship between novice science learners' spontaneous use of gesture during the course of science learning. Gesturing might be particularly important for novices who lack the domain knowledge and spatial reasoning abilities of highly trained scientists. The present study focuses on the use of gesture and its relation to children's understanding of elementary geoscience concepts, which is one of the most spatially intensive amongst the scientific disciplines (Hegarty, Crookes, Dara-Abrams, & Shipley, 2008; Jee et al., 2010; Kastens, Liben, & Agrawal, 2008; Liben, Kastens, & Christensen, 2011). We first review literature outlining how gesture influences spatial thought, and then we discuss the role that gestures may play in the acquisition of early geoscience concepts.

1.1 Gesture and Spatial Reasoning

Prior research has revealed at least three ways in which gesturing augments spatial reasoning. The first is that gesture promotes attention to spatial information (Alibali, 2005; Alibali, Spencer, Knox, & Kita, 2011; Rimè, Shiaratura, Hupet, & Ghysselinckx, 1984). For example, Sauter and colleagues showed that eight- to ten-year-old children who used gestures in communicating relations among locations tended to produce more spatial information in their speech than children who did not use gesture (Sauter, Uttal, Alman, Goldin-Meadow, & Levine, in press). In addition, children who produced gesture-speech mismatches when predicting which way a balance beam will fall – that is, their gestures reflected distance information but their speech reflected only weight information – were more likely than children who did not produce such gesture-speech mismatches to explicitly recognize the importance of both weight and distance information later on in learning (Pine, Lufkin, & Messer, 2004). Thus, recruitment of gesture can cue attention to spatial information.

Another way in which gesture can augment spatial thinking is that it can allay demands placed on working memory. De Ruiter (1998) found that speakers were more likely to gesture when they needed to convey spatial information of objects and when visual representations of those objects were unavailable. This finding was replicated with both objects that were difficult to verbally describe (e.g., patterns of lines as shapes), as well as with objects that were easily verbalized (e.g., a flower, a clock, etc; Morsella & Krauss, 2004). Taken together, these studies suggest that gesture acts as a representational tool that allows speakers to more fluently and accurately convey spatial content (Alibali, 2005; Wesp, Hess, Keutmann & Wheaton, 2001).

Finally, gesture appears to facilitate the spatial reasoning process itself. A number of studies have found that participants who spontaneously gesture during spatial tasks perform better at those tasks than individuals who do not gesture (e.g., Cook & Goldin-Meadow, 2006). Rauscher, Krauss, and Chen (1996) found that participants who were prohibited from gesturing while describing a series of action cartoons verbally produced less spatial content than participants who were allowed to gesture. Another study showed that even preschool-age children benefit from gesturing in spatial transformation tasks (Ehrlich, Levine & Goldin-Meadow, 2006; Ping, Ratliff, Hickey, & Levine, 2001).

In sum, gesturing can act as a useful representational tool for thinking about spatial information for both children and adults. Next we consider how gesture may influence students' reasoning in the highly spatial domain of geoscience.

1.2 Gesture and Geoscience Learning

Prior research suggests that expert geoscientists frequently utilize gesture during the course of scientific reasoning. For example, Kastens, Liben, and Agrawal (2008a) documented geoscientists' use of gesture as they attempted to integrate 3-D models of geological structures with their observations of artificial rock outcrops. This investigation revealed that geoscientists repeatedly made deictic (i.e., pointing) and iconic (i.e., hand movements intended to represent concrete entities) gestures to refer to and describe geological phenomena. Similar findings are reported when structural geology experts were asked to read and explain a geologic map (Resnick, Atit, Goksun, & Shipley, 2011)

To our knowledge, however, only a handful of studies have addressed whether novice geoscience learners spontaneously utilize gesture. One case study followed a group of three 6th-grade students in depth over the course of a unit on plate tectonics (Singer, Radinksy, & Goldman, 2008) and found that students used gestures to create a shared representation, sometimes correcting or modifying their peers' gestures during the course of learning. In addition, Liben, Christensen, and Kastens (2010) asked university students to complete tasks related to the geologic concepts of strike and dip (i.e., of methods of describing the orientation of tilted layers of rock in three-dimensional space) and found that students who had no prior experience with the geologic terms were the only group of participants who gestured during the reading task.

Though these studies provide valuable process descriptions of how experts and novices incorporate gestures when learning geoscience, the nature of the relationship between gesturing and geoscience learning is still unclear: do novice geoscience learners gesture more frequently? Or do they gesture less and simply make better use of gestures that they produce? In this paper, we report an analysis of novice learners' gesturing in a laboratory investigation.

1.3 The Present Study

The primary aims of the present study were to explore 1) whether there is a relationship between gesturing and children's geoscience understanding, and 2) to document the nature of this relationship. This research was conducted within the context of teaching children about an important concept in elementary geoscience education: plate tectonics. Plate tectonics is the study of how the earth's plates are driven and shaped by geological forces that keep them in constant motion, which is a fundamental mechanism involved in the formation of volcanoes and mountains. Despite its importance, however, children have been shown to exhibit a variety of misconceptions in this domain (Gobert, 2004; Matlen, Vosniadou, Jee, & Ptouchkina, 2011; May, Hammer, & Roy, 2006).

Given that expert scientists commonly gesture, and that gesturing facilitates spatial reasoning in cognitive tasks (e.g., Alibali et al., 2011; Cook & Goldin-Meadow, 2006), we hypothesized that children who spontaneously produce gestures would exhibit better understanding of geoscience overall than children who do not use gestures.

2 Method

The study reported in the present paper was part of a larger experiment that investigated the use of instructional text and graphics on the teaching of geoscience concepts. Here, we report the methods and results relevant to our investigation of gesturing and geoscience learning.

2.1 Participants

Participants were 27 eight- to sixteen-year-old children (*M* = 11.79, *SD* = 2.29, 14 girls, 13 boys) recruited from the Pittsburgh area. We recruited children from this age range to represent a broad sample of K – 12 students.

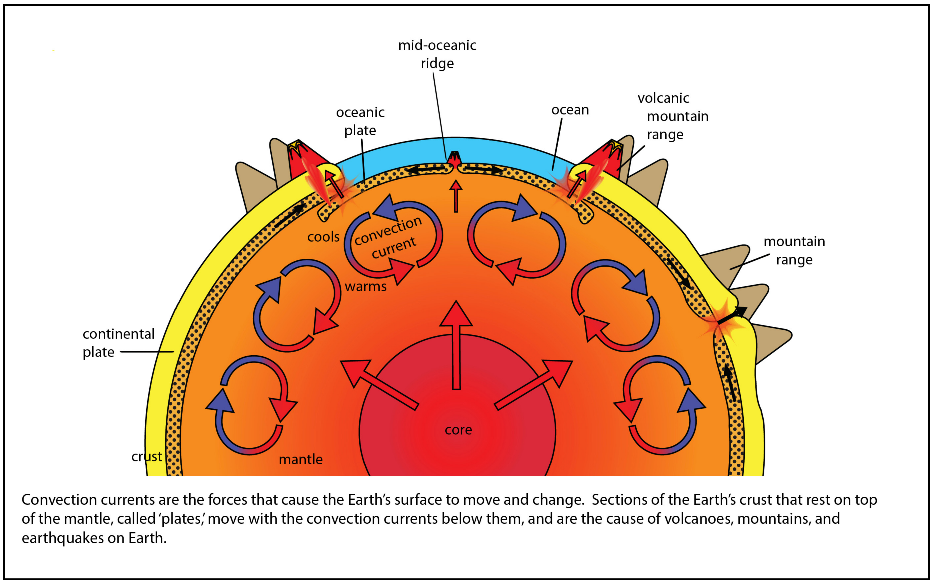
2.2 Materials and Procedure

All children were tested individually in a laboratory at Carnegie Mellon University. The experiment was comprised of two phases – the instruction and interview phases – that are described in detail below.

Instruction Phase. Children were asked to view instruction on a computer screen that consisted of both pictures and words that pertained to the topic of plate tectonics. Children were allowed to take as long as they needed to read and study the instruction. The instructional material comprised 15 slides, each slide included a short instructional text and a static picture designed to illustrate the geological phenomena mentioned in the text[1]. An example of one of the slides is provided in Figure 1[2].

[1] Subjects received one of three versions of the pictures: 1) an abstract version that was devoid of color, 2) a relevant concrete version that consisted of colors for relevant concepts (pictured in Figure 1), and 3) a concrete version that consisted of colors for relevant concepts as well as other non-relevant pictures, such as airplanes or clouds surrounding the Earth. No differences were found in children's interview performance, gestures produced, or motivation produced as a function of the type of pictures they were instructed with (all *p*s >.10), therefore, we collapse students' performance across these groups.

[2] This diagram is intended to be schematic in nature and is therefore simplified such that it ignores the issue of scale and conveys only very basic concepts in plate tectonics (i.e., that plate movement – caused by heat in the Earth's interior – causes Earth's geological change). The design of the diagram was based on an informal review of graphics commonly used in elementary-school science textbooks.



Convection currents are the forces that cause the Earth's surface to move and change. Sections of the Earth's crust that rest on top of the mantle, called 'plates,' move with the convection currents below them, and are the cause of volcanoes, mountains, and earthquakes on Earth.

Fig. 1. An example slide from the instruction

The instruction covered three important boundary types: 1) oceanic – oceanic divergent boundaries where mid-oceanic ridges form, 2) continental – continental convergent boundaries where mountain ranges form, and 3) continental – oceanic convergent boundaries where volcanic mountain chains form.

After reading the instruction, children filled out a motivation questionnaire that consisted of six statements. Students were asked to rate, on a scale from 1 – 7, how much they agreed with each of the statements, with 7 meaning "strongly agree" and 1 meaning "strongly disagree". The statements pertained to the extent to which children considered plate tectonics to be 1) exciting, 2) fun, 3) important, 4) useful, 5) desirable to learn more about, and 6) desirable to take as a class at their school.

Interview Phase. During the interview phase, children were videotaped while they verbally answered questions from the experimenter about plate tectonics. Children were asked a total of ten questions in a fixed order. The first six questions pertained to concepts that children had learned about during the instruction (e.g., what causes the Earth's plates to move?). The final four questions consisted of showing children pictures of actual geological formations on Earth (e.g., the Himalayas). Then, children were provided a short description of the geological formation and were asked how they thought it formed (e.g., "*This is the Himalayan Mountain Range located in India,*" [Experimenter points to the field-photograph depicting the Himalayas] "*it is the tallest mountain range in the world. How do you think the Himalayan mountain range formed?*").

2.3 Scoring

To code for accuracy during the interview phase, an ideal answer was generated for each question and then broken down into individual knowledge components (henceforth referred to as "KC's"; see Koedinger, Corbett, & Perfetti, in press)[3]. For example, for the question "How do mountains form?" the associated knowledge components were 1) two continental plates, 2) collide, and 3) produced an upward force. The first and third authors coded a random selection of 25% of the videos for the presence of KC's in each child's responses. Overall, the raw inter-rater agreement was $r = .94$, kappa = .85. The first author then coded the remainder of children's responses. The score on the motivational questionnaire was the sum of the points for each question.

2.4 Gesture Coding

In order to analyze children's spontaneous use of gesture during the interview, we coded children's hand and arm movements into one of three categories: 1) KC-relevant gestures, 2) KC-irrelevant gestures, and 3) unrelated gestures. Both KC-relevant and KC-irrelevant gestures were "iconic" in that they referred to concrete entities (Roth & Lawless, 2002) in the domain of geoscience, where KC-relevant gestures pertained to geoscience phenomena that corresponded to a KC of a given question (e.g., a circular hand-motion to represent a convection current in response to the first question) and KC-irrelevant gestures pertained to concepts in geology, but did not correspond to any of the KC's of a given question (e.g., short, rapid movements of the hands to represent an earthquake). Unrelated gestures were either iconic gestures referring to concrete entities not related to geoscience content (e.g., a ship). The first and third authors coded a random selection of 25% of the videos for the presence of each type of gesture. On average, the raw inter-rater agreement was $r = .94$, kappa = .84. The first author then coded the remainder of the videos for the presence of each gesture type.

3 Results

3.1 Correlational Analyses

In total, we identified 270 KC-relevant gestures, 160 KC-irrelevant gestures, and 56 unrelated gestures. We first conducted correlations to see if children's age, gender, and motivation scores correlated with the proportion of KC-relevant gestures produced (i.e., relative to all gestures they produced) and the proportion of KC's children correctly identified during the interview (henceforth referred to as "interview accuracy"). There were no significant correlations between children's motivation scores, gender, interview accuracy, and proportion of KC-relevant gestures produced (all *p*s > .44). However, age was significantly correlated both with interview accuracy ($r = .453$, $p < .05$) and with the proportion of KC-relevant gestures produced ($r = .446$, $p < .05$). In order to control for children's age, motivation, and gender, partial correlations were conducted for all subsequent correlational analyses.

[3] Knowledge components are equivalent to concepts, principles, facts, or skills.

Of primary interest to us was whether the proportion of KC-relevant gestures that children spontaneously produced relative to all gestures produced would correlate with understanding of plate tectonics. Thus, we investigated the correlation on the proportion of KC-relevant gestures children produced by their interview accuracy, which revealed a significant, positive correlation ($r = .668$, $p < .001$) (see Figure 2). There was also a significant positive correlation between interview accuracy and the raw numbers of KC-relevant gestures children produced ($r = .575$, $p < .005$).

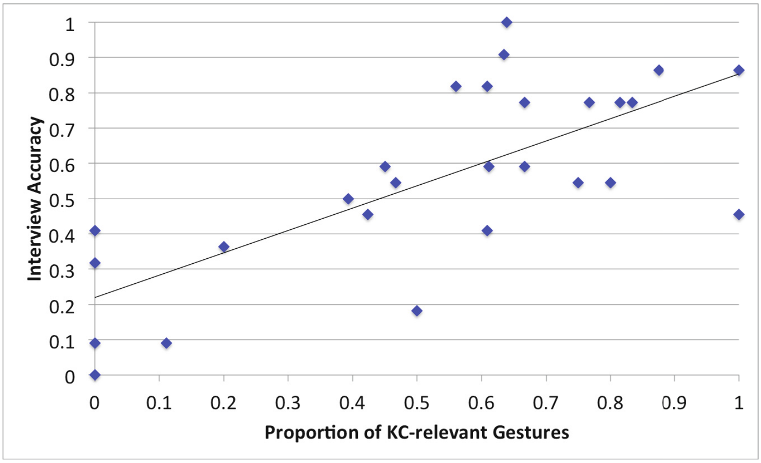


Fig. 2. Proportion of KC-relevant gestures produced as a function of interview accuracy

To examine whether other types of gestures correlated with geological understanding, we computed two more correlations, one on the proportion of KC-irrelevant gestures and interview accuracy, and another on the proportion of unrelated gestures and interview accuracy. These analyses revealed no significant relationship between the proportion of unrelated gestures and interview accuracy ($p > .52$). However, there was a significant, negative correlation between the proportion of KC-irrelevant gestures and interview accuracy ($r = -.663$, $p = .001$).

3.2 High- vs. Low-KC-Gesturers

To further explore the robustness of the relationship between KC-relevant gestures and geology understanding, we parsed children using a median split based on the proportion of KC-relevant gestures produced (*Med* = .61). This division created two groups: a “High-KC-gesturers” group and a “Low-KC-gesturers” group (High-KC-gesturers produced a significantly higher proportion of KC-relevant gestures $M = .77$, $SD = .13$ than Low-KC-gesturers $M = .33$, $SD = .23$; $t(24) = 5.87$, $p < .001$)[4]. There were no significant differences between High- and Low-KC-gesturers with regard to their motivation scores or gender (all *p*s > .24). There was a significant difference

[4] One child never gestured and therefore was not included in this analysis.

between the ages of each group, with the High-KC-gesturers (*M* = 12.78 yrs, *SD* = 1.89 yrs) slightly older on average than the Low-KC-gesturers (*M* = 10.97 yrs, *SD* = 2.4 yrs; *t*(24) = 2.08, *p* < .05). Importantly, High-KC-gesturers evidenced significantly higher interview accuracy (*M* = .73, *SD* = .17) than Low-KC-gesturers (*M* = .41, *SD* = .26; *t*(24) = 3.75, *p* < .001).

Do High-KC-gesturers simply gesture more often than Low-KC-gesturers? To directly explore this possibility, we conducted an independent samples t-test on the total number of gestures (i.e., KC-relevant, KC-irrelevant, and unrelated gestures) produced by both High- and Low-KC-gesture groups. This analysis revealed no differences between the groups (for the High group *M* = 18.54, for the Low group *M* = 18.08) *t*(24) = .08, *ns*. Additionally, to directly test whether there were differences in the types of gestures produced by High- and Low-KC-gesturers, we conducted a 2 (KC-gesturer: High vs Low) x 3 (gesture type: KC-relevant, KC-irrelevant, and unrelated) mixed ANOVA on the raw number of gestures produced, which revealed a significant effect of gesture type *F*(2,48) = 19.59, *p* < .001, qualified by a significant interaction *F*(2,48) = 8.11, *p* = .001 (see Figure 3). Post-hoc tests revealed that Low-KC-gesturers produced a significantly higher number of KC-irrelevant gestures (*M* = 8.31, *SD* = 7.51) than High-KC-gesturers (*M* = 3.23, *SD* = 3.42); *p* < .05), that High-KC-gesturers produced significantly more KC-relevant gestures than they did KC-irrelevant or unrelated gestures (all *p*s > .005), and that Low-KC gesturers produced significantly more KC-relevant and KC-irrelevant gestures than they did unrelated gestures (all *p*s < .01).

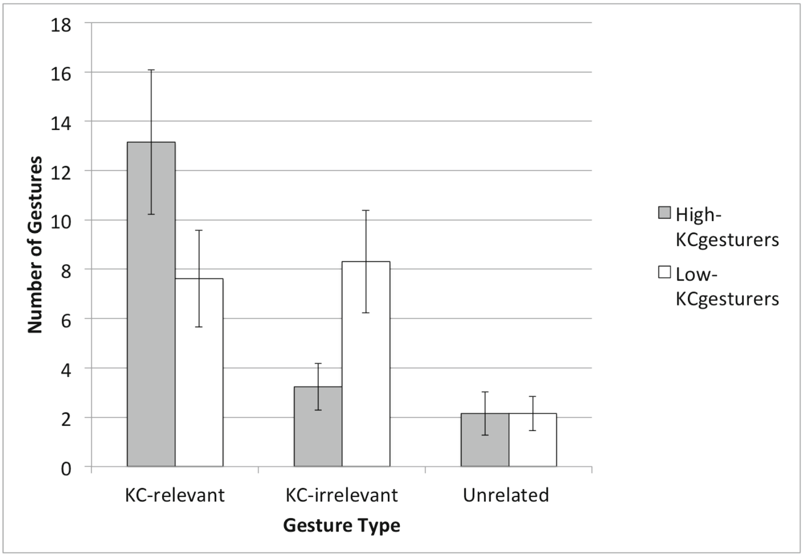


Fig. 3. Mean number of raw types of gestures produced as a function of whether children were categorized as High- or Low-KC-gesturers. Error bars represent standard errors of the mean.

4 Discussion

The primary aim of the present study was to determine whether children's gesturing was associated with their understanding of geoscience concepts. We found that students who produced a higher proportion of KC-relevant gestures were more likely to understand geoscience-related concepts, even when controlling for children's age, motivation, and gender. Moreover, both high- and low-KC-gesturers produced a similar amount of gestures overall, suggesting that it was not the *amount*, but rather, the *content* of children's gestures that predicted geoscience knowledge. This study is among the first to report a quantitative relationship between the frequency of children's gesturing and the understanding of a spatially demanding scientific concept. Our findings suggest that gesturing may even facilitate the process of learning science concepts, an insight that could have important implications for learning and instruction in science education.

However, due to the correlational nature of the present study, it is difficult to determine whether gestures caused or simply reflected geoscience understanding. Since in our task, children were asked to explain geoscience concepts to the experimenter, gesture may have assumed primarily a communicative role: those children who demonstrated better understanding of plate tectonics may have been better able to convey those concepts in gesture. Since a number of qualitative studies have shown that gesturing plays an important role in the acquisition of scientific concepts (e.g., Crowder, 1996; Roth, 2000), we surmise that children's gesturing may also have facilitated scientific understanding, but future research is needed to further examine this issue.

As the present study cannot tease apart the causal nature of gesturing and geoscience understanding, our future aim is to directly examine whether encouraging gesture causes increased geoscience understanding. Specifically, we are currently conducting a follow-up study in which we systematically compare the learning and transfer of children who are directly encouraged to gesture during the learning phase vs. those who are inhibited from gesturing. If gesturing does influence understanding, we would expect the gesture group to show stronger performance - as well as more frequent use of relevant gestures – on a post-test interview, similar to the one reported in this study.

In sum, though the present study is preliminary in nature. It is the first to our knowledge to document a quantitative relationship between gesturing and geoscience understanding in children. Although this relationship is correlational, these findings raise the possibility that incorporating and directly teaching gestures within the classroom will offer support for struggling students. At minimum, our results provide an empirical basis for the future investigation of this possibility.

Acknowledgements. This research was jointly supported by the Pittsburgh Science of Learning Center grant funded by the National Science Foundation award number SBE-0836012, the Spatial Intelligence and Learning Center grant funded by the National Science Foundation award number SBE-1041707, and by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education (# R305B040063). We thank Nora S. Newcombe, Kenneth R. Koedinger, Thomas F. Shipley, David Klahr, and Anna V. Fisher for their guidance and support throughout this project.

References

1. Ainsworth, S., Prain, V., Tytler, R.: Drawing to learn in science. Science 333, 1096–1097 (2011)
2. Alibali, M.W., Spencer, R.C., Knox, L., Kita, S.: Spontaneous gestures influence strategy choices in problem solving. Psychological Science 22, 1138–1144 (2011)
3. Alibali, M.W.: Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. Spatial Cognition and Computation 5, 307–331 (2005)
4. Beilock, S.L., Goldin-Meadow, S.: Gesture changes thought by grounding it in action. Psychological Science (in press)
5. Coleman, S.L., Gotch, A.J.: Spatial perception skills of chemistry students. Journal of Chemical Education 75, 206–209 (1998)
6. Cook, S.W., Goldin-Meadow, S.: The role of gesture in learning: Do children use their hands to change their minds? Journal of Cognition and Development 7, 211–232 (2006)
7. Crowder, E.M.: Gestures at work in sense-making science talk. The Journal of the Learning Sciences 5, 173–208 (1996)
8. de Ruiter, J.P.: Gesture and speech production. Ph.D. Dissertation, Nijmegen University (1998)
9. Ehrlich, S.B., Levine, S.C., Goldin-Meadow, S.: The importance of gesture in children's spatial reasoning. Developmental Psychology 42, 1259–1268 (2006)
10. Goldin-Meadow, S.: Beyond words: The importance of gesture to researchers and learners. Child Development 71, 231–239 (2000)
11. Gobert, J.D.: The effects of different learning tasks on model-building in plate tectonics: Diagramming versus explaining. Journal of Geoscience Education 53, 444–455 (2004)
12. Goodwin, C.: Environmentally coupled gestures. In: Duncan, S.D., Cassell, J., Levy, E.T. (eds.) Gesture and the Dynamic Dimension of Language, pp. 195–212. John Benjamins, Amsterdam (2007)
13. Hegarty, M., Crookes, R.D., Dara-Abrams, D., Shipley, T.F.: Do All Science Disciplines Rely on Spatial Abilities? Preliminary Evidence from Self-report Questionnaires. In: Hölscher, C., Shipley, T.F., Olivetti Belardinelli, M., Bateman, J.A., Newcombe, N.S. (eds.) Spatial Cognition VII. LNCS, vol. 6222, pp. 85–94. Springer, Heidelberg (2010)
14. Jee, B.D., Uttal, D.H., Gentner, D., Manduca, C., Shipley, T., Sageman, B., Ormand, C.J., Tikoff, B.: Analogical thinking in geoscience education. Journal of Geoscience Education 58, 2–13 (2010)
15. Kastens, K.A., Agrawal, S., Liben, L.S.: How students and field geologists reason in integrating spatial observations from outcrops to visualize a 3-D geological structure. International Journal of Science Education 31, 365–393 (2009)
16. Kastens, K.A., Agrawal, S., Liben, L.S.: Research methodologies in science education: The role of gestures in geoscience teaching and learning. Journal of Geoscience Education 56, 362–368 (2008)
17. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The knowledge-learning-instruction (KLI) framework: Toward bridging the science-practice chasm to enhance robust student learning. Cognitive Science (in press)
18. Kozhevnikov, M., Motes, M.A., Hegarty, M.: Spatial visualization in physics problem solving. Cognitive Science 31, 549–579 (2007)
19. Liben, L.S., Christensen, A.E., Kastens, K.A.: Gestures in Geology: The Roles of Spatial Skills, Expertise, and Communicative Context. In: Hölscher, C., Shipley, T.F., Olivetti Belardinelli, M., Bateman, J.A., Newcombe, N.S. (eds.) Spatial Cognition VII. LNCS, vol. 6222, pp. 95–111. Springer, Heidelberg (2010)

20. Liben, L.S., Kastens, K.A., Christensen, A.E.: Spatial foundations of science education: The illustrative case of instruction on introductory geological concepts. Cognition and Instruction 29, 45–87 (2011)
21. Matlen, B.J., Vosniadou, S., Jee, B.G., Ptouchkina, M.: Enhancing the comprehension of science text through visual analogies. In: Carlson, L., Holscher, C., Shipley, T. (eds.) Proceedings of the XXXIV Annual Meeting of the Cognitive Science Society (2011)
22. May, D.B., Hammer, D., Roy, P.: Children's analogical reasoning in a third-grade science discussion. Science Education 90, 316–330 (2006)
23. Morsella, E., Krauss, R.M.: The role of gestures in spatial working memory and speech. American Journal of Psychology 117, 411–424 (2004)
24. Nersessian, N.J.: How do engineering scientists think? Model-based simulation in biomedical engineering research laboratories. Topics in Cognitive Science 1, 730–757 (2009)
25. Novick, L.R.: The Importance of Both Diagrammatic Conventions and Domain-Specific Knowledge for Diagram Literacy in Science: The Hierarchy as an Illustrative Case. In: Barker-Plummer, D., Cox, R., Swoboda, N. (eds.) Diagrams 2006. LNCS (LNAI), vol. 4045, pp. 1–11. Springer, Heidelberg (2006)
26. Orion, N., Ben-Chaim, D., Kali, Y.: Relationship between earth-science education and spatial visualization. Journal of Geoscience Education 45, 129–132 (1997)
27. Pine, K.J., Lufkin, N., Messer, D.: More gestures than answers: Children learning about balance. Developmental Psychology 40, 1059–1067 (2004)
28. Ping, R., Ratliff, K., Hickey, E., Levine, S.C.: Using manual rotation and gesture to improve mental rotation in preschoolers. In: Carlson, L., Holscher, C., Shipley, T. (eds.) Proceedings of the XXXIV Annual Meeting of the Cognitive Science Society (2011)
29. Rauscher, F.H., Krauss, R.M., Chen, Y.: Gesture, speech, and lexical access: The role of lexical movements in speech production. Psychological Science, 7, 226–231 (1996)
30. Resnick, I., Atit, K., Göksun, T., Shipley, T.: Experts' and novices' use of gesture in explaining geologic maps. Poster Presented at The Annual Meeting of the Cognitive Science Society 2011, Boston, MA (July 21, 2011)
31. Rimè, B., Shiaratura, L., Hupet, M., Ghysselinckx, A.: Effects of relative immobilization on the speaker's nonverbal behavior and on the dialogue imagery level. Motivation and Emotion 8, 311–332 (1984)
32. Roth, W.-M.: From gesture to scientific language. Journal of Pragmatics 32, 1683–1714 (2000)
33. Roth, W.-M., Lawless, D.: Scientific investigations, metaphorical gestures, and the emergence of abstract scientific concepts. Learning and Instruction 12, 285–304 (2002)
34. Sauter, A., Uttal, D., Alman, A.S., Goldin-Meadow, S., Levine, S.C.: Learning what children know about space from looking at their hands: The added value of gesture in spatial communication. Journal of Experimental Child Psychology (in press)
35. Singer, M., Radinsky, J., Goldman, S.R.: The role of gesture in meaning construction. Discourse Processes 45, 365–386 (2008)
36. Wesp, R., Hess, J., Keutmann, D., Wheaton, K.: Gestures maintain spatial imagery. The American Journal of Psychology 114, 591–600 (2001)

Integration of Spatial Relations across Perceptual Experiences

Marios N. Avraamides[1], Christina Adamou[1], Alexia Galati[1], and Jonathan W. Kelly[2]

[1] Department of Psychology, University of Cyprus
P.O. Box 20537, 1678 Nicosia, Cyprus
[2] Department of Psychology, Iowa State University
W112 Lagomarcino Hall, Ames, IA 50010
{mariosav,adamou.d.christina,galati}@ucy.ac.cy,
jonkelly@iastate.edu

Abstract. People often carry out tasks that entail coordinating spatial information encoded in temporally and/or spatially distinct perceptual experiences. Much research has been conducted to determine whether such spatial information is integrated into a single spatial representation or whether it is kept in separate representations that can be related at the time of retrieval. Here, we review the existing literature on the integration of spatial information and present results from a new experiment aimed at examining whether locations encoded from different perspectives in the same physical environments are integrated into a single spatial representation. Overall, our findings, coupled with those from other studies, suggest that separate spatial representations are maintained in memory.

Keywords: Integration of spatial information, Reference frames, Spatial memory organization, Perspective taking.

1 Introduction

Much of our everyday activity relies on retrieving spatial information from memory. For example, when planning a route prior to navigating a familiar environment we typically consider where the goal location is relative to our starting point, but also how landmarks along the route relate spatially to each other. Also, during navigation we must monitor our orientation by determining where we are in relation to immediate and distal landmarks. To carry out such tasks effectively, we must construct accurate spatial representations when we experience the space (e.g., when navigating a city for the first time) and maintain those representations in memory. Whereas in some cases people construct spatial representations by experiencing multiple locations simultaneously or near simultaneously from a fixed standpoint (e.g., when looking at a small room from its entrance, or when inspecting a table-top arrangement of objects from a specific direction), in other cases, they do so by experiencing locations at different times and typically from different standpoints (e.g., viewing objects by moving within a multi-room house).

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 416–430, 2012.

Previous research has established that people can successfully compute the relations between locations acquired through distinct experiences (e.g., they can point to unseen distal landmarks with above chance accuracy). However, what is not yet clear is whether people integrate spatial information experienced at different points in time into a single spatial representation or maintain it in distinct representations that can be related at task execution. In the present paper we review literature that can shed light on this issue. In section 1, we provide a brief introduction on how people organize in spatial memory information experienced simultaneously (or near-simultaneously). This introduction highlights that objects that can be viewed at once, typically as part of a layout that is external to the observer, are stored in a single representation maintained in memory from a preferred direction. Next, in section 2, we present evidence that locations encoded sequentially within the same spatial environment are also remembered from a preferred direction, which suggests that they are integrated into a single representation. Finally, in subsequent sections we discuss the results from studies that have examined, using different paradigms, spatial memory for temporally and/or spatially separated layouts. Findings from these studies generally suggest that spatial information is not integrated into a single representation, although often experiencing one layout may influence the way objects in subsequent layouts are encoded.

2 Memory for Locations Viewed Simultaneously

Mounting evidence suggests that locations in spatial layouts are encoded on the basis of allocentric reference frames that are maintained in memory in a preferred direction (McNamara, 2003; Mou & McNamara, 2002). Such evidence comes from studies that examine the organizational structure of spatial memories by having participants study a layout from an external standpoint and then, in a different laboratory room, make Judgments of Relative Direction (JRD); that is, respond to statements of the form “Imagine standing at x facing y, point to z”, where x, y, and z are objects from the memorized layout. These studies generally show that pointing performance is faster and/or more accurate from one or more imagined perspectives. This is typically interpreted as evidence that during learning, participants created a spatial representation that was stored in memory from a particular orientation, axis, or set of axes (see McNamara, 2003 for a review).

Many studies in this area have focused on identifying the factors that determine the preferred direction(s) from which spatial memories are maintained. For example, environmental cues play an important role in selecting a preferred direction in memory (Shelton & McNamara, 2001). In this study, participants studied a layout of 7 objects placed on a square mat within a rectangular room. In one experiment the edges of the mat were aligned with the walls of the room and participants viewed the objects from two standpoints: one that was aligned with the mat and the walls of the room (0°) and one that was not (135°). Viewing order was counterbalanced across participants. Subsequent JRD testing revealed that, for both viewing orders, performance was better when responding from the aligned 0° than from the misaligned 135° perspective, which was no better than the remaining non-experienced perspectives. That participants used the array's alignment with respect to the room over their own misaligned orientation as an organizing axis, highlights how powerful environmental cues are when selecting a preferred direction in memory. Similarly, other studies have provided evidence that other cues available during learning may determine the preferred direction. These cues include the presence of an axis of bilateral symmetry (Mou, Zhao, & McNamara, 2007), instructions (Greenauer & Waller, 2008), and other things being equal, egocentric experience (Shelton & McNamara, 1997).

In summary, research with scenes external to the observer indicates that spatial information is maintained in memory from a preferred direction that is selected during learning based on available cues. This suggests that if locations encoded in distinct experiences are integrated into a single spatial representation, this representation should have a preferred direction observable in subsequent testing. In the next section we discuss whether locations encoded sequentially are indeed maintained in a single representation.

3 Memory for Locations Viewed Sequentially

In contrast to studies with table-top displays and other layouts that can be viewed at once, studies have also examined spatial memories for room-size environments in which not all objects can be viewed simultaneously. A typical set-up involves objects that are placed around the observer at different angles (e.g., Hodgson & Waller, 2006; Kelly, Avraamides, & Loomis, 2007). Being positioned within or internal to the layout, the observer must thus move her head or body in order to inspect all locations. Despite this additional requirement, which results in processing locations sequentially, to the best of our knowledge, no study has reported any differences in the organizational structure of memories in which the observer's position is interval vs. external to the layout.

In fact, people's memories for scenes viewed sequentially, while being internal to the scene, seem to be organized around the same principles as their memories for table-top scenes and scenes viewed at once. For example, Kelly et al. (2007) extended findings regarding the organization of memories for externally viewed scenes to scenes viewed internally, within a virtual-reality environment. In this study, participants learned the locations of 8 objects placed in the corners of an octagonal virtual room. All participants began inspecting objects from the same orientation but were then allowed to freely rotate and study the layout from any orientation and for as long as they wanted. Due to the narrow horizontal field of view of the Head-Mounted-Display (HMD) that was used, no more than 2 objects could be viewed simultaneously from any orientation, ensuring their sequential viewing. Following learning, participants were tested using JRD while standing in either the same room in which learning took place or in a different room, and while assuming an orientation that was offset by 90° to the left or right of the learning orientation. Participants tested in a different room exhibited superior performance when pointing to objects from an imagined orientation that was aligned with the initial orientation they had during learning. This result replicates previous findings from studies with external scenes showing that in the absence of other cues or instructions, egocentric experience is used to determine a preferred direction in memory (Shelton & McNamara, 1997). In addition, those tested in the same room also did well when they responded from an orientation that was aligned to their actual testing orientation. This advantage for people's actual orientation at the time of retrieval has also been documented with external layouts (Mou, McNamara, Valiquette, & Rump, 2004).

The parallel results from studies with internal vs. external layouts suggest that people can easily integrate into a single spatial representation locations that are encoded sequentially. This conclusion is further corroborated by the findings of a study that manipulated the temporal presentation of to-be-learned targets. Avraamides, Loomis, Klatzky, and Golledge (2004) asked participants to indicate the relative direction and distance between pairs of objects that they had previously encoded by vision or spatial language. For our purposes, the experiments where objects were encoded through vision are pertinent. In one experiment, four visual targets were presented simultaneously in the frontal visual field of participants, whereas in a second experiment, the same targets were presented sequentially and in isolation (i.e., the previous object was removed before a new one was placed). Regardless of how participants had encoded visual targets across the two experiments, both their response latency and the standard deviation of their signed pointing errors were equivalent.

Overall, results from studies in which the encoding of spatial locations occurs incrementally within the boundaries of the same physical space suggest that people have no difficulty integrating information within a single spatial representation. In the next section, we discuss studies that have examined whether people integrate spatial locations encoded with greater temporal separation.

4 Integration of Layouts with Extended Temporal Separation

As our review so far suggests, when people encode spatial locations sequentially (typically as observers internal to the scene) they easily integrate into a single representation. However, it is unclear whether they also do so when the temporal separation between locations is greater than the time needed to turn their head to view an object. Do they keep these locations in distinct representations or do they integrate them in a single representation?

Studies with large-scale environments provide converging evidence that these environments are also represented in memory from preferred directions. For example, Werner and Schmidt (1999) showed that people in Göttingen, Germany pointed faster and more accurately to landmarks in their city when imagining themselves at orientations that were aligned than misaligned with the two streets of a main intersection. This suggests that their spatial memory was maintained from preferred directions that were determined by the structure of the environment. Findings from McNamara, Rump, and Werner (2002) corroborate further this conclusion. In this study participants navigated a park following one of two paths. One path was aligned with the intrinsic axes of a salient landmark (i.e., a Parthenon replica) and the other one was misaligned. Then, while in the lab, they pointed towards various park objects from imagined perspectives. Results revealed that participants walking the aligned path pointed more accurately from (1) perspectives aligned with the legs of the path and the intrinsic axes of the Parthenon, and (2) a perspective oriented towards a second salient landmark (i.e., a lake). Those walking the misaligned path pointed more accurately from the perspective oriented towards the lake, with accuracy for the remaining perspectives decreasing with increasing angular disparity from that perspective. Thus, in both cases participants organized their memories on the basis of a reference frame that was intrinsic to the layout, with the preferred direction being influenced by the alignment of the path.

At this point, it should be noted that a small number of studies have provided evidence that spatial reasoning about large-scale or even room-size navigable environments is orientation-independent (e.g., Evans & Pezdek, 1980; Presson, DeLange, & Hazelrigg, 1987; 1989). For example, in one experiment, Evans and Pezdek had college students judge the depicted spatial relations of triads containing either campus landmarks or American States, with the triads shown at various orientations. When participants judged relations among triads of States, their response latency increased linearly as a function of the angular deviation of the triad from the upright orientation typically shown in a map. However, when judging campus landmarks, which were presumably encoded in memory through active exploration as opposed to observing a map, this wasn't the case: participants judged the campus triads equally fast from every presented orientation. Although this finding may suggest that the spatial representation containing campus landmarks was orientation-free, an alternative possibility is that students had experienced campus landmarks from various orientations and constructed a representation with multiple preferred directions. But there is evidence that even unfamiliar environments may be represented in orientation-independent representations, as suggested by participants' memory performance after walking long paths in the laboratory (Presson, DeLange, & Hazelrigg, 1987; 1989). However, subsequent studies have failed to replicate such orientation-independence for unfamiliar environments, suggesting that they may be limited to the specific testing situations employed by Presson and colleagues (see Roskos-Ewoldsen, McNamara, Shelton, & Carr, 1998; Waller, Montello, Richardson, & Hegarty, 2002). In general, the majority of studies on both outdoor and indoor navigable environments suggest that the constructed memories for these environments are orientation-dependent, just like the memories for locations that are experienced either simultaneously or in close temporal proximity.

A number of studies in which participants experience locations with extended temporal separation allow for further insight into whether the resulting representations are orientation-dependent and whether they involve the integration of locations. In the following subsections we review relevant findings from such studies using different paradigms: studies examining memories for nested environments, studies assessing integration by comparing responses for within- and between-layout judgments, and studies using the *transfer of reference frames* paradigm.

4.1 Integration Assessed across Nested Environments

One approach to examining whether people integrate locations they have encoded with temporal separation is to evaluate their representations of locations in nested environments. Wang and Brockmole (2003a, 2003b) did precisely that by having people reason both about a newly learned local environment (e.g., the research laboratory) and a familiar, large-scale environment in which the new environment was nested (e.g., the campus in which the laboratory is located).

In one study (Wang & Brockmole, 2003a), they examined whether participants arriving at the laboratory would integrate new knowledge about the locations of laboratory objects into their existing spatial representation of the campus. In a first experiment, participants were exposed to a number of laboratory objects and following a brief rotation, they pointed towards both laboratory objects and campus landmarks. Participants made larger configuration errors[1] when pointing to campus landmarks than laboratory objects, suggesting that they held objects from the two environments in distinct representations. Moreover, while heading error [2] was uniformly distributed for laboratory objects, it was randomly distributed for campus landmarks. This suggests that participants remained oriented within the laboratory but failed to relate their orientation to the more distal campus landmarks. In a follow-up experiment, participants walked a route from the laboratory to the campus and back, and pointed towards objects and landmarks along the way. Participants could point correctly to the direction of a campus landmark only when they exited the room. Conversely, once they were on campus grounds they lost track of their orientation relative to the room layout. These findings indicate the newly acquired spatial knowledge for a local layout is not readily integrated during learning into an existing representation of a larger scale environment. Rather, separate representations are maintained. Although this is consistent with accounts of hierarchical representations of space (Hirtle & Jonides, 1985), it is evident that in this study participants did not represent in memory the directional relation between the two spatial representations.

In another study, Wang and Brockmole (2003b) investigated whether people automatically update spatial relations in one environment when rotating with respect to the objects of another environment. Participants, sitting on a swivel chair, learned the locations of laboratory objects and brought to mind the locations of familiar campus landmarks. Then they were asked to physically rotate to various orientations relative to either laboratory objects or campus landmarks, depending on the condition. At the end of a series of rotations they pointed to both laboratory objects and campus landmarks. When participants turned relative to laboratory objects, they were faster to point to these objects than to campus landmarks. In contrast, when they rotated relative to campus landmarks they were equally fast at pointing to campus landmarks and laboratory objects. These findings indicate that participants held laboratory

[1] Configuration Error is the standard deviation of the signed pointing errors. It is a measure of the internal consistency of the spatial representation, i.e., how accurate an object is localized relative to the other objects.

[2] Heading Error is the average of the signed pointing errors. Its value is close to 0° when participants are oriented but it is randomly distributed when they are disoriented.

objects in a distinct representation that was automatically updated with rotational movement. This is in line with arguments that spatial updating is limited to objects in one's immediate surroundings (Wang & Spelke, 2000) that are maintained in a transient sensorimotor representation (Avraamides & Kelly, 2008; Mou et al., 2004; Waller & Hodgson, 2006).

Overall, the results from studies with nested environments suggest that people keep spatial information for each environment in separate representations. Relating information across representations, which presumably takes place when required by the task, may take place but at a considerable performance cost.

4.2 Integration Assessed in between vs. within-Layout Judgments

Another approach to examining whether people integrate locations from multiple layouts that they have encoded with temporal separation is to examine their judgments for spatial relations within the same layout vs. between layouts (e.g., Giudice, Klatzky, & Loomis, 2009; Ishikawa & Montello, 2006). If information in the two layouts is integrated into a single spatial representation at encoding then no performance differences are expected when comparing within- and between-layout judgments.

This paradigm has been used to investigate different types of layouts, from large scale environments in studies investigating navigation (Moar & Carleton, 1982; Montello & Pick, 1993; Golledge, Ruggles, Pellegrino & Gale, 1993; Ishikawa et al., 2006) to table-top scenes (Giudice et al., 2009). However, findings from these studies are contradictory.

A number of studies indicate that even though participants are able to relate information derived from separate experiences by performing well above chance, they do better at within- than between-layout trials. For example, Montello and Pick (1993) had participants walk two routes within a building. Participants learned the two routes separately but they were then either verbally informed about the connection between the two routes, or experienced the relationship through navigation. In subsequent testing, participants pointed to non-visible landmarks on the two routes while walking in one of them. Although participants could point to objects in both routes with above chance accuracy, they did better when pointing to objects from the route they were travelling on than the other one. This pattern of results suggested that people did not integrate the spatial information from the two routes into a single representation. Instead, they computed intra-route information at the time of responding. This was also the case in a study by Ishikawa et al., (2006), where participants learned two separate routes by being driven along each route ten times. After the first three experiences a connecting path between the two routes was experienced in the learning routine. At the end of each learning experience participants had to estimate the direction and route distance between four landmarks that were previously experienced along the routes. After the fourth session, participants provided straight line distance and direction estimates between landmarks within and across routes. Participants' direction estimates across routes were above chance performance, although distance estimates did not differ significantly from guessing. Moreover, as participants gained more experience with the routes their performance improved. Thus, participants were able to relate to some degree the spatial information from the two layouts. But even so, the performance for between route landmarks was worse than within route landmarks suggesting that the two routes were not integrated into a single spatial representation in memory.

On the other hand, other studies have provided results compatible with integrating information from separate layouts into a single spatial representation (Holding & Holding, 1989; Moar & Carleton, 1982). For example, Moar and Carleton (1982) investigated whether people would integrate information from separately-learned, intersecting routes. Navigation was simulated by showing participants photographs taken from the routes. Participants then had to estimate the distance and direction between two locations presented in two slides on a screen that were either from the same route or from different routes. Participants performed comparably when providing estimates for within-route and cross-route pairs of places, suggesting that they were able to integrate the two routes into a single spatial representation.

A complicating factor from interpreting findings from studies comparing between-layout and within-layout judgments to determine whether single or distinct representations are maintained is that previous research has indicated that spatial performance is influenced by the temporal and spatial separation of locations (McNamara, Halpin, & Hardy, 1992). Studies using within vs. between-layout judgments typically involve learning layouts of objects separately in time (e.g., Ishikawa et al., 2006; but see Greenauer & Waller, 2010). Also, although some studies have controlled for spatial separation (e.g., Giudice et al., 2009; Montello & Pick, 1993), others have not (e.g., Greenauer & Waller, 2010). Furthermore, a study by Greenauer and Waller (2010) demonstrated that despite comparable performance for within- and between-layout judgments, layouts were maintained in distinct representations. In this study, participants studied objects placed in the center of a room forming two adjacent arrays. The two arrays were viewed simultaneously (Exp.1, 2, and 4) or sequentially (Exp.3) and were distinguished by colored disks. Importantly, each array had its own axis of bilateral symmetry which was misaligned with the learning view of the observer, and participants were instructed to learn the layouts along their symmetry axes. JRD responses for within-layout trials indicated that each array was maintained in memory from a different preferred direction that was determined by its axis of bilateral symmetry. In contrast, between-layout responses were facilitated along the direction determined by the learning view of the participant. Thus, despite the similar performance in overall accuracy and latency in within- and between-layout trials, these findings suggest that, in line with theories of hierarchical encoding (Hirtle & Jonides, 1985), the two layouts were organized around distinct microreference frames whose relation was specified by a more global macroreference frame.

One paradigm that controls for the spatial separation of layouts is the transfer of reference frames, which we present in the next section.

4.3 The Transfer of Reference Frames across Layouts

Kelly and McNamara (2010) developed a new method to study how people organize in memory distinct layouts that are learned in sequence (see also Kelly, Avraamides, & McNamara, 2010). This method examines whether the reference frame that is used to organize the first studied layout is transferred to the encoding of the second layout. It should be noted that finding that a common reference frame is used to encode two layouts does not necessarily mean that the layouts have been integrated into a single representation. However, the opposite finding -- i.e., that each layout is associated with a distinct reference frame (e.g., Greenauer & Waller, 2010)-- is hard to accommodate with a single representation account.

In one study Kelly and McNamara (2010) had participants study an external layout of 7 objects from one of two perspectives (0° or 135°). Following learning, 7 new objects were added to the scene which participants studied from a fixed perspective (135°). Testing with JRD involving only within-layout locations revealed that performance for the second layout was facilitated for imagined perspectives aligned with the study viewpoint of the first layout (0° or 135° depending on condition). According to the authors, participants established a reference frame from the study viewpoint of the first layout and subsequently used it to encode the locations of the second. This is congruent with findings that reference frames established through vision can be later used to encode haptic locations (Kelly & Avraamides, 2011) and vice-versa (Kelly, Avraamides, & Giudice, 2011).

The studies on the transfer of reference frames control for spatial separation of objects by using overlapping layouts. However, as their primary goal was to assess the reference frames used for encoding each layout, they have not included any between-layout judgments. Thus, although their findings are compatible with a single-representation for distinct layouts, they are not conclusive. In the next section we present results from a new experiment aimed at assessing the transfer of reference frames for layouts learned from different perspectives, while also using both within- and between-layout judgments.

4.4 Integration of Layouts Encoded from Distinct Perspectives

We have conducted an experiment to examine whether spatial locations in the same physical space but experienced as separate layouts from different perspectives are integrated into a single representation (Adamou, 2011). Meilinger, Berthoz, and Wiener (2011) have also examined the integration of spatial information that was viewed from different perspectives. In their study, participants learned two spatial layouts each containing 3 locations by either viewing both layouts from the same standpoint, or by viewing the second layout upon walking to a different standpoint that was offset by 90° from the first. Following learning, participants in both conditions were instructed to walk the shortest path that linked the 6 locations. Participants were capable of relating spatial information across experiences in order to compute a path: path-planning performance in both conditions was only 3.7% longer than the shortest possible path. Additionally, participants made more errors when walking to targets of the first layout. The authors interpreted this finding as evidence that the locations of the first layout were transformed to the reference frame of the second.

The path-walking task used by Meilinger et al. (2011) relies strongly on participants' actual orientation and may have encouraged participants to update spatial information during movement from the first standpoint to the other. Unlike Meilinger et al., we used JRD, an off-line test of spatial memory that is less dependent on the observer's actual orientation (see Avraamides & Kelly, 2008 for a discussion). Additionally, in contrast to previous studies on the transfer of reference frames (Kelly et al., 2011, 2010), in this study participants were internal to the layout. Participants first learned a layout of four objects from one orientation (0°) while standing in the center of a featureless round room in virtual reality (Fig.1). Once they memorized the layout, the objects were removed and participants rotated to the left to study a second layout from a different orientation (210°). Following learning, participants moved to a different laboratory and carried out a series of JRD trials (i.e., "imagine facing x, point to y") using a computerized pointer. Following the logic of previous studies (e.g., Ishikawa & Montello, 2006), trials involved pairs of objects from the same or different layouts.

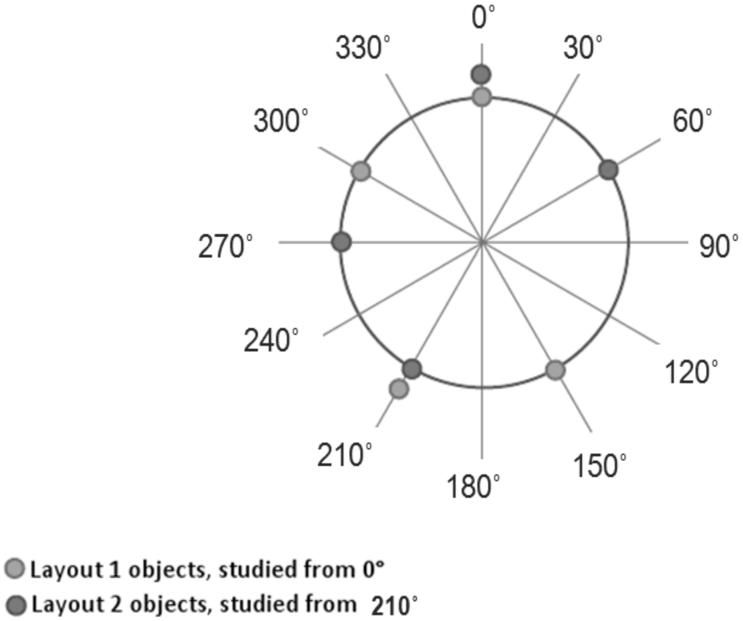

0°
330°
30°
300°
60°
270°
90°
240°
120°
210°
180°
150°
● Layout 1 objects, studied from 0°
● Layout 2 objects, studied from 210°

Fig. 1. A schematic illustration of the two spatial layouts. Participants studied one layout while facing 0° and another while facing 210°. The assignment of layout identity (1 or 2) to learning order (first or second) was counterbalanced across participants. The study viewpoint was always 0° for the layout studied first and 210° for the layout studied second.

Results indicated that performance was above chance in between-layout trials confirming that participants were able to use information from the different layouts to carry out the task. However, they were considerably slower to respond in between-layout than within-layout trials suggesting that the two layouts were not integrated into a single spatial representation. Compatible with this conclusion were the analyses of latency to respond from imagined perspectives: for within-layout trials participants

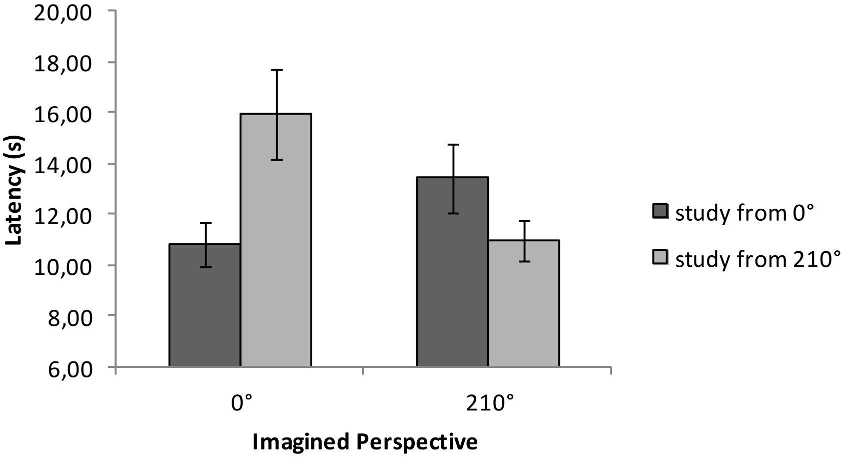


Fig. 2. Latency as a function of imagined perspective in within-layout trials. Only performance for the two viewpoints that were common in the two layouts is shown. Separate statistical analyses considering all imagined perspectives confirmed the presence of a preferred direction aligned with the study viewpoint of each layout.

were faster to respond from the perspective that was aligned with the study viewpoint of the layout (0° or 210° depending on learning order) than the other imagined perspectives (Fig. 2).

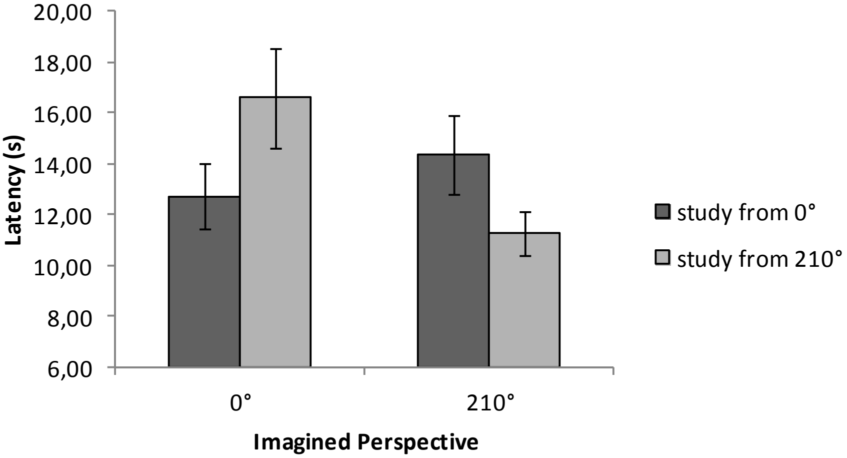


Fig. 3. Latency as a function of imagined perspective and the study viewpoint of the layout from which the orienting object was sampled in between-layout trials

In between-layout trials, participants were faster to respond from imagined perspectives aligned with the study viewpoint of the layout from which the orienting object (i.e., the one defining the imagined perspective) came (Fig. 3).

The findings from this study indicate that when participants learn separate layouts of objects from different viewpoints, they keep these layouts in distinct spatial representations each with its preferred direction, even when the objects are dispersed within the boundaries of the same physical environment. This is compatible with the claims of Greenauer and Waller (2010) that layouts are maintained in distinct microreference frames. The transfer of reference frame observed in previous studies (e.g., Kelly & McNamara, 2010) did not take place here, but numerous methodological differences could account for this failure to replicate. For example, the first layout was occluded when participants studied the second layout, and the circular room provided no global orientation cues. Further research is needed to determine the necessary conditions for reference frame transfer.

5 Summary and Conclusions

The findings from the studies we have reviewed here suggest that spatial locations encoded in memory as part of distinct perceptual experiences are kept in separate spatial representations. Each spatial representation can be organized around a different preferred direction on the basis of cues that are available during learning (Mou & McNamara, 2002). People are flexible at picking up cues to establish a preferred direction and often transfer such cues from one experience to another (Kelly & Avraamides, 2011). Furthermore, the spatial relation between two or more spatial representations is sometimes directly represented in memory (Greenauer & Waller, 2010) and sometimes not (Wang & Brockmole, 2003a).

Representing objects in memory in small clusters may be beneficial for everyday tasks that typically rely on processing only a small number of locations at a time. For example, on-line tasks such orienting ourselves in the local environment entail processing only a small number of immediate locations. Similarly, off-line tasks such as describing the layout of our house to a colleague who visits our office requires considering only the locations of the distal household objects and ignoring any objects in the immediate environment of the office. Clustering objects into small meaningful representations may thus allow us to activate only the spatial information that is needed at a given moment enabling us to operate within the capacity limits of working memory. Locations at our home can thus be clustered into smaller meaningful units (e.g., defined by rooms) that can be managed more easily during retrieval.

Although people may represent spatial information in appropriate, distinct representations, they are typically efficient at relating spatial information across representations. In the studies we have reviewed, participants could localize objects between layouts well above chance. Thus, when a task requires coordinating information across perceptual experiences, people seem capable of doing so at the time the information is needed (Meilinger et al., 2011). Although a performance cost is sometimes observed, this cost is by no means dramatic. In fact this cost may be modulated by a number of factors, such as the requirements of the task at hand. Different tasks place different demands on encoding and maintaining spatial relations among locations learned with extended spatial and temporal separation. For example, memorizing a route that one is travelling requires relating spatially the locations encountered on the route (e.g., landmarks, decision-points, etc) and storing these relations in memory. Indeed, neuroimaging studies employing subliminal priming methods have provided evidence that people encode and maintain in memory functional links between locations experienced along travelled routes (e.g., Janzen & Westeijn, 2007; Shinazi & Epstein, 2010). Thus, although people may normally default to maintaining separate representations for spatial representations derived from distinct perceptual experiences, they can integrate information into a single representation either at encoding or at a later stage if the task requires them to do so.

Acknowledgments. Support was provided by research grant OSSMA-206912 from the European Research Council to M.N.A.

References

1. Adamou, C.: Integration of visuo-spatial information encoded from different modalities. Unpublished MSc Thesis. University of Cyprus (2011)
2. Avraamides, M.N., Kelly, J.W.: Multiple systems of spatial memory and action. Cognitive Processing 9, 93–106 (2008)
3. Avraamides, M.N., Loomis, J.M., Klatzky, R.L., Golledge, R.G.: Functional equivalence of spatial representations derived from vision and language: evidence from allocentric judgments. Journal of Experimental Psychology. Learning, Memory, and Cognition 30, 804–814 (2004)
4. Giudice, N.A., Klatzky, R.L., Loomis, J.M.: Evidence for amodal representations after bimodal learning: integration of haptic-visual layouts into a common spatial image. Spatial Cognition & Computation 9, 287–304 (2009)
5. Golledge, R.G., Ruggles, A.J., Pellegrino, J.W., Gale, N.D.: Integrating route knowledge in an unfamiliar neighborhood: Along and across route experiments. Journal of Environmental Psychology 13, 293–307 (1993)
6. Greenauer, N., Waller, D.: Intrinsic array structure is neither necessary nor sufficient for nonegocentric coding of spatial layouts. Psychonomic Bulletin & Review 15, 1015–1021 (2008)
7. Greenauer, N., Waller, D.: Micro- and macroreference frames: Specifying the relations between spatial categories in memory. Journal of Experimental Psychology. Learning, Memory, and Cognition 36, 938–957 (2010)
8. Hirtle, S.C., Jonides, J.: Evidence of hierarchies in cognitive maps. Memory & Cognition 13, 208–217 (1985)
9. Hodgson, E., Waller, D.: Lack of set size effects in spatial updating: Evidence for offline updating. Journal of Experimental Psychology. Learning, Memory, and Cognition 32, 854–866 (2006)
10. Holding, C.S., Holding, D.H.: Acquisition of route network knowledge by males and females. Journal of General Psychology 116, 29–41 (1989)

11. Ishikawa, T., Montello, D.R.: Spatial knowledge acquisition from direct experience in the environment: individual differences in the development of metric knowledge and the integration of separately learned places. Cognitive Psychology 52, 93–129 (2006)
12. Janzen, G., Weststeijn, C.G.: Neural representation of object location and route direction: an event-related fMRI study. Brain Research 1165, 116–125 (2007)
13. Kelly, J.W., Avraamides, M.N., Loomis, J.M.: Sensorimotor alignment effects in the learning environment and in novel environments. Journal of Experimental Psychology. Learning, Memory, and Cognition 33, 1092–1107 (2007)
14. Kelly, J.W., Avraamides, M.N.: Cross-sensory transfer of reference frames in spatial memory. Cognition 118, 444–450 (2011)
15. Kelly, J.W., McNamara, T.P.: Reference frames during the acquisition and development of spatial memories. Cognition 116, 409–420 (2010)
16. Kelly, J.W., Avraamides, M.N., Giudice, N.A.: Haptic experiences influence visually acquired memories: reference frames during multimodal spatial learning. Psychonomic Bulletin & Review 18, 1119–1125 (2011)
17. Kelly, J.W., Avraamides, M.N., McNamara, T.P.: Reference Frames Influence Spatial Memory Development within and Across Sensory Modalities. In: Hölscher, C., Shipley, T.F., Olivetti Belardinelli, M., Bateman, J.A., Newcombe, N.S. (eds.) Spatial Cognition VII. LNCS (LNAI), vol. 6222, pp. 222–233. Springer, Heidelberg (2010)
18. McNamara, T.P., Halpin, J.A., Hardy, J.K.: Spatial and temporal contributions to the structure of spatial memory. Journal of Experimental Psychology. Learning, Memory, and Cognition 18, 555–564 (1992)
19. McNamara, T.P.: How Are the Locations of Objects in the Environment Represented in Memory? In: Freksa, C., Brauer, W., Habel, C., Wender, K.F. (eds.) Spatial Cognition III. LNCS (LNAI), vol. 2685, pp. 174–191. Springer, Heidelberg (2003)
20. McNamara, T.P., Rump, B., Werner, S.: Egocentric and geocentric frames of reference in memory of large-scale space. Psychonomic Bulletin & Review 10, 589–595 (2003)
21. Meilinger, T., Berthoz, A., Wiener, J.M.: The integration of spatial information across different viewpoints. Memory & Cognition 39, 1042–1054 (2011)
22. Mou, W., McNamara, T.P.: Intrinsic frames of reference in spatial memory. Journal of Experimental Psychology. Learning, Memory, and Cognition 28, 162–170 (2002)
23. Moar, I., Carleton, L.R.: Memory for routes. Quarterly Journal of Experimental Psychology 34, 381–394 (1982)
24. Montello, D.R., Pick, H.L.: Integrating Knowledge of Vertically Aligned Large-Scale Spaces. Environment and Behavior 25, 457–484 (1993)
25. Mou, W., McNamara, T.P., Valiquette, C.M., Rump, B.: Allocentric and egocentric updating of spatial memories. Journal of Experimental Psychology. Learning, Memory, and Cognition 30, 142–157 (2004)
26. Mou, W., Zhao, M., McNamara, T.P.: Layout geometry in the selection of intrinsic frames of reference from multiple viewpoints. Journal of Experimental Psychology. Learning, Memory, and Cognition 33, 145–154 (2007)
27. Presson, C.C., DeLange, N., Hazelrigg, M.D.: Orientation-specificity in kinesthetic spatial learning: the role of multiple orientations. Memory & Cognition 15, 225–229 (1987)
28. Presson, C.C., DeLange, N., Hazelrigg, M.D.: Orientation specificity in spatial memory: what makes a path different from a map of the path? Journal of Experimental Psychology. Learning, Memory, and Cognition 15, 887–897 (1989)
29. Roskos-Ewoldsen, B., McNamara, T.P., Shelton, A.L., Carr, W.: Mental representations of large and small spatial layouts are orientation dependent. Journal of Experimental Psychology. Learning, Memory, and Cognition 24, 215–226 (1998)

30. Schinazi, V.R., Epstein, R.A.: Neural correlates of real-world route learning. NeuroImage 53, 725–735 (2010)
31. Shelton, A.L., McNamara, T.P.: Multiple views of spatial memory. Psychonomic Bulletin & Review 4, 102–106 (1997)
32. Shelton, A.L., McNamara, T.P.: Visual memories from nonvisual experiences. Psychological Science 12, 343–347 (2001)
33. Werner, S., Schmidt, K.: Environmental reference systems for large-scale. Spatial Cognition and Computation 1, 447–473 (2000)
34. Waller, D., Hodgson, E.: Transient and enduring spatial representations under disorientation and self-rotation. Journal of Experimental Psychology. Learning, Memory, and Cognition 32, 867–882 (2006)
35. Waller, D., Montello, D.R., Richardson, A.E., Hegarty, M.: Orientation specificity and spatial updating of memories for layouts. Journal of Experimental Psychology: Learning, Memory, and Cognition 28, 1051–1063 (2002)
36. Wang, R.F., Brockmole, J.R.: Human navigation in nested environments. Journal of Experimental Psychology. Learning, Memory, and Cognition 29, 398–404 (2003a)
37. Wang, R.F., Brockmole, J.R.: Simultaneous spatial updating in nested environments. Psychonomic Bulletin & Review 10, 981–986 (2003b)
38. Wang, R.F., Spelke, E.S.: Updating egocentric representations in human navigation. Cognition 77, 215–250 (2000)

An iPad App for Recording Movement Paths and Associated Spatial Behaviors

Nick Sheep Dalton[1], Ruth Conroy Dalton[2], Christoph Hölscher[3], and Gregory Kuhnmünch[3]

[1] The Open University, Walton Hall, Milton Keynes, UK MK7 6AA
[2] Northumbria University, Newcastle upon Tyne, UK NE1 8ST
[3] University of Freiburg, Friedrichstr. 50, 79098 Freiburg, Germany
n.dalton@open.ac.uk, ruth.dalton@northumbria.ac.uk,
{hoelsch,Gregory}@cognition.uni-freiburg.de

Abstract. This paper describes an iPad App, known as 'PeopleWatcher' created for the real-time recording of wayfinding behaviors in buildings/outdoor environments. Initially the paper reviews other spatial-temporal behavioral recording programs and compares their features to the PeopleWatcher App, which is introduced in the next section. The third section presents a pilot study in which the App was tested and discusses the resultant user feedback. It concludes that the iPad is a particularly useful device for behavioral observations in the field, but that further development, the inclusion of post-experiment data-analyses, could be beneficial for future versions of the App.

Keywords: iPad App, Wayfinding and navigation, Wayfinding task, Direct observation, Building usability.

1 Introduction

Since the early 20th Century, researchers of human (and animal behavior) have sought ways to make their observations of behaviors more objective, accurate and less prone to error (particularly see Ittleson et al [16] who developed and popularized behavior mapping/tracking). In the field of spatial behavior, of which wayfinding and navigation research is one sub-category, observational accuracy is required in two dimensions, the accuracy of the spatial location and the precise time of the event being observed (in other words spatial-temporal accuracy). Objectivity can be achieved through the creation of precise definitions and classifications of potential behaviors in conjunction with sufficient training of researchers so that they are able to reliably recognize such events. However, the third requirement, the reduction of human-error in making those observations, is harder to achieve, especially when the environment being observed may be complex or noisy and the numbers of potential participants and different classes of behaviors are large. The cognitive load of an observer, however well trained, might reach the point where human errors can easily

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 431–450, 2012.

occur[1]. Although pen-and-paper methods partially facilitated making accurate and error-free observations, the recent development of mobile computing solutions has served to fill a need for making spatial-temporal observations in the field.

2 Precedents

2.1 Pre-digital and Early Computerized Methods

This section will start by describing antecedents to the PeopleWatcher App (see figure 1 for a diagrammatic time-line representation of antecedent software), and will cover hand-based and computer-based methods formerly employed by researchers investigating wayfinding and navigational behavior in complex environments. This section will then focus on the small number of recent and contemporary software programs that most closely approximate PeopleWatcher with respect to their intended use and functionality.

The first means of recording wayfinding behavior were manual [16]: the participant would be accompanied by a researcher who could simply record (by drawing) their path onto a pre-prepared map attached to a clipboard. Task times could be recorded in parallel using a stopwatch. The primary form of data was the resultant paths or trajectories (which might later be transcribed or digitized for further analysis) and the associated task-durations times. The simple clipboard and stopwatch approach was remarkably robust and amenable to different settings. This 'movement tracing' method was also used extensively by space syntax researchers for observations of 'natural movement' where high-volumes of pedestrian paths were unobtrusively observed without the participant's knowledge. These methods were documented in a handbook [13]. Often, wayfinding tasks will include a pointing task where, typically, a participant is instructed to point to specified locations. These were typically recorded either with a magnetic compass or with a circular dial and then transcribed manually. It is very easy to see how such basic methods continued to be used for a surprisingly long period of time, being both highly accessible, inexpensive yet relatively effective research tools. The problem with such hand-based methods, however, is that it is challenging to record any additional behavioral information other than paths, durations and compass directions.

In order to overcome these deficiencies, wayfinding researchers occasionally employed 'Thinking Aloud Protocols' [10, 23] in which the participant is instructed to verbalize their thoughts and to comment upon environmental features that have caught their attention. In order to record such a potentially rich dataset, there was a need to move from pure hand-based methods to some form of automated recording: initially audio, and later video recording. However, if recording audio alone, the concomitant problem of reconciling the traced-path and the precise location of a pertinent comment or remark arises. This can only be solved through the recording and 'time-stamping' of spatial events.

[1] It could be argued that the cognitive abilities of a human observer will always prove to be a limiting factor in behavioral observations.

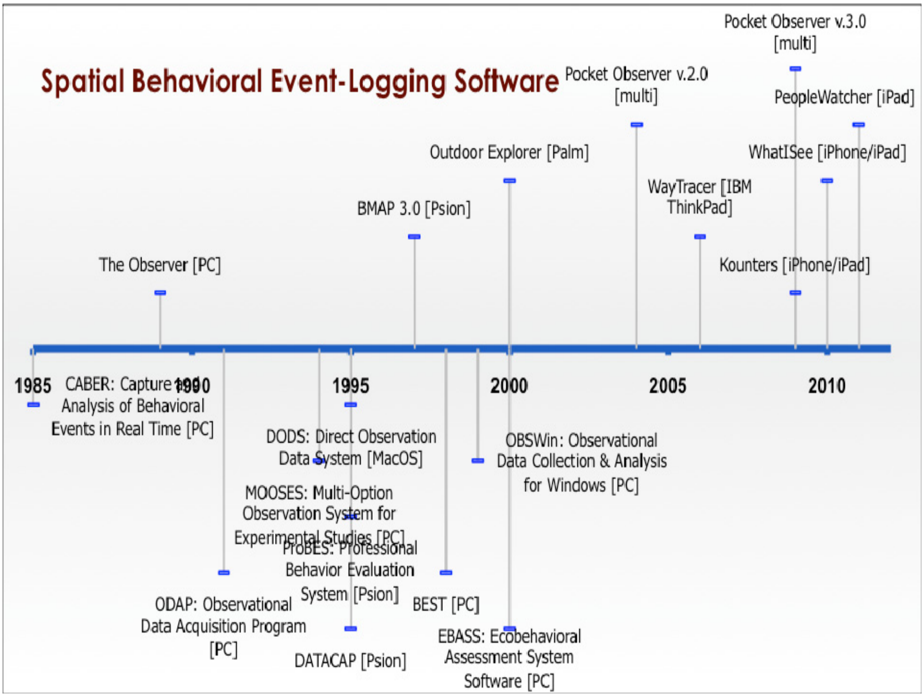
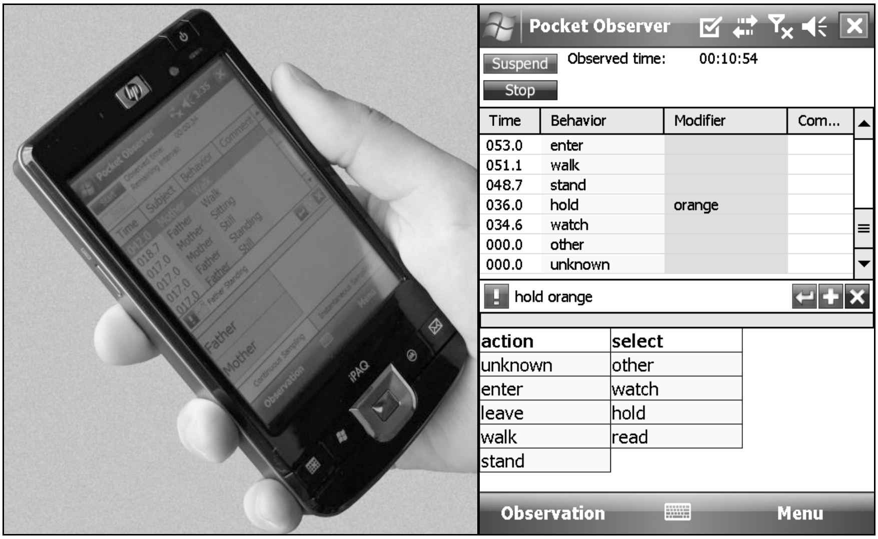


Fig. 1. Timeline showing approximate release dates of known spatial-behavior event-logging software. Applications discussed in section 2 lie above the dateline.

Behavioral researchers were early adopters of computerized methods of behavioral data collection (see figure 1 above): for the most part software was developed within academic environments and rarely commercialized; therefore it is particularly difficult to research the precedents to PeopleWatcher. Most software consisted of the pre-definition of the types of 'events' to be observed, a means of identifying the observee and recording the associated time (and often location as well) of the events. Most behavioral researchers (which include human and animal behavioral researchers) were primarily interested in observing and recording natural behavior in the field and therefore there was also a strong, early impetus to explore mobile computing solutions for fieldwork. Previous reviews of early behavioral recording software include Farrell's paper [12] and the papers by Kahng and Iwata [17, 18] where they review 15 programs for behavioral data collection, including a number of applications for handheld devices (Psion Series 3, Psion Organizer, Psion Workabout, Palm Pilot, and Apple Newton). One of programs they review is the Behavioral Evaluation Strategy and Taxonomy (BEST) software also reviewed in Sidener et al's paper [29] and The Observer software [1, 4, 8, 15, 24, 25]. The Observer software is particularly noteworthy for its longevity, dating back to the late 1980s [25], and is still available today; in the following sections we will begin by reviewing the current mobile solution, Pocket Observer, offered as part of The Observer suite of software solutions.

2.2 Pocket Observer

As mentioned above, The Observer suite of software dates back to the late 1980s (ibid), was produced by Noldus Information Technology and, very early in its development-history, versions for handheld computers were produced [15]. The original desktop version only supported live observations [25] but rapidly developed into supporting post-hoc analyses of video-based observations. In this respect, most of the software solutions provided as part of The Observer family can now be held to be extremely sophisticated video-coding software. However, in 2004 Noldus produced a new version for the handheld computing market, called Pocket Observer 2.0 [9] which ran on a large range of available handheld computers and permitted the time-stamped encoding of up to 250 different participants, 100 behavioral classes and 250 sub-behaviors (ibid). This was updated to Pocket Observer v.3.0 in 2009. Unlike its fully featured, video-based counterpart, Pocket Observer is intended for the collection of real-time data, rather than post-hoc video encoding. Given the mobile and small-screen nature of this version of the Observer suite, there was some criticism that the large number of available behavioral classes meant that scrolling to select the correct one was cumbersome and time-consuming and that, overall, there was a steep learning curve to learning to use the software (ibid). However, noteworthy features of the software is the ability to create codes/classes of new behaviors 'on the fly' if they are observed in the field, and the ability to store notes with individual instances of behavioral events. It should, however, be noted that Pocket Observer does not store the spatial location of any observed behavioral events (although, given the high number of available sub-classes of behaviors that can be defined, location-descriptors could be added as event-modifiers. See figure 2 for screenshots of Pocket Observer.



Observed time: 00:10:54

Time	Behavior	Modifier	Com...
053.0	enter		
051.1	walk		
048.7	stand		
036.0	hold	orange	
034.6	watch		
000.0	other		
000.0	unknown		

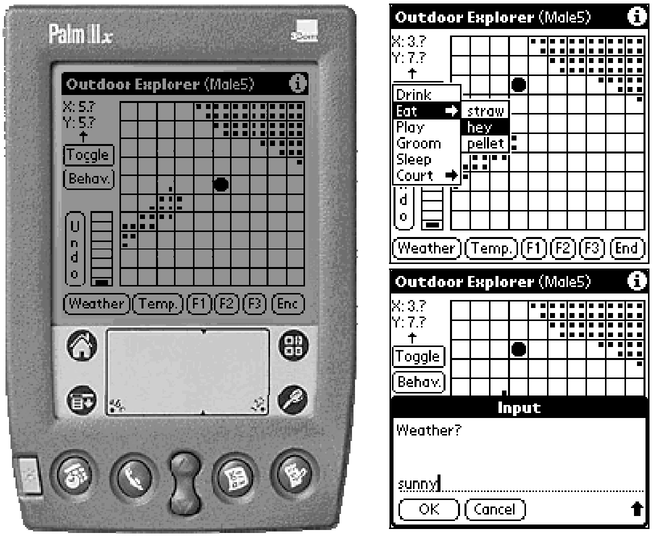
hold orange

action	select
unknown	other
enter	watch
leave	hold
walk	read
stand	

Fig. 2. Pocket Observer software: screenshot on handheld device (left) and example of the data entry screen (right). Image source: Noldus Information Technology.

2.3 Outdoor Explorer

Outdoor Explorer was written by Lars Lewejohann at University of Muenster and appears to have been written for his own academic research circa 2000 [22]. It was written for any handheld computer running Palm OS. What is noteworthy about Outdoor Explorer is its early attempt to record not only the time and type of observed spatial behavior, but also its approximate locations (accurate to within 1m). This is achieved by superimposing an imaginary 10x10 grid upon the observation-setting: this can be edited to include features/boundaries (see figure 3). 15 behavioral classes with 15 behavior-modifiers can be user-customized. When a behavior is observed, it is recorded by tapping the grid-square in which the observee is located (this can be zoomed to a finer 10x10 sub-grid) and this location then linked to the associated behavior/sub-behavior and additional observations/notes.



Weather?

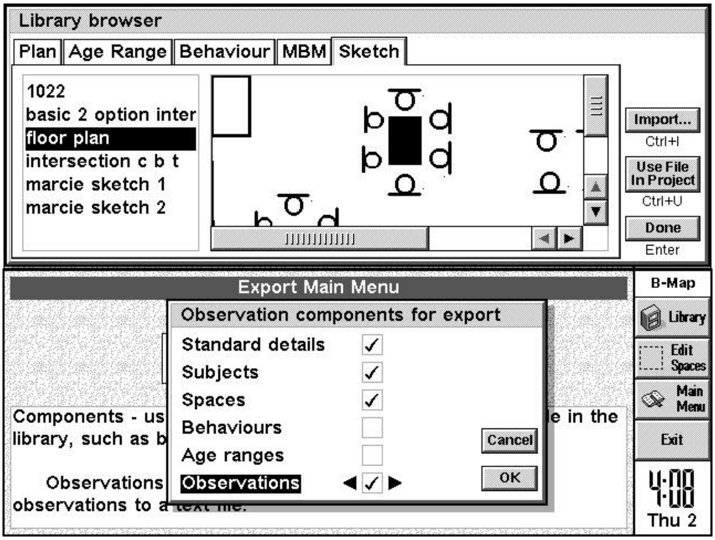
sunny

Fig. 3. Palm interface to Outdoor Explorer software. Image source: [22].

2.4 BMAP

BMAP 3.0 was an application written for the Psion handheld computer for tracking, mapping and time-stamping behavioral events developed by Wener[2] at the Polytechnic Institute of New York University [30–32]. It permitted the import of raster plans/maps as backgrounds and then, when the observer tapped onto a point on the map/plan, the program automatically recorded the location and time of the event and, if set up to include associated data, the tap could prompt a pop-up menu to record other data, such as participant attributes (participant gender or age-range, for example) and types or sub-types of behaviors.

[2] The first distributable version of BMAP was programmed by Alex Wilbur under the guidance of Richard Wener [31].



1022
basic 2 option inter
floor plan
intersection c b t
marcie sketch 1
marcie sketch 2

Observation components for export

Standard details ☑
Subjects ☑
Spaces ☑
Behaviours □
Age ranges □
Observations ◄ ☑ ►

Components - us... e in the library, such as b...

Observations ... observations to a text file.

Fig. 4. Screenshots (top) showing BMAP's floor-plan import facility and (bottom) showing the export data selection screen. Source: [30].

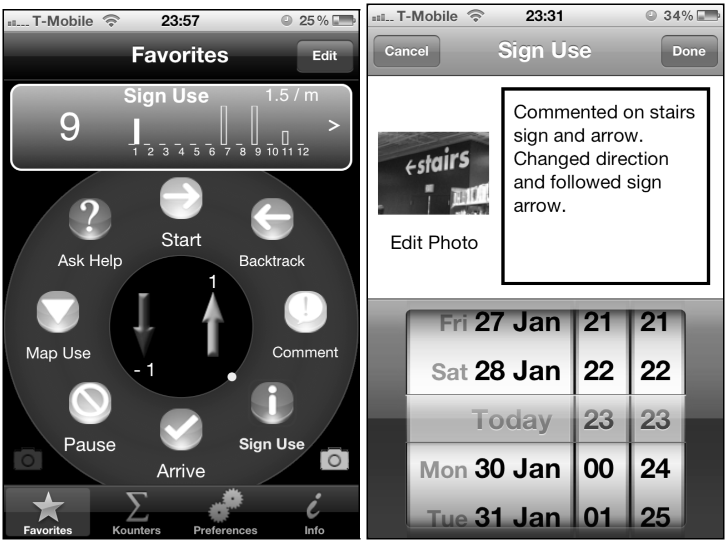
The fact that BMAP 3.0 could simultaneously record locations with event-classes, participant information and times is highly significant, and is the reason why it has been included in this review. However, BMAP was only able to record discrete locations of events: it had a "tracking module" but could not draw continuous lines. It was possible to 'tap-tap-tap' as a person progressed through a space (each tap recorded time and place[3]), which provided a reasonable facsimile for their continuous movement path[4] [31]. BMAP was subsequently tested in several studies [3, 19, 11].

2.5 Kounters

There is a growing number of Apps that keep a tally of user-defined events; we have chosen to review one of the first to be released of these, Kounters. Kounters (currently version 1.3) is an iPhone App released on December 15, 2009 by iPinsoft. It is compatible with iPhone, iPod Touch, and iPad and requires iOS 3.1 or later. Kounters was never intended to serve as a spatial behavior event-logging App; rather it was designed as a generic 'counting' app but with the ability to simultaneously track a number of different 'counters'. However, the reason why it is included in this review is that it highly customizable and is surprising effective at logging wayfinding behavior. Figure 5 shows a screenshot of Kounters, adapted with custom-icons for gathering wayfinding data.

[3] And as opposed to the Outdoor Explorer, the position records of BMAP were not snapped to a grid, and thus could be placed more precisely.

[4] In subsequent analysis (outside the BMAP program environment) it was possible to smooth the sequence points into a continuous line, although this was never directly implemented in the BMAP software as the Psion Series 5 went out of production [31].



Commented on stairs sign and arrow. Changed direction and followed sign arrow.

Fig. 5. Screenshots of Kounters iPhone App, customized for wayfinding experiments. The rightmost screen shows the ability to link a photograph and observer-notes with a time-stamped event. Source: author's iPhone.

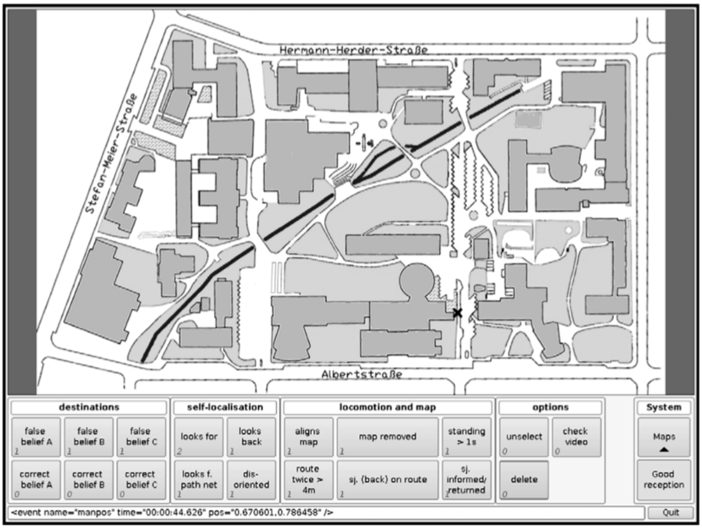
Eight events can be logged, and when the appropriate button is tapped[5], that specific event can either be logged 'positively' or 'negatively' (Kounters permits a negative 'count' or reduction in the tally of recorded events), in which case the event is time-stamped, the total count incremented, and the results are displayed on a small bar chart above the event-buttons. Although not inherently 'spatial', Kounters has one feature that permits the linking of a time-stamped event to a spatial location: every event recorded can be linked to a photograph and observer-notes stored alongside the time-stamped event. This can be particularly useful if, for example, the participant explicitly uses signage. The rightmost screen of figure 5 shows an image and notes associated with a single time-stamped event. Furthermore, if a more precise location is required, it is possible to record the GPS location of the phone in the background, using a second App (there are currently several Apps that will perform this task) and then integrate the time-stamped event log with the GPS log. Furthermore the ability to location-stamp as well as time-stamp events is intended to be included in the next version of Kounters [26].

2.6 WayTracer

WayTracer was developed by Kuhnmünch et al [20, 21] (DFG: SFB/TR8 Spatial Cognition, project I2-MapSpace). It is written in C++ with LINUX as an operating

[5] Events can also be recorded by shaking the iPhone or upon 'sensing' a noise instead of tapping the screen: this has the potential to be useful in certain experimental settings.

system and tested on an IBM ThinkPad X41 tablet PC using a pen-input interface [21]. The functionality of WayTracer is considerable and we are only able to review a small proportion of its functionality in this review. WayTracer consists of a map screen with a series of pre-defined event buttons below it. See figure 6 below for a screenshot of the WayTracer interface.



```
<event name="manpos" time="00:00:44.626" pos="0.670601,0.786458" />
```

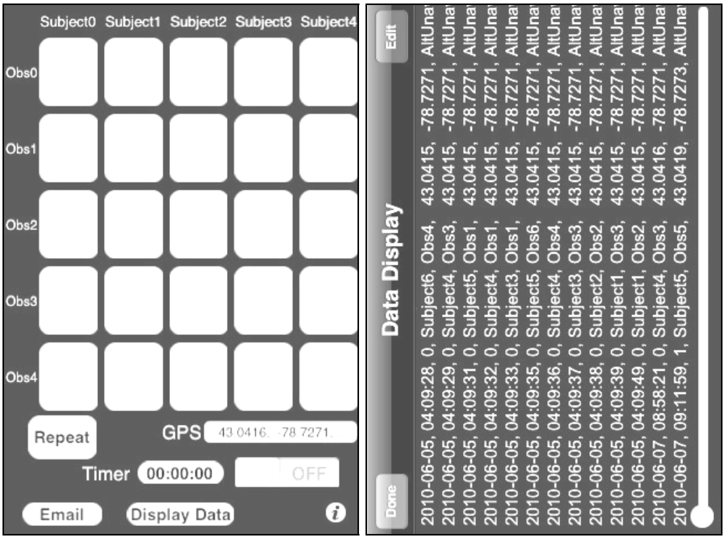
Fig. 6. WayTracer's event entry screen. Source: [20]

The position of the observer is automatically recorded through an attached, external GPS receiver. For indoor use or when GPS signals are not available, the locations of individual events can also be noted manually by pressing the button and then indicating the spatial location of that event (the event is time-stamped on the initial button-press). Event buttons can be configured and grouped by the user under titles of their choice. WayTracer permits switching between multiple maps rendering it useful for multi-floor buildings or large areas. In addition, system feedback on the status of buttons and the GPS signal's quality ease working with WayTracer.

WayTracer was tested extensively in several field experiments and projects within SFB/TR8 [21]; they report testing results of the first experiment in this series, amongst them a high observer agreement between well-trained experimenters of 0.92 (percentage agreement when allowing for a temporal offset of 3s). This indicates that the event-button/stylus interface together with the chosen method of data entry is well-suited for recording such spatio-temporal data and supports reports by Wener about the earlier Psion-based BMAP system [30, 32]. It was noted by Kuhnmünch & Strube [21] that in their tests the manual spatial input often proved more accurate than the accompanying GPS data. This is due to inherent imprecisions of GPS data when signal quality is diminished by the environment (e.g., street canyons). PeopleWatcher shares a significant number of these key functions of WayTracer, and it is officially credited that PeopleWatcher was developed using central design features of WayTracer.

2.7 WhatISee

WhatISee (currently version 2.0) is another iPhone App originally released on June 21, 2010 by Heuser. It is compatible with iPhone, iPod touch, and iPad and requires iOS 3.2 or later. WhatISee is a straightforward App with a simple interface. It does not display maps/plans of any kind, simply presents the experimenter with a matrix of different participants and potential actions/events (this array can be customized for the number of participants/actions and the associated labels can be edited). When an event is observed, by clicking in the correct cell in the matrix, the participant, event/action, time, date and elapsed time (since start of session) as well as location is recorded (see figure 7). The location is recorded by GPS that, unlike the Kounters iPhone App, is integrated into the App. Although a map interface is not provided, the resultant spatial locations of actions can be visualized easily on a map. The locations are obviously discrete locations (instances of discrete events) rather than a continuous path or track through an environment, but the strength of this simple App is its ability to track simultaneously the actions of multiple participants, something neither WayTracer nor PeopleWatcher is able to do.



GPS 43.0416, -78.7271

Timer 00:00:00 OFF

```
2010-06-05, 04:09:28, 0, Subject6, Obs4, 43.0415, -78.7271, AltUna
2010-06-05, 04:09:29, 0, Subject4, Obs3, 43.0415, -78.7271, AltUna
2010-06-05, 04:09:31, 0, Subject5, Obs1, 43.0415, -78.7271, AltUna
2010-06-05, 04:09:32, 0, Subject4, Obs1, 43.0415, -78.7271, AltUna
2010-06-05, 04:09:33, 0, Subject3, Obs1, 43.0415, -78.7271, AltUna
2010-06-05, 04:09:35, 0, Subject5, Obs6, 43.0415, -78.7271, AltUna
2010-06-05, 04:09:36, 0, Subject4, Obs4, 43.0415, -78.7271, AltUna
2010-06-05, 04:09:37, 0, Subject3, Obs3, 43.0415, -78.7271, AltUna
2010-06-05, 04:09:38, 0, Subject2, Obs2, 43.0415, -78.7271, AltUna
2010-06-05, 04:09:39, 0, Subject1, Obs3, 43.0415, -78.7271, AltUna
2010-06-05, 04:09:49, 0, Subject1, Obs2, 43.0415, -78.7271, AltUna
2010-06-07, 08:58:21, 0, Subject4, Obs3, 43.0416, -78.7271, AltUna
2010-06-07, 09:11:59, 1, Subject5, Obs5, 43.0419, -78.7273, AltUna
```

Fig. 7. Two screenshots of the WhatISee App. Source Heuser

Table 1. Comparison of primary features of recent spatial behavioral recording software[6]

Software	Time-stamped events	Multi-classes of event	Discrete spatial locations of events	Continuous paths tracked	Parallel multi-participant logs	Audio transcript	Compass direction for pointing task
BMAP	✓	✓	✓	(✓[4])	✓	✗	✗
PeopleWatcher	✓	✓	✓	✓	✗	✓	✓
Kounters	✓	✓	(✓[8])	✗	✗	✗	✗
Outdoor Explorer	✓	✓	✓	✗	✓	✗	✗
Pocket Observer	✓	✓	✗	✗	✓	✗	✗
Waytracer	✓	✓	✓	✓	✗	✗	✗

3 The PeopleWatcher Approach

PeopleWatcher emerged out of a research program designed specifically conduct building-level[7] (rather than urban-based) usability experiments. Our previous architectural work relied on a paper-based approach but this limited the amount of behavioral data that could be collected which was central to our research aims. We were aware of a number of previous digital approaches to mobile observation (Section 2) but the hardware required was quickly becoming dated (e.g the Psion) or often required complex installation and maintenance. One solution to these problems is the proper choice of the hardware platform, with an easy-to-use software environment. Thus, we were drawn to the iPhone and the iPad. The potential use of the iPad created many possibilities, the long battery life meant a number of experiments could be performed in a single period without pauses to recharge. The iPad is an off-the-shelf consumer-level product that reduces cost and minimizes the technical knowledge necessary to operate and maintain it. Given that one user group for the tool might be student research assistants we prioritized simplicity of use early in the design phase. Furthermore the computing power of the iPad is significantly higher than previous generations of hardware permitting more computing power to be used.

In terms of form-factor, the iPad is light and so less fatiguing to the experimenter during prolonged usage and is light enough to be easily carried in the hand without a

[6] Spatial location of events in Kounters (using GPS) to be included in the next version (Pinzon, 2012).

[7] The focus on architectural usability immediately required an alternative to relying on GPS for recording the paths of participants as GPS is usually unavailable or inaccurate inside a building. It was clear that one primary requirement of any software solution was to be able to trace the path of a participant onto a plan, and to time-stamp the coordinate of the drawn route.

need for straps or other supports. iPads are rapidly becoming a common sight in public so the presence of a participant accompanied by an experimenter holding an iPad does not arouse attention or attract interruption (which can invalidate a task). Experience with building evaluation has shown that building owners prefer that experimenters conduct themselves as discreetly to as possible and minimize disruption to a building's occupants.

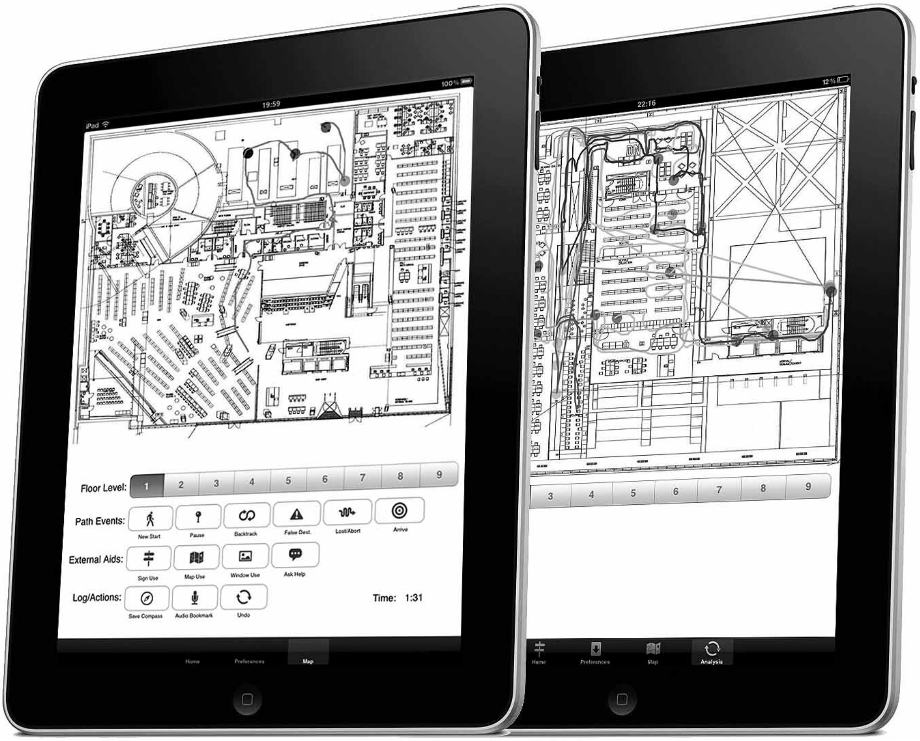


Fig. 8. Screenshots of the PeopleWatcher user interface, showing the 'Map Page' (left) and 'Analytic Page' demonstrating the superimposition of multiple paths (right)

Our original intention was to evaluate and utilize a number of pre-existing iPad applications[8] and use them simultaneously, each performing a different, specialized task. The iPad hardware comes preinstalled with many useful elements: a multi-touch screen, digital compass, Wi-Fi connection, motion sensor, audio recorder, GPS receiver, Bluetooth, LED light and still/video camera, all within one single, small package which would eliminate the need for carrying and connecting several devices.

[8] An event logger (Kounters), a GPS tracker (Location Tracking GPS), a compass (Compass or Direction Known), a drawing program (i.e. Sketches 2, Brushes or Sketch Memo) which would permit a floor-plan/map to be imported as a background image and then drawn upon/annotated and, finally, an audio recorder (Voice Memos))

We encountered two problems with the ‘collection of Apps’ approach. First, it proved impossible to bundle them together into a single display. This meant that a user would need to have some Apps running in the background (i.e. the GPS tracker and audio recorder) whilst rapidly flipping between other Apps as required (i.e. toggling between the event logging and drawing App). This would clearly cause problems for the subsequent accuracy of recording observations and increasing the complexity of the user-interface. Second, this approach requires considerable technical skill to weave together the resultant, disparate data. The lack of a common ‘timeline’ or ‘timeframe’ or even coordinate system between different Apps prolongs the subsequent analysis phase. Our approach became the development of a single App that contained all of the functionality that we had found useful in the pre-existing Apps we had already evaluated, but to combine then in a single application which was simple and robust to use.

3.1 The ‘Home Page’ of PeopleWatcher’s Interface

This section of the paper will describe the primary features of the PeopleWatcher App. The PeopleWatcher app is designed following the standard IOS look and feel. It consists of a lower ‘control bar’ which has four buttons. ‘Home’, ‘Preferences’, ‘Map’ and ‘Analysis’. Each of these pages prescribes a different area of activity. The Home Page is the starting point for both the App and each experiment or experimental session. This page is where new participant records are entered and where the experiment-recording phase is initialized). In PeopleWatcher an ‘Experiment’ typically consists of an experimenter or observer discreetly following a participant. Each experimental session has only one participant[9] but may consist of a number of differing wayfinding tasks. All the tasks for one participant are stored in a single file. The Home Page (functionally analogous to the XML-based ‘metadata’ configuration file used by Waytracer) simply provides an area for the experimenter to set-up a new experiment, initiate and ultimately end, the ‘recording’ phase. An area is also provided on the Home Page to allow the experimenter to record notes, which are subsequently appended to the text-based log-file. A label displaying the automatically assigned, individual participant-ID[10] is displayed on the Home Page, which permits the experimenter to use the participant ID to annotate manually collected information such as additional questionnaire data.

Once an experiment has been initiated the recording begins and each action or event is time-stamped[11] relative to the beginning of the experiment. The Home Page also provides the option for an audio recording track to be recorded in synchronization with the event log-file. This permits the experimenter to either

[9] In this respect PeopleWatcher is differs from Pocket Observer, BMAP and Outdoor Explorer, which permit simultaneous observations of multiple participants.

[10] The automatic ID feature was intended to prevent a human error of assigning an ID number twice and hence overwriting valuable observational data.

[11] Time-stamping is relative, not absolute, as in many of the applications reviewed in the first section of this paper. This was intended to facilitate the easy comparison of the same task performed serially by different participants.

simultaneously record audio notes or to directly record the participant's voice if employing a 'Thinking Aloud Protocol' [10].

Once the recording-phase is initiated, the experimenter typically switches over to the Map Page until the experiment is finished. Once the experiment is over the experimenter must return to the Home Page in order to 'stop' the recording at which point all the relevant data is output to log-files (see section 3.5).

3.2 The 'Map Page' of PeopleWatcher's Interface

The Map Page is the primary area for spatial behavior recording/encoding. The Map page screen is further divided into two sections: the upper 'map' section and the lower 'events' section. The map section displays the current floor level (for a multi-level building) and is a 'drawable' part of the screen, allowing the experimenter to manually trace the path[12] of a participant onto the screen as they observe their progress through an environment. (See figure 8 for a screenshot of PeopleWatcher's Map Page.)

The coordinates of the participant's location are recorded in real-time. The lower half of the screen consists of a series of buttons permitting actions to be logged. The buttons are classified as changes in floor level (at which point the displayed map will be updated accordingly), as path events (starting a new task, pausing, backtracking, arriving at a false destination, becoming lost or giving up the task), the use of external aids (signage, maps, external views to the outside or equivalent invariant views, asking for help) and other log/action events (saving a compass direction in a pointing task, recording the location of a significant remark, if simultaneously recording an audio transcript and 'undo' which deletes any of the previously recorded actions). See figure 8 for a screenshot. Every time an event is logged a colored 'dot' on the traced-path is created: it is time-stamped and its location[13] is noted in the log-file. The text-based log-files, annotated maps and any associated audio files are saved for subsequent retrieval.

To facilitate the use of other recording mechanisms (video, still cameras, paper notes) elapsed time is also displayed during the experiment. There are two timer displays, the first is the overall time which indicates the time from the beginning of the overall experiment-session and coordinates with the times recorded in the log-file

[12] GPS data tracking can be recorded in conjunction with manual tracing for GPS-enabled iPads. However, since it was envisioned that this would be used primarily indoors, where GPS reception is typically poor or unavailable, the development of the App focused on the manual path entry mode. For discrete position entries (either GPS or manual) not to produce artifacts, the positions have to be recorded regularly, especially in curves and when participants repeat sections of the path. Whether recorded manually or by GPS, the main precondition for interpolation of a continuous path by connecting discrete positions hence is a sufficient frequency of position entries. PeopleWatcher overcomes this when users trace paths continuously by hand.

[13] Since the path of the participant is already being drawn manually, the location of the observed action is held to be co-located with the last recorded position of the participant. The time of the event is held to be the point at which the event-button is pressed.

and the second timer indicates the elapsed time from when the 'New Start' (a new sub task) button is clicked. This permits notes to be taken relative to the start of the most recent sub-task: a feature that emerged from the first usability testing of PeopleWatcher.

3.3 The 'Settings Page' of PeopleWatcher's Interface

The Settings Page is used infrequently: its purpose is to import and manage the image-files of maps representing the individual floors (or tiles of a larger map). The pages consist of an array of buttons which, when clicked, prompts the experimenter to upload an image from their photo library. Prior to setting up the experiment, the experimenter is expected to upload a maximum of nine floor plans from the iPad's Photo Library. Experiments have shown that navigating (panning and zooming) a large-scale map is difficult in a mobile context [27] so images should cover the entire navigable area of a building's floor-plan. Multiple-floor maps can also be adapted to complex single-floor settings, for example an airport, in which case the 'floor' images can be used to represent different adjacent zones at 'zoomed-in', higher resolutions.

If preparing for a multi-floor experiment, while preparing the floor plan images experience has shown it is useful to take care to ensure that the individual maps are vertically registered. That is to say, any vertical circulation elements such as the lift/elevators and the staircases should align between floors. On the Map Page, when a new floor plan is selected (as a participant moves from one floor to another) then the last point on the previous floor plan is registered (displayed as a dot) on the current floor plan. This acts as a visual prompt for the experimenter to rapidly reorient him/herself (on the new plan) while simultaneously observing participant-behavior during the often-important vertical transition of the participant. This registration between floor plans is also important to calculate inter-floor metrics from the log files. For example, if an experimenter calculates walking speed from the spatial data-stream and time then there will be an implicit assumption of scaling factor (from screen pixels to meters or feet) if the individual floors do not represent the same scale.

It is planned that for future versions of PeopleWatcher the Setting Page will also permit users to change the labels and icons of the individual buttons, for example customizing the events to be logged.

3.4 The 'Analytic' Page of PeopleWatcher's Interface

Once a number of experiments with different participants have been completed it is usual to undertake some form of post-experiment analysis. This can be achieved in one of two ways. The Analytic Page is designed to provide an early-stage, rapid analysis in order to identify any possible experimental problems or areas for more in-depth research. The Analytic Page resembles the Map Page (see figure 8) and displays one of the available floor plans (user-selected from a radio button control). The analytic view merges all of the experimental tracks from each participant on the selected floor together. Each participant is assigned their own unique color facilitating the identification of points/zones of commonality (similar routes taken or zones where multiple participants paused). This composite view permits the simple observation of emergent behavior events common to all participants. In addition, when placing two fingers on the screen on the map (on the Analytic Page) a line or ‘virtual gate’ is created between the two touch points. The system then counts the number of routes crossing this virtual threshold and displays the total in the bottom left-hand corner of the map (see [13] for more background on the ‘gate-counts’ observation method).

In future versions it is intended that the user will be able to draw a polygonal ‘zone’ onto the plan displayed in the Analytic Page and PeopleWatcher will calculate how many paths pass through this zone and enumerate the number of different events that take place within the selected area. For more sophisticated or specific analysis PeopleWatcher reports the observation data can be exported to the desktop environment (see section 3.5 for PeopleWatcher’s file format).

3.5 PeopleWatcher File Format

PeopleWatcher has the potential to generate a number of output files. The primary output file is the event log-file and this takes the form of a simple .csv (comma separated values) format. There are a separate event log-files for each participant and they are named as ‘PathXXX.cvs’ where XXX is the participant number automatically generated and displayed on the home page when the ‘new experiment’ button was clicked. Accompanying each event file is a number of .pdf vector files: one for each floor/zone of the building. Each of these files contains a high resolution, vector version of the information displayed on the Map Page for participant XXX. This displays the participants’ recorded path along with dots representing the location of recorded events (such as pauses). The original map image-file is stored as a background for the vector path permitting the path to be viewed against its context (without the need to re-generate the path in third-party mapping software). If the experimenter has used the optional audio recording an additional file “AudioXXX.caf” is also saved.

The PeopleWatcher file format is designed to be open and simple enough to permit other tools to be written to facilitate the further processing of the data held. The text-based .csv format can be simply imported into packages from Excel to large SQL databases and is simple enough to be edited by hand in a simple text editor. The file is in the form: “event-type, floor, time, x-coordinate[14], y-coordinate”.

4 Pilot Testing the PeopleWatcher App

4.1 Procedure

The PeopleWatcher App was tested in the field in November 2011 as part of a genuine wayfinding experiment. Experiment participants, the majority of whom were attendees at the 52nd Annual Meeting of The Psychonomic Society, were invited to

[14] The coordinate system is in pixels, relating to the image-display of the map, in the Map Page.

participate in a the experiment that took place in Seattle Public Library. 28 participants (13 M; 15 F) aged between 21 and 68 years (mean age of 35) presented themselves at the library's Fourth Avenue entrance, level 1, at pre-arranged time-slots, throughout a single day (Nov 3, 2011). All participants were, hitherto, unfamiliar with the building. They were tested individually. The tasks consisted of four separate wayfinding tasks: two took place within a single floor (levels 1 and 3) and therefore involved no vertical travel and two tasks involved navigating from one level to different level. Of the two 'within floors' tasks, one was intended to be relatively easy and one harder and this pattern was mirrored exactly by the 'between floors' tasks (L1 to L4; L5 to L7). The tasks (E='easy'; H='hard'; W='within floors'; B='between floors') were as follows:

- E&W: Starting from the Fourth Avenue entrance (L1) and finding the boys/girls restrooms in the Children's Center (L1);
- H&B: From the Story Hour Room (L1) and finding Meeting Room 6 (L4);
- H&W: From the far end of the Teen Center (adjacent to the staff-only meeting room) (L3) and finding a book (Sherlock Holmes by Arthur Conan Doyle) located in the Mystery Fiction section located behind the red staircase (L3);
- E&B: Starting in front of the Info Desk (L5) and finding the non-fiction DVDs located on level 7 of the Book Spiral (L7). Please refer to figure 9 for an illustration of the four routes. The tasks were designed to be 'chained', that is to say, once a participant had concluded one task in the sequence they were led, by one of the experimenters, to the starting location of the subsequent task, ready to begin again. In this way, four participants could be tested simultaneously, each beginning at one of the four separate starting points and being observed by one of four experimenters, and then moving in rotation between the tasks, so as not to inadvertently 'overlap' with one another.

It is clear from the description above, that the experiment set-up served to be a particularly challenging first test of the PeopleWatcher App. Not only were four participants being tested simultaneously (and therefore the four experimenters needed to use PeopleWatcher installed on four separate iPads) but the routes frequently involved changes in floor-level by different modes (stairs, escalators and elevators), a proportion of the tasks were intentionally designed to be spatially complex and, finally, the library staff requested that the experiment be conducted discreetly and not disturb the library's patrons. In addition to this, the four experimenters came from three different institutions, and had little or no familiarity of PeopleWatcher in advance of the experiment. Due to the nature of the event, only limited training in its use was possible (1-4 hours per observer): although all the PeopleWatcher users had conducted similar experiments in the past, and therefore were familiar with general procedures for indoor navigation experiments.

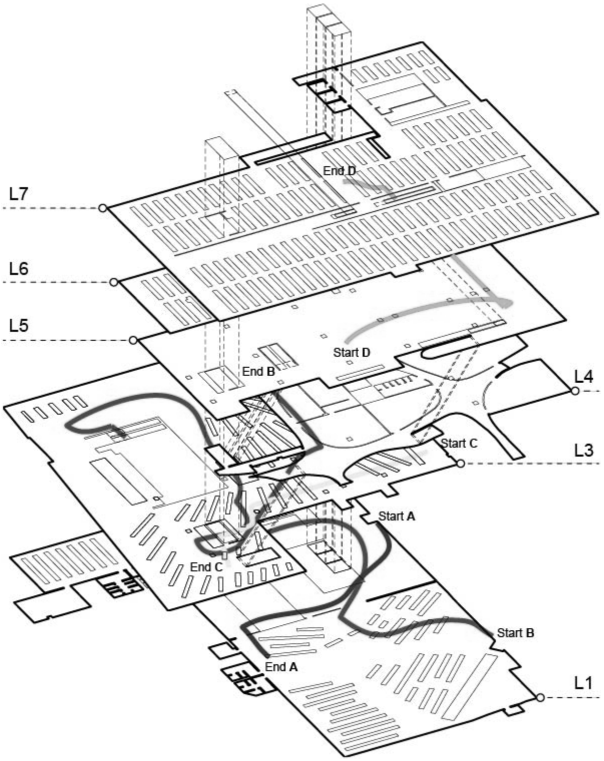


Fig. 9. An illustration of part of Seattle Public Library (levels 1 to 7 only) showing the paths of the routes used as wayfinding tasks in PeopleWatcher's test-study

PeopleWatcher enabled the experimenters to observe and record the path taken by the participants for each of the four tasks, whilst simultaneously logging actions or behaviors observed en route, such as pausing, using signs, backtracking etc. These actions were recorded and logged in real time, with the location and time of the behavior/action being recorded by the App and was subsequently downloadable as a single data-file. Such real-time, electronic collection of the data enabled the unusually rapid analysis[15] of the participant data and constitutes one of the notable aspects of this study.

The detailed results of the experiment in the Seattle Public Library will be published separately [5]. But the available preliminary analysis [6, 7] already indicate that the data collected with PeopleWatcher is sensitive to systematic variations in the wayfinding tasks and that it captures behavioral differences that can be traced to differences in psychometric tests like the SBSOD [14] and mental rotation [28]. In this test-study we concentrated on collecting trajectory data, pauses and sign use. For

[15] The experiment was conducted at the beginning of the Annual Meeting of The Psychonomic Society; the results were presented at the end of the Annual Meeting.

each of these measures significant statistical effects were obtained in line with study hypotheses, as well as significant correlations to individual difference measures[16].

5 Usability Feedback

As evidenced in Kuhnmünch and Strube's paper [21], they found a high degree of inter-observer agreement when testing this aspect of WayTracer's usability. Given the similarity of the user-interface of PeopleWatcher to WayTracer (they both utilize map- and button-pressing event-sampling user-inputs) we saw no reason to re-produce their inter-observer agreement study, as we would expect the results to be broadly comparable. Rather, we decided to focus on the overall usability of the App. Since one of the reasons for using the iPad (see section 3) was the perceived ease of use of the iPad plus accompanying benefits of its lightweight form, we wished to test whether PeopleWatcher did, as intended, result in an easy-to-use solution for gathering spatial behavior data.

Post-experiment, the four experimenters who took part in the Seattle Public Library study were asked to gauge the usability of the PeopleWatcher App by filling in a usability questionnaire based on System Usability Scale (SUS) and selecting descriptive words from a version of Microsoft's Product Reaction Cards [2]. In addition to this, they were invited to make any other comments on the use of PeopleWatcher in an open question. This section of the paper will discuss the usability of the PeopleWatcher App based on the results of and feedback of the experiment and will discuss any implications for the automation of data gathering of human behavior in the future.

While the number of users in this field-test was low, the feedback does give an indication of the attention to usability that influenced the development of the App and the early results of this process. The SUS yields a single number representing a composite measure of the overall usability of the system. SUS cores have a range of 0 to 100 with the mode (of a number of commonly used industrial products and websites along with research projects) between 71-80. Our early system had a score of 71 putting it in the top 50th percentile or on a par with everyday products. One question that many users gave a low score for was 'I need to learn a lot of things before I could get going with this app' which may reflect either the experience of the users of conducting wayfinding experiments or may reflect the usability of application itself: this will be further investigated in future usability tests.

The Microsoft Product reaction cards highlighted the words '*Straightforward*', and '*Valuable*' with '*Unpredictable'* as the only negative response[17]. Almost all of the responses to the open question concerned suggestions for additional functionality for future versions. Many of these suggestions have already been incorporated into the current version of PeopleWatcher. The only comment that did not specifically address

[16] Gender differences have yet to be analyzed for this dataset and hence no conclusive results are currently available. It should however be stated that the 'Note' field located on the App's 'Home Page' can be used to record gender data, if required as part of the experimental setup.

[17] This needs further research to reveal the origins of this response.

future functionality requests, stated, "*Overall I thought it was easy to use and well suited for the task.*" The next stage of PeopleWatcher's development will include further user-testing and continued use of the above methods to investigate its usability.

6 Conclusions

It is clear from the review of previous software solutions that behavioral researchers benefit from easy-to-use, mobile and discreet software solutions to gather real-time data in the field. We would also like to suggest that the iPad, with its many inbuilt-features and lightweight form, is ideally suited to this task. The aim of PeopleWatcher was to take advantage of as many of these inbuilt features as possible and to bring them together in a single, easy-to-use package. The Seattle Public Library study proved a challenging first test of the software. The App is currently in its second stage of development and is in the process of responding to the feedback from the study.

Future developments will focus on reliably concerning the hard- and software and enhancing the in-App data-analysis functionality. Further test-studies are planned and ongoing usability studies will form an integral part of this process. As part of this commitment to long term usability an off-the-shelf package called 'Flurry', which annotates the software and anonymously sends problems/errors to a central sever for later diagnosis, has been included in the current version of PeopleWatcher.

Acknowledgments. Christoph Hölscher is and Gregory Kuhnmünch was supported by DFG within the SFB/TR8 Spatial Cognition. The Seattle Public Library study is an ongoing project initiated by Laura Carlson and Amy Shelton with Christoph Hölscher and Ruth Dalton and we are grateful for their collaboration.

References

1. Baber, J.: Tools and techniques: The Observer. Applied Ergonomics 25, 398–399
2. Benedek, J., Miner, T.: Measuring Desirability: New methods for evaluating desirability in a usability lab setting. In: Proceedings of Usability Professionals Association, pp. 8–12 (2002)
3. Benne, M.R.: Methods for assessing influences of the visual-spatial environment on museum display attraction (2001)
4. Boccia, M.L.: Practical Computing. Practicing Anthropology 17, 59–61
5. Carlson, L.A., et al.: Conducting wayfinding experiments within buildings: The Seattle Public Library (in preparation)
6. Carlson, M.L., Shelton, A.: Wayfinding in the Seattle Public Library: What Can We Learn About Navigational Styles? In: Proceedings of the Psychonomics Society, Seattle, WA, pp. 49–51
7. Dalton, R.C., Hölscher, C.: Navigating the Seattle Public Library: Usability, Cognition and Building Analysis. In: 8th International Space Syntax Symposium, Santiago de Chile (2012)

8. Davis, A.: The Observer: An integrated software package for behavioural research. Journal of Animal Ecology 62, 218–219 (1993)
9. Eckhardt, G., Waterman, J.: Pocket Observer 2.0:: By Noldus Information Technology. Animal Behaviour 67(4), 805–806 (2004), `http://www.noldus.com`
10. Ericsson, K., Simon, H.A.: Verbal reports as data. Psychological Review 87(3), 215 (1980)
11. Farbstein, J., Wener, R.: A comparison of "direct" and "indirect" supervision correctional facilities. National Institute of Corrections, Washington, DC (1989)
12. Farrell, A.D.: Computers and behavioral assessment: Current applications, future possibilities, and obstacles to routine use. Behavioral Assessment (1991)
13. Grajewski, T., Vaughan, L.: Space Syntax Observation Manual. UCL Bartlett and Space Syntax Limited, London (2001)
14. Hegarty, M., et al.: The Santa Barbara Sense of Direction Scale (2001)
15. Hile, M.: Hand-held behavioral observations: The Observer. Behavioral Assessment 13, 187–196 (1991)
16. Ittelson, W.H., et al.: The use of behavioral maps in environmental psychology. In: Environmental Psychology: Man and his Physical Setting, pp. 658–668 (1970)
17. Kahng, S., Iwata, B.A.: Computer systems for collecting real-time observational data. In: Behavioral Observation: Technology & Applications in Developmental Disabilities, pp. 35–45 (2000)
18. Kahng, S.W., Iwata, B.: Computerized systems for collecting real-time observational data. Journal of Applied Behavior Analysis 31(2), 253 (1998)
19. Kantrowitz, M., Farbstein, J.: POE delivers for the post office. In: Building Evaluation Techniques. McGraw-Hill, New York (1996)
20. Kuhnmünch, G., et al.: WayTracer-A mobile assistant for logging navigation behaviour. In: Spatial Cognition 2006: Poster Presentations, vol. 17 (2006)
21. Kuhnmünch, G., Strube, G.: Waytracer: a mobile assistance for real-time logging of events and related positions. Computers in Human Behaviour 25(5), 1156–1164 (2009)
22. Lewejohann, L.: `http://www.phenotyping.de/Outdoorexplorer.html`
23. Lynch, K.: The image of the city. MIT Press, Cambridge (1960,1971)
24. Noldus, L.P.J.J.: The Observer: a software system for collection and analysis of observational data. Behavior Research Methods 23(3), 415–429 (1991)
25. Noldus, L.P.J.J., et al.: The Observer Video-Pro: new software for the collection, management, and presentation of time-structured data from videotapes and digital media files. Behavior Research methods 32(1), 197–206 (2000)
26. Pinzon, I.: (2012)
27. Setlur, V., et al.: Towards designing better map interfaces for the mobile: experiences from example. In: Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application, p. 31 (2010)
28. Shepard, R.N., Metzler, J.: Mental rotation of three-dimensional objects. Science 171, 3972, 701 (1971)
29. Sidener, T.M., et al.: A review of the Behavioral Evaluation Strategy and Taxonomy (BEST®) software application. Behavioral Interventions 19(4), 275–285 (2004)
30. Wener, R.: BMAP 3.0: Handheld Software to Support Behavior Observations. In: 33rd Annual Conference of The Environmental Design Research Association, Philadelphia, PA (2002)
31. Wener, R.: (2012)
32. Wener, R.: (2011)

Detection of Object Onset and Offset in Naturalistic Scenes

Maria J. Donaldson and Naohide Yamamoto

Department of Psychology, Cleveland State University
2121 Euclid Avenue, Cleveland, OH 44115, USA
{m.j.donaldson,n.yamamoto}@csuohio.edu
http://academic.csuohio.edu/n_yamamoto/lab/

Abstract. The present study was conducted to investigate whether observers are equally prone to overlook any kinds of visual events in change blindness. Capitalizing on the finding from visual search studies that abrupt appearance of an object effectively captures observers' attention, the onset of a new object and the offset of an existing object were contrasted regarding their detectability when they occurred in a naturalistic scene. In an experiment, participants viewed a series of photograph pairs in which layouts of seven or eight objects were depicted. One object either appeared in or disappeared from the layout, and participants tried to detect this change. Results showed that onsets were detected more quickly than offsets, while they were detected with equivalent accuracy. This suggests that the primacy of onset over offset is a robust phenomenon that likely makes onsets more resistant to change blindness under natural viewing conditions.

Keywords: Change blindness, change detection, scene, onset, offset.

1 Introduction

Vision can provide rich information about details of an environment, if observers specifically attend to them. Conversely, when sufficient attention is absent, it is surprisingly easy to miss large changes in the visual field. For example, people often do not notice anything when an object in a movie scene suddenly appears or disappears from one clip to the next [1]. Typically, such failure in noticing changes occurs when they take place during a brief disruption of visual access to the environment (e.g., while blinking or during saccadic eye movements [2]) or when they are accompanied with other prominent visual events [3]. This inability to detect potentially salient changes is called change blindness, and it has been well demonstrated that observers can be unaware of a variety of changes such as displacement of an object [4] and color change in an existing object [5], to name but a few (for review, see [6,7]). However, what remains unclear is whether observers are equally prone to overlook any kinds of change or they are more resistant to change blindness with certain kinds of change. The purpose of the present study was to address this issue by focusing on two major types

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 451–460, 2012.

of visual change that happen in everyday environments: onset (appearance) and offset (disappearance) of an object. An experiment was conducted to investigate whether onset and offset in a naturalistic scene differ in their susceptibility to change blindness.

An important clue for approaching this question has been provided by studies on visual search. In these studies, observers were typically presented with an array of simple objects such as alphanumeric characters and geometric shapes, and then instructed to indicate the presence of a target among distractors. It has been found that how quickly and accurately the observers detect a target is strongly influenced by stimulus attributes the target possesses. For example, targets that display unique shape, color, or movement in the search array tend to be detected with greater speed and accuracy [8,9]. Importantly, among these different types of targets, one of the most robust targets that consistently elicits enhanced detection is an object that abruptly appeared in the search array [10]. Although sudden disappearance of an object also makes an effective target [11], direct comparisons between appearing objects and disappearing objects often showed that targets defined by their appearance induce more efficient detection than those defined by their disappearance [12,13,14]. These results indicate that observers can notice the onset of a new object more easily than the offset of an existing object, suggesting that onsets are more resistant to change blindness than offsets.

However, great caution must be taken when findings from visual search studies are used to predict that onset should have higher resistance to change blindness than offset. This is due to the fact that previous studies on visual search and those on change blindness have utilized substantially different visual stimuli to carry out their investigation. In visual search studies, search arrays were constructed by arranging simple objects (e.g., letters) in a restricted manner so that various variables that could affect search efficiency (such as color, luminance, and locations of objects) were tightly controlled. As a result, the primacy of onset over offset was established under conditions in which there were few other visual features that could concurrently influence observers' performance. On the other hand, change blindness has typically been investigated by using photographs of real-world scenes that contained diverse visual features [4]. Thus, in order to make predictions regarding change blindness by applying the finding from visual search studies, it needs to be examined first whether the onset primacy is robust enough to be observed under viewing conditions in which many concomitant variables are less controlled.

Initial efforts have already been made to bridge the gap between visual search studies and change blindness studies. Cole and colleagues [13] presented colored pictures of an array of household objects and asked participants to detect a change that occurred in the array. The change was either the onset of a new object or the offset of an existing object. Consistent with their previous experiments in which an array of simple geometric shapes was used, onsets were detected more accurately than offsets in this experiment. This provided important first evidence that the onset primacy holds true when the search array contains

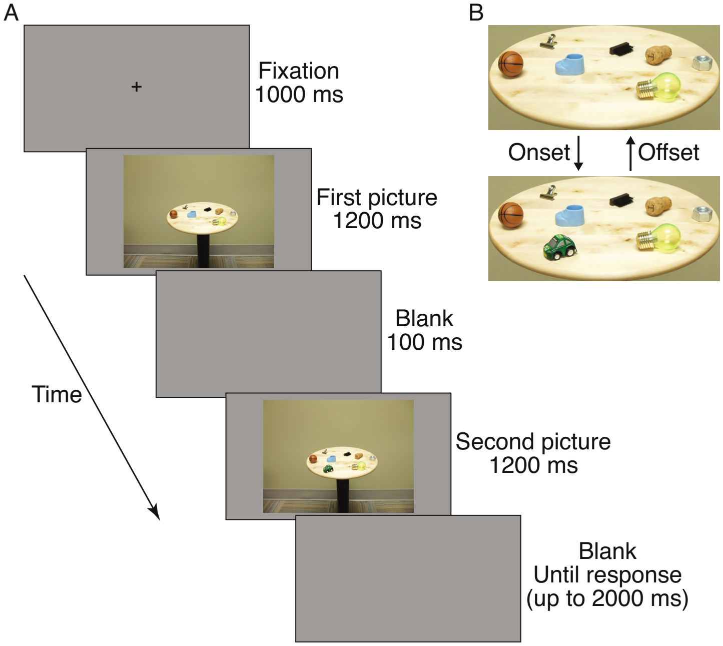


Fig. 1. (A) The trial sequence and examples of scene stimuli used in the present experiment. In this case, an object (the car) is added to the second image on the left-hand side. (B) Close-up views of the object array.

a wide variety of visual features. However, in order to make the array of household objects look sufficiently similar to search arrays used in other experiments, Cole et al. [13] arranged the objects in an arbitrary configuration and took the pictures from an aerial perspective. As a result, although the array of household objects successfully facilitated cross-experiment comparisons, it did not afford naturalistic viewing conditions that observers have in typical change blindness studies.

There were two studies that used photographs of real-world scenes to examine whether onsets and offsets are detected with different degrees of efficiency. In Mondy and Coltheart's study [15], an object was digitally added to or deleted from an existing photograph and participants tried to detect the onset or offset of the altered object. Mondy and Coltheart found that offsets were detected more accurately than onsets in this study. Brockmole and Henderson [16], on the other hand, took a different approach by making two new photographs of a scene with or without an additional object. These two photographs were presented successively and seamlessly while participants were viewing them, thereby creating an onset or offset of the object. Participants' eye movements were recorded during viewing, and it was found that they tended to fixate the area occupied by the changed object sooner when it appeared in the scene than when it disappeared from the scene. This result suggests that onsets were detected more quickly than offsets.

In summary, findings have been mixed when actual scenes were used to contrast onsets and offsets regarding their vulnerability to change blindness [15,16]. This discrepancy cannot be resolved by experiments using complex real-world scenes because each of those scenes contains a unique set of visual features and semantic contexts. In other words, there is always a possibility that observed findings stemmed from some peculiar characteristics that happened to be possessed by particular scenes used in the experiments. Thus, in order to investigate whether the primacy of onset over offset is still observable under naturalistic viewing conditions, it is necessary to conduct an experiment that utilizes more controlled but still realistic scene stimuli. The present study was designed to carry out this investigation by developing simple three-dimensional scenes that afforded much more natural viewing conditions than the bird's-eye views of arbitrarily arranged objects [13] while maintaining good control over various visual features that might affect detectability of objects. If onsets were detected with greater speed or accuracy than offsets in the experiment reported below, it would increase the likelihood that onsets are more immune to change blindness than offsets even under naturalistic viewing conditions. On the other hand, if the experiment failed to find the advantage of onsets over offsets in change detection, it would support a claim that the onset primacy is a subtle phenomenon that can be observed reliably only under tightly controlled viewing conditions, and therefore the susceptibility of onsets and offsets to change blindness is mostly equivalent in everyday scene viewing.

2 Method

2.1 Participants

Thirteen participants (4 males and 9 females, 19–47 years of age) from the Cleveland State University community volunteered in return for extra credit in psychology courses. All participants were right-handed and reported normal or corrected-to-normal vision.

2.2 Materials

Experimental stimuli presented to participants were color digital pictures that depicted a wooden round tabletop on which six to nine objects were placed in various different arrangements. The objects were toys and small household goods that were approximately 4 cm in width, 2 cm in depth, and 3 cm in height. The tabletop was 38 cm in diameter and supported by a table base that was 75 cm tall. The objects were spread across the entire tabletop such that every object was visible in its entirety. The wall behind the table was also visible, as was the carpet on which the table stood. These additional items in the stimuli provided rich depth cues, thereby making the depicted scenes more naturalistic. For examples of the stimuli, see Fig. 1.

These images were presented on a 21.5-inch liquid crystal display by using the PsyScope program [17]. The screen was positioned vertically in front of the participant. The distance between the participant and the screen was approximately 60 cm. The images were presented in the center of the screen against a gray background and subtended approximately 29.3° (horizontal) × 22.3° (vertical) of visual angle. The tabletop measured 14.3° horizontally and 6.7° vertically. When presented on the screen, the center of the tabletop was approximately at the center of the screen.

2.3 Design

The experiment was modeled after the paradigm developed by Cole et al. [13]. Participants viewed a series of photograph pairs in one block. Each pair constituted either an onset trial in which a new object was added to the second image or an offset trial in which one of the objects in the first image was deleted in the second image (see Fig. 1). The same 64 photograph pairs, in which images contained either seven or eight objects, were used to create onset and offset trials by reversing the order of presentation of two images in each pair. The 128 trials (64 onset trials and 64 offset trials) composed of these pairs were the experimental trials from which data were collected. By using the same photograph pairs for these experimental trials, it was ensured that any particular properties of objects or their configurations (such as their color, location, or semantic salience) would equally affect onset and offset trials. In the experimental trials, each object was used the same number of times to create an onset trial or an offset trial (i.e., all objects were presented an equal number of times). The participant's task was to detect the change as accurately and quickly as possible by indicating whether it occurred in the right half or the left half of the tabletop. The location of the change was counterbalanced such that in half the onset trials objects in the left side changed and in the remaining half the objects in the right side changed. The same was done for offset trials.

Additional 32 pairs of photographs were used to create filler trials in which either an eight-object image was followed by a nine-object image or a seven-object image was followed by a six-object image. These filler trials (16 onset trials and 16 offset trials) were added to make the first image of each pair unpredictable of trial types: Without the filler trials, participants could potentially figure out whether it would be an onset trial or an offset trial just by looking at the first image (e.g., a seven-object image as the first image could indicate that an eight-object image would follow). Given that these numbers of objects (seven or eight) exceed the typical capacity of visual short-term memory (approximately four objects or six locations under optimal conditions [18,19]), it is unlikely that participants were actually able to predict the trial type. Furthermore, it was not readily clear whether this predictability would differentially affect performance in onset and offset trials. However, the filler trials reduced any such biases if they existed. As a result of adding the filler trials, participants performed 80 onset trials and 80 offset trials in total that were randomly intermixed. Data from the filler trials were not included in the analysis.

After receiving instructions about how to perform the task, participants did 16 practice trials (eight onset trials and eight offset trials randomly intermixed).

Photograph pairs that were not used in the experimental or filler trials were presented in these practice trials. Participants were not given any feedback on their performance throughout the experiment.

2.4 Procedure

Participants were told that they were going to view a series of photograph pairs in which an object would somehow change between two images of each pair; however, they were not informed of the nature of the change. They were instructed to press either the F key or the J key of a standard American English computer keyboard, depending on where on the tabletop the change occurred: The F key was for changes in the left and the J key was for changes in the right. They were also instructed to place their index fingers on these two keys all the time. They were cautioned to be as quick and accurate as possible. They were run individually.

Fig. 1A illustrates the trial sequence. In each trial, participants first viewed a fixation cross for 1000 ms that was presented at the center of the tabletop. They were instructed to keep fixating on the cross while it was displayed. However, eye movements were not constrained while participants viewed subsequent images. They then viewed the first image for 1200 ms, which was followed by a brief (100 ms) blank gray screen. The second image immediately followed and was displayed for 1200 ms. After the second image, the screen turned into gray until participants made a response or 2000 ms passed, whichever came faster. Subsequently, participants were presented with the fixation cross again, which indicated the beginning of the next trial. Reaction time was defined as time elapsed between the appearance of the second image and the participants' key press. Accuracy in the left/right judgment was also measured. When incorrect responses were made, reaction time data from those trials were excluded from the analysis. When no response was made, it was regarded as an incorrect response.

3 Results

Data from two participants were excluded from the analysis because their overall accuracy (8.6% and 63.1%) was substantially lower than that of the rest of the group (96.7%). Thus, data from 11 participants were analyzed below. Reaction time and accuracy were separately compared between two trial types (onset or offset) by paired *t*-tests.

3.1 Reaction Time

After removing trials in which incorrect responses were made, mean reaction time was computed for each participant and for each trial type. Outliers were defined as data points that were more than three standard deviations away from each participant's mean reaction time to each trial type. Twenty-eight data points

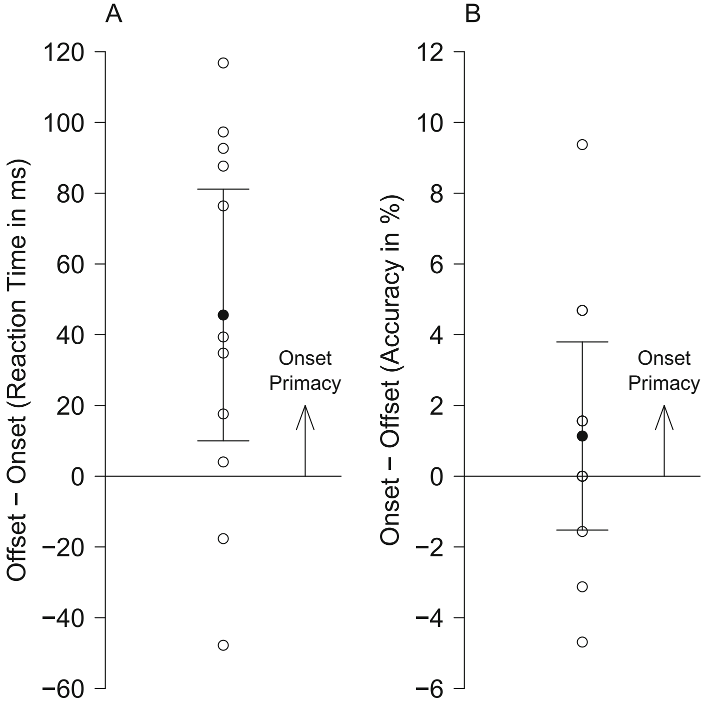


Fig. 2. Differences between onset and offset trials in (A) reaction time and (B) accuracy. Each open circle represents one participant. Filled circles represent the means of all participants. Error bars represent 95% confidence intervals. Note that the direction of subtraction between onset and offset trials is reversed in reaction time and accuracy so that positive values indicate onset primacy in both panels.

were defined as outliers and removed from the analysis. They constituted 2.0% of the data.

Fig. 2A plots differences between onset and offset trials (defined by offset − onset) shown by each participant. As illustrated in this figure, participants responded to onsets with significantly shorter reaction time, $t(10) = 2.85, p = .017$. Mean reaction time (M) and standard deviation (SD) for each trial type were as follows: $M = 668.3$ ms, $SD = 310.2$ ms (onset); $M = 713.9$ ms, $SD = 284.2$ ms (offset).

3.2 Accuracy

Fig. 2B plots differences between onset and offset trials (defined by onset − offset) shown by each participant. Although participants responded to onsets with higher accuracy than offsets, this difference was not statistically reliable, $t(10) = 0.95, p = .36$. Mean accuracy and standard deviation for each trial type were as follows: $M = 97.3\%, SD = 2.1\%$ (onset); $M = 96.2\%, SD = 3.4\%$ (offset).

Mean accuracy for each trial type was reliably different from 100%, $t(10) = -4.25, p = .0017$ (onset); $t(10) = -3.77, p = .0037$ (offset), confirming that there were no ceiling effects.

4 Discussion

The present study was conducted to investigate whether the onset of a new object is detected with greater speed or accuracy than the offset of an existing object when they occur in a naturalistic scene. Participants responded more quickly to onsets than to offsets, while they noticed both types of change equally accurately. These results suggest that the primacy of onset over offset regarding their detectability is robust enough to endure in a complex visual environment, thereby making onsets less susceptible to change blindness than offsets.

Why should the detection of onsets be more enhanced than that of offsets? One possibility is that onsets draw observers' attention more strongly than offsets. When a new object appears, it requires observers to newly form a mental representation of the object (so-called object file [20]). It has been proposed that the creation of an object file causes observers' attention to be directed to the new object in an exogenous manner [21]. When an object disappears, on the other hand, it only entails the deletion of an existing object file. It is likely that this process does not always result in the increase of attention to the disappearing object, making observers less efficient in detecting offsets.

Another possibility is that onsets are easier to detect than offsets because onsets are usually accompanied with a greater amount of sensory transients. When an object appears in a scene, it locally creates a large change in luminance (e.g., from plain gray to a complex combination of various patterns in the current experiment). On the other hand, when an object disappears, it tends to reveal only a background that is filled with a relatively simple pattern (e.g., from the gray to the wooden board pattern). Thus, it is possible that a larger amount of luminance change allowed observers to accumulate sufficient information sooner for determining the presence of an onset than an offset. The results from the present experiment are consistent with both possibilities, and in fact, there has been an active debate in the visual search literature as to whether it is a new object itself or the contribution of accompanying sensory transients that makes the detection of onsets especially efficient [22,23]. It would be an important challenge for future research to examine (and resolve) this issue in the context of change blindness as well.

Given that many of the previous studies and the present study provided evidence for the primacy of onset over offset, it is not readily clear why Mondy and Coltheart [15] found that offsets were detected more accurately than onsets. One possibility is that some idiosyncratic characteristics of the stimuli used by Mondy and Coltheart caused their unique finding. For example, it is conceivable that digital alteration of the photographs produced sharper-than-usual contrasts between an altered object and its surroundings, and they somehow biased detection accuracy in favor of offsets. This possibility did not exist in the Brockmole and Henderson's study [16] and in the present study because two photographs were taken with or without an additional object to create onsets and offsets. Another point of note is that Mondy and Coltheart only measured accuracy. This leaves open the possibility that there were speed-accuracy tradeoffs in their study. That is, it is possible that offsets were indeed more difficult to detect in their experiments too, but the increased difficulty led participants to slow down in offset trials, resulting in higher accuracy for offsets than onsets. The present study excluded this possibility by measuring both reaction time and accuracy.

In conclusion, the present study showed that the onset of a new object can be detected more efficiently than the offset of an existing object, even when scenes are more realistic and thus contain more visual noises than abstract arrays of simple objects. This helps bridge the gap between visual search studies and change blindness studies, facilitating the use of rich knowledge gained through visual search studies for understanding change blindness. In particular, it is suggested that not all visual events are equal in their susceptibility to change blindness; rather, as results from visual search studies indicate, abrupt onset of a new object is especially resistant to change blindness. Future studies should build on this finding to investigate the mechanisms with which the detection of onsets is enhanced under naturalistic viewing conditions.

Acknowledgments. The authors thank Kamil Klinkosz for assistance with stimulus development and data collection.

References

1. Levin, D.T., Simons, D.J.: Failure to Detect Changes to Attended Objects in Motion Pictures. Psychon. Bull. Rev. 4, 501–506 (1997)
2. Grimes, J.: On the Failure to Detect Changes in Scenes across Saccades. In: Akins, K. (ed.) Perception. Vancouver Studies in Cognitive Science, vol. 5, pp. 89–110. Oxford University Press, New York (1996)
3. O'Regan, J.K., Rensink, R.A., Clark, J.J.: Change-Blindness as a Result of 'Mudsplashes'. Nature 398, 34 (1999)
4. Rensink, R.A., O'Regan, J.K., Clark, J.J.: To See or Not to See: The Need for Attention to Perceive Changes in Scenes. Psychol. Sci. 8, 368–373 (1997)
5. Arrington, J.G., Levin, D.T., Varakin, D.A.: Color Onsets and Offsets, and Luminance Changes Can Cause Change Blindness. Perception 35, 1665–1678 (2006)
6. Rensink, R.A.: Change Detection. Annu. Rev. Psychol. 53, 245–277 (2002)
7. Simons, D.J., Rensink, R.A.: Change Blindness: Past, Present, and Future. Trends Cogn. Sci. 9, 16–20 (2005)
8. Franconeri, S.L., Simons, D.J.: Moving and Looming Stimuli Capture Attention. Percept. Psychophys. 65, 999–1010 (2003)
9. Theeuwes, J.: Stimulus-Driven Capture and Attentional Set: Selective Search for Color and Visual Abrupt Onsets. J. Exp. Psychol. Hum. Percept. Perform. 20, 799–806 (1994)

10. Yantis, S., Jonides, J.: Abrupt Visual Onsets and Selective Attention: Evidence from Visual Search. J. Exp. Psychol. Hum. Percept. Perform. 10, 601–621 (1984)
11. Theeuwes, J.: Exogenous and Endogenous Control of Attention: The Effect of Visual Onsets and Offsets. Percept. Psychophys. 49, 83–90 (1991)
12. Cole, G.G., Kentridge, R.W., Heywood, C.A.: Visual Salience in the Change Detection Paradigm: The Special Role of Object Onset. J. Exp. Psychol. Hum. Percept. Perform. 30, 464–477 (2004)
13. Cole, G.G., Kentridge, R.W., Heywood, C.A.: Detectability of Onsets versus Offsets in the Change Detection Paradigm. J. Vis. 3, 22–31 (2003)
14. Cole, G.G., Kuhn, G.: Attentional Capture by Object Appearance and Disappearance. Q. J. Exp. Psychol. (Hove) 63, 147–159 (2010)
15. Mondy, S., Coltheart, V.: Detection and Identification of Change in Naturalistic Scenes. Vis. Cogn. 7, 281–296 (2000)
16. Brockmole, J.R., Henderson, J.M.: Object Appearance, Disappearance, and Attention Prioritization in Real-World Scenes. Psychon. Bull. Rev. 12, 1061–1067 (2005)
17. Cohen, J., MacWhinney, B., Flatt, M., Provost, J.: PsyScope: An Interactive Graphic System for Designing and Controlling Experiments in the Psychology Laboratory Using Macintosh Computers. Behav. Res. Methods Instrum. Comput. 25, 257–271 (1993)
18. Jiang, Y., Olson, I.R., Chun, M.M.: Organization of Visual Short-Term Memory. J. Exp. Psychol. Learn. Mem. Cogn. 26, 683–702 (2000)
19. Luck, S.J., Vogel, E.K.: The Capacity of Visual Working Memory for Features and Conjunctions. Nature 390, 279–281 (1997)
20. Kahneman, D., Treisman, A., Gibbs, B.J.: The Reviewing of Object Files: Object-Specific Integration of Information. Cogn. Psychol. 24, 175–219 (1992)
21. Yantis, S.: Stimulus-Driven Attentional Capture. Curr. Dir. Psychol. Sci. 2, 156–161 (1993)
22. Davoli, C.C., Suszko, J.W., Abrams, R.A.: New Objects Can Capture Attention without a Unique Luminance Transient. Psychon. Bull. Rev. 14, 338–343 (2007)
23. Hollingworth, A., Simons, D.J., Franconeri, S.L.: New Objects Do Not Capture Attention without a Sensory Transient. Atten. Percept. Psychophys. 72, 1298–1310 (2010)

Wayfinding in Real Cities: Experiments at Street Corners

Beatrix Emo

Bartlett School of Graduate Studies, UCL, UK
b.emo@ucl.ac.uk

Abstract. Experimental evidence sheds new light on the role of spatial geometry for wayfinding in real urban environments. Eye-tracking is used in a desktop-based experiment to study where people look during wayfinding decisions when let to look freely or asked to find a taxi rank. Gaze patterns from these two tasks are compared with a subsequent recall task and analyzed in light of the topology of the street grid. Results show that decisions strongly favor more connected streets, and that fixation patterns respond to the spatial geometry of the stimuli in both the spatial decision-making and recall tasks. Controls single out the impacts of lighting and affordances in both the behavioral responses and gaze bias patterns; the presence of people and traffic serve as particularly strong attractors. The paper highlights the role of spatial geometry for individual spatial decision-making in real urban environments.

Keywords: Wayfinding, spatial configuration, space syntax, gaze bias, individual.

1 Introduction

The paper provides initial experimental evidence for the role of spatial geometry during wayfinding in real urban environments. A desktop-based eye tracking experiment shows where people look when making wayfinding decisions, either when looking freely or when searching for a taxi rank. The stimuli used in the experiment were photographs of urban street corners. Task-related viewing patterns are controlled for through a recall task. Two strands of analysis are explored. First, the behavioral responses of the spatial decision-making tasks are compared with the spatial configuration of the urban layout; this is recorded using the space syntax measures of Integration and Choice at global and local scales. Second, the gaze bias patterns are analyzed; fixations in the spatial task conditions are compared with those in the recall task. The similar gaze bias pattern between the wayfinding tasks and the recall task are an indicator that subjects are responding to the spatial geometry of the stimuli.

The paper begins with an overview of relevant previous work. It then reports the experiment design, the procedural methods and the nature of the data analyses. The paper discusses the behavioral results in light of the space syntax model of street connectivity. Results show that the space syntax model is effective when assessing individual spatial decision-making. The paper then turns to the gaze bias patterns and highlights the similarities between the spatial tasks and the recall task. Results in this paper provide initial experimental evidence for the role of spatial geometry in real

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 461–477, 2012.

world individual spatial decision-making; future work should provide specific visuo-spatial measures that model the gaze bias recorded.

1.1 Wayfinding Behavior

The effect of environmental variables on wayfinding behavior has been discussed in various studies. Wayfinding can be defined as the decision-making process stage of navigation, where navigation is composed of locomotion and wayfinding [1]; wayfinding is necessarily related to choices made by the individual. Weisman's seminal study identified floorplan configuration, architectural differentiation, visual access and signage as environmental variables affecting wayfinding behavior [2]. The role of spatial configuration on individual spatial decision-making is particularly interesting because it suggests that the layout of the environment itself affects the choices that individuals make. The term spatial configuration refers to the way every space in the built environment relates to every other; at an urban scale, it defines the connectivity between all the streets in a network. Peponis, Zimring and Choi's study of wayfinding behavior suggested a positive influence of spatial configuration on wayfinding performance [3]. This was substantiated in a later study that examined a large number of participants whilst wayfinding in hospitals [4]. A recent paper finds that measures relating to spatial configuration can be used to explain wayfinding behavior [5]. All of these studies use space syntax-related techniques to measures spatial configuration. This study, building on results in [6], aims to test the hypothesis that wayfinding behavior in an urban environment is largely affected by the spatial configuration of the street grid.

Several other factors have been shown to affect wayfinding behavior. Wayfinding strategies are based on spatial knowledge, of which three types have been identified: landmark, route and survey [7]. This paper looks at path search activities, with no information pertaining to either route or survey knowledge. Wayfinding is affected by familiarity; several studies have noted a change in wayfinding behavior when subjects are familiar with the environment (eg. [8], and more recently [9]). Familiarity is not a condition tested in this study, although it is controlled for, as all subjects were not familiar with the environment. Wayfinding behavior is also affected by attractors, such as people and traffic (eg. [10], [11], and [12]); this study specifically controls for both factors. Another variable is the spatial ability of the subject, which can be tested by recording the wayfinding performance (eg. by using the Santa Barbara Sense of Direction Scale, see [13]); this is not a feature of this study, as there was no correct wayfinding behavior.

The explicit or implicit purpose of wayfinding is a crucial factor in explaining wayfinding behavior. It follows that studies wishing to examine wayfinding behavior must be acute in their identification of a task; a taxonomy of wayfinding tasks has been proposed to help classify studies that use different tasks [14]. The simplest wayfinding task is to ask, 'which way would you go?'. It forces a choice based on a purpose that has not been made explicit, although the subject, in reaching a decision, will have acknowledged some type of purpose. It is suggested that a basic human characteristic, when faced with an open question of which way to go, is to assess how connected the path alternatives are. Thus a subject looking at the same scene in two instances, with a different purpose in mind in each case, might well reach a different conclusion as to which way to go. This paper explores the role of a task in wayfinding behavior by examining directed and undirected path search activity.

1.2 Using the Space Syntax Model to Measure Spatial Configuration

The paper tests the role of spatial configuration on wayfinding behavior. It measures spatial configuration using the space syntax environmental model (see [15] for an introduction to space syntax). At an urban scale, the space syntax model reduces the street grid to the set of longest and fewest lines of sight with potential for movement in the system; this is known as an axial map (eg. figure 5). These lines of sight are then broken at each junction into segments, so that each segment begins and ends at an intersection with another line. This produces a network of intersecting lines which can be analyzed as a graph, with the segments as nodes and the intersections as the links connecting them (see [16] on graph theory). The benefit of using the segment model is the ability to weight the graph for angular displacement, which seems to reflect individual spatial decision-making [17].

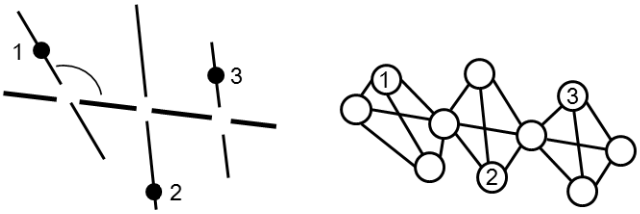


Fig. 1. Segment angular model of a street network (left); its associated graph (right) indicates which are the more connected segments in the model

The model is solely based on the connectivity of streets, and does not include any other type of information such as land use or traffic flow. The model has proved accurate at the aggregate level, although it seems likely that it is also relevant at the individual level, given that it seems to reflect the way humans interact with their surroundings [18]. The issue of developing the space syntax model has gained importance in the space syntax community [19]. The specific merits of the space syntax model for individual spatial decision-making were highlighted by Hillier and Iida, in which they argued for the geometric and topological capabilities of the model over existing metric models [17]. Only few studies have attempted to relate real world navigational decisions with spatial configuration (eg. [4] and [5]), and these tend to be in an indoor setting. This paper provides experimental data that can be used to test the effectiveness of the space syntax model on choices made by the individual in urban settings.

1.3 Wayfinding and Gaze Bias

Using an eye tracker to record gaze bias is a useful method in navigation research to avoid the subjectivity of asking participants how they chose which way to go. Several studies use eye tracking devices to record the viewing patterns in navigation-related research; however only few studies have used eye tracking to examine path search behavior during wayfinding. A recent paper reveals the benefit of recording fixation during when analyzing wayfinding behavior in a study based on virtual stimuli [20]. Another study records wayfinding behavior in a real world setting using a mobile eye tracker [21]; however the difficulty in identifying fixations suggests that real world studies need not necessarily collect data in situ. This paper uses color photographs of the real world.

Several models of the visual perception of real world stimuli exist, against which collected gaze bias data has to be tested; these span top-down (task related) and bottom-up (stimulus derived) influences (see [22] for an overview). The most widely recognized bottom-up approach uses saliency maps, which model the orientation, color and intensity properties of the stimulus (eg. [23]). Purely bottom-up models offer a reliable explanation for covert attention patterns, but often fail to account for contextual elements in more complex viewing behavior. To this end, the contextual guidance model includes information that is likely to be relevant when viewing the stimulus [24]. At the other end, top-down models suggest that viewing patterns can be largely explained through task-related influences; Yarbus' pioneering study offers evidence for this model [25].

Both of these models are addressed in this study. To account for the influence of low-level properties on viewing behavior, lighting conditions are controlled for in some of the stimuli. Specifically, in the instance where one path alternative is far brighter than the other, a control condition is presented where both path alternatives have the same brightness. To account for task-related influences, the experiment design accounts for spatial and non-spatial tasks. By having a change of task, it will be possible to assess to what extent the viewing behavior is determined by the nature of the task.

1.4 Spatial Complexity

Another benefit of recording fixation data is the ability to examine in greater detail the role of spatial complexity on wayfinding behavior. Factors relating to spatial complexity play a role in determining wayfinding, although it is sometimes difficult to distinguish between spatial complexity and configuration. Several studies acknowledge the need to study the complexity of the configuration; for Weisman the two factors go hand in hand [2]; O'Neill discusses the measure of topological complexity [26] and, more recently, Hölscher et al. [5] examine floorplan complexity.

Spatial complexity can be measured by looking at the number of objects in a scene, which provide directly deducible information as to which way to go. This type of complexity is specifically accounted for in this experiment through the choice of stimuli used. In addition, this study tests the hypothesis that people and traffic act as affordances.

Another way to measure spatial complexity comes from isovist analysis[1]. An isovist is a 2D polygon that represents the vista around one point, its generating location (see [29] for a pioneering study of the use of isovists in the built environment). Different properties of an isovist represent geometric information in the environment (see for example ([29], [30] and [31]); some of these have been shown to correlate with navigation behavior (eg.[5], [32] and [33]).

Whilst isovist analysis and visibility graph analysis (which examines the interrelation between individual isovists, see [30]) provide an accurate measurement of the geometric properties of a viewshed, they tend to be based on the architectural representation of the environment. For studies undertaken in a virtual environment, the viewsheds match the perceptual information the participant is presented with. However, in the real world, the different forms of isovist analyses do not match a subject's sensory information; street furniture, moving obstacles, contrasting light conditions and overhead obstructions are all examples of how the structural properties of a real world viewshed might differ from the viewshed drawn off an architectural representation of the environment. A recent innovation in isovist analysis seeks to adapt existing techniques to the challenges of real world experiments. Six image properties have been proposed that represent geometric properties of the urban environment as perceived by the viewer: depth of view, visual connectivity, percentages of visible sky and floor areas, the ratio of sky to floor area and the longest permeable route [34]. In a similar approach, but using virtual stimuli, Wiener et al. [20] propose the depth profile in conjunction with the number of edges as a useful measure of geometric information in the environment as perceived by the viewer.

Although the type of spatial complexity addressed by isovist-inspired analyses is not controlled for in this study, it will be a salient component of a subsequent study that aims to model the gaze bias reported in this study through visuo-spatial measures.

1.5 Specific Aims and Hypotheses

The paper addresses two specific research questions.

- To test whether people understand the global configuration of the built environment when viewing a local viewpoint during wayfinding. This is tested by comparing the number of wayfinding decisions that can be modeled using space syntax measures of spatial configuration.
- To examine what people look at when making wayfinding decisions. The hypothesis is that the spatial geometry of the local viewpoint is crucial, although the presence of attractors is likely to be paramount; this should be measured using isovist-related techniques. Only initial evidence towards this research question is provided here, which should be tested more fully in subsequent research.

[1] Note that the concept of an isovist was originally used in landscape analysis by Hardy [27] and the term coined by Tandy [28].

2 Methods

An experiment was designed to test the role of spatial geometry during wayfinding. Participants were asked to choose between path alternatives, either when looking freely or when searching for a taxi rank. In a subsequent task, participants were ask to recall whether a particular stimulus had already been shown. The stimuli used in the experiment were photographs of urban street corners.

2.1 Stimuli

Spatial Tasks. The stimuli were 28 photographs taken at urban street corners in the City of London, which is the historical core, and an important financial and legal center, of London. Each stimulus presents a decision point with a distinct binary choice of one left and one right path alternative. The photographs were taken for the study using a tripod and a camera specific head to ensure that the camera itself (and not just the tripod) was level with the ground. The height of the lens was 160cm, the average eye height for people in the UK, with minimal discrepancies to ensure that the camera was level.

The photographs aim to present the viewer with a clear understanding of the spatial geometry of the environment. Elements of a scene that reflect a temporary quality, such as people, vehicles or building works were avoided. This leaves an image in which the spatial geometry of the scene is intact – a necessary feature of a study on the spatial geometry of wayfinding in urban environments.

Another merit of the stimuli is that they evoke wayfinding activity; responses in a post-study interview of a pilot study suggest that participants were responding to the photographs as if they were in situ. This is important because i) it confirms the role of wayfinding as a static condition of navigation that occurs at decision points and ii) this study promotes the benefits of modeling real world wayfinding behavior using photographs.

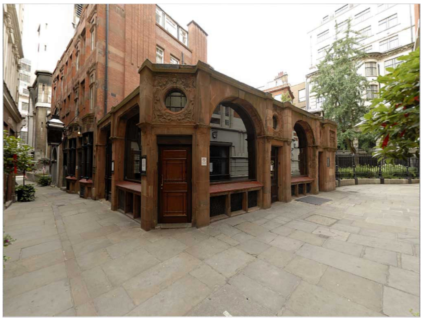


Fig. 2. Example stimulus

Several of the variables identified in the literature were controlled for. Attractors were controlled for in the choice of stimuli by including some stimuli where people and vehicular traffic are present.

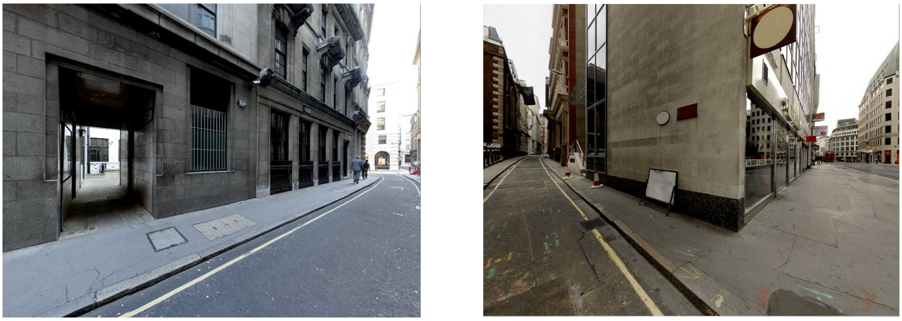


Fig. 3. Example stimuli for the control conditions of attractors; people (left) and a bus (right) are visible

Top-down viewing properties of the stimuli were controlled for by altering the light conditions in two of the images, such that the two path alternatives were seen to have equal light settings. The images were chosen to test the effect of light conditions because the path alternatives feature an alleyway leading an opening onto a subsequent street.

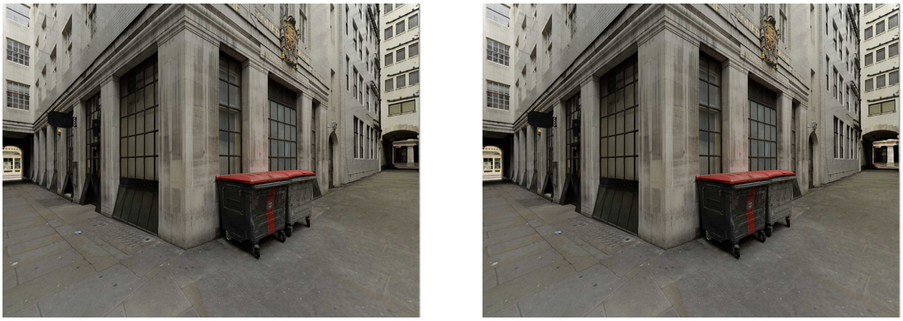


Fig. 4. Example of a control condition for light. The left image shows one path alternative that is considerably darker than the other; this bias is not present in the right image.

Recall Task. For the recall task, random subsets of stimuli already used in the experiment were interspersed with different stimuli, from which the initial set were originally derived. This ensured that the stimuli in the recall task were similar in nature to the original stimulus set.

2.2 Tasks

Spatial Tasks. The experiment examines path search behavior, where no information is provided of route or survey knowledge. In order to fully investigate this kind of wayfinding activity, two instances are assessed: i) an undirected instance where participants looked freely, responding to the question ‘which way would you go’ with no other information being provided and ii) a directed instance where subjects were asked to look for a taxi rank. Another benefit of examining these two types of wayfinding activity is to be able to assess how important the introduction of a specific task is.

A pilot study identified what kind of task would be suitable for examining the directed path search behavior. The aim was to have a task that was spatially defined as well as visually neutral, to promote a natural viewing behavior during the eye tracking experiment. By having a task that promotes the use of spatial configuration to aid the decision process, it is more likely that space syntax measures will be able to model those decisions. The hypothesis is that the specific task will facilitate the decision-making process – it is suggested that lower response times will reflect this.

Recall Task. In order to control for the top-down viewing properties of the stimuli, a non-spatial task was introduced into the experiment design. Should the viewing pattern of the stimuli be determined by top-down effects, a change of task should produce vastly different viewing patterns. A commonly used non-spatial task is a recall task, implementing learning processes. Participants were asked to recall whether a stimulus had already been shown in the experiment; they expressed a positive or a negative response, corresponding to an accurate or an inaccurate recall response.

2.3 Participants

15 participants took part in the experiment, of which 10 were female. The average age was 29.9 ±5.9years. None of the participants were familiar with the area used in the study; this can be deduced from responses in the post-study interview in which none of the participants said they recognized the images, nor were they able to state with any accuracy which urban environment they were taken from.

2.4 Procedure

A desktop-based pan/tilt ASL Eye-Trac 600 was used, with a screen refresh rate of 100Hz. Participants sat in front of the monitor at a distance of 60cm so that the resulting visual angle was 35 degrees on the horizontal scale and 27 degrees on the vertical scale. A nine-point calibration grid was used before and after data was collected to ensure that the eye movements were fully calibrated with the equipment. Subjects viewed the stimuli at a resolution of 1024 x 768 pixels on a 20” CRT monitor. Stimuli were shown until a response was made. A white screen with a central black cross was shown for 2 seconds in between stimuli to foster similar viewing behavior for each stimulus.

Participants filled in a brief questionnaire informing them of the nature of the study. This included information asking them to look carefully at the stimuli and to choose to go either left or right according to a task; responses were recorded by using the arrow keys on the keyboard. They were told to follow on-screen instructions relating to a change of task during the experiment. Participants were asked to make their choice as soon as they were confident with that decision.

To control for any left/right bias a version of each stimulus was generated that was mirrored on the vertical axis. Each participant was shown the full set of stimuli including the mirrored version (n=28) in random order. The spatial tasks were blocked, with the undirected instance being shown first; this was the necessary procedure to avoid participants responding to the more specific directed task in the undirected instance. Participants viewed half the stimuli when responding to the undirected task; following on-screen instructions, participants viewed the remaining half of the stimuli while responding to the directed task.

Subsequently, on-screen instructions asked participants to recall whether they had already been shown any of the following stimuli. Five randomly chosen stimuli from the full stimulus set were interwoven with the same number of images not previously viewed; these images (n=10) were then shown in random order.

2.5 Space Syntax Measures

The space syntax model was used to provide the measure of spatial configuration for each path alternative in the stimuli. The space syntax model (refer to section 1.2 for more detail) is reached by reducing the street grid to a series of interconnecting lines that form a network (eg. figure 5). This network can be analyzed as a graph, with the street segments as nodes and the junctions between segments as the links connecting the nodes (figure 2). The graph is measured according to different mathematical properties (see [16] that are believed to reflect urban movement (for a greater discussion of this refer to [17]).

The space syntax measures allow for a scientific definition of the more connected street, derived from graph centrality measures. For each stimulus, the connectedness of both path alternatives was compared and the more connected path alternative identified; two connectivity measures were used at two scales. The measure of closeness, or Integration as it is commonly known in space syntax research, reflects how likely it is that a segment is an origin or destination segment. The measure of betweenness, or Choice, reflects how likely it is that a segment features as an intervening space in between an origin and destination. These measures can be recorded at different scales, reflecting the variable length of journeys in the urban network; at a basic level it is helpful to distinguish between local and global scales. The scale is recorded in terms of a radius, centered on the segment in question, measured in metric distance.

The space syntax model used in this study was based on an axial map of Greater London, restricted to the area of interest (the center of London) plus a catchment area of three miles to avoid any edge effects. All the analysis was segment angular analysis. Integration and Choice were measured at the global scale (r=n) and at the local scale of radius 100m; note that it is customary in space syntax research to record the natural logarithm of Choice. Below are the graphical representations of the measures from the space syntax model; the darker the line, the more connected the street is.

The selection of the local radius of 100m was reached through a pilot study that measured the average depth of view in a large set of photographs from which the stimuli used in the present study were ultimately derived. In addition, it is reasonable to assume that most spatial decisions are reached based on information within 100m from the current standpoint.

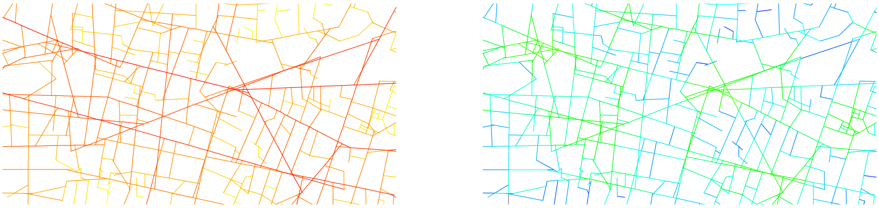


Fig. 5. Connectivity graph of the center of London based on the space syntax measure of Integration at global (left) and local (right) radii

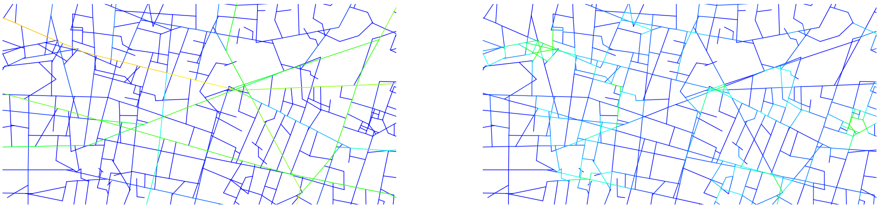


Fig. 6. Connectivity graph of the center of London based on the space syntax measure of Choice at global (left) and local (right) radii

2.6 Analysis

Two strands of analysis are explored: analysis of the behavioral data is followed by the gaze bias data. In each case the spatial task data is compared with the recall task data.

Behavioral Analysis

Spatial Tasks. The behavioral analysis of the spatial task data is divided into three sections: response time, connectivity and controls.

Response Time

Connectivity. The number of decisions each participant made that favored the more connected street was recorded. The concept of the more connected path alternative refers to the space syntax measures of spatial configuration, which are mathematical measures based on a topological model of the urban street grid (see above). Two space syntax measures were used, each at local and global scale, resulting in four variables:

- Integration r=n;
- Integration r=100m;
- Choice r=n;
- Choice r=100m.

For each stimulus, the more connected path alternative for the above measures was identified. The number of decisions made by each participant that favored the more connected street was recorded. The data was tested against the null hypothesis that decisions were random using a one-sample t-test. Note that both the undirected and the directed search task in this experiment encouraged the participant to choose the more connected of the two path alternatives; the merit of having both tasks is to be able to assess the impact of the specific task (taxi rank) from free roaming activity.

Controls. Any Left/Right bias was controlled for. The number of decisions each participant made that went the same way, regardless of whether the path alternative had been shown on the left or right hand path choice was recorded. A one-sample t-test was used to test the probability that participants' choices did not fit a 50:50 model. Additionally, decisions in the controls for light and attractors are described; given the small sample size for these control conditions, no statistical tests are provided.

Recall Task. Time-course analyses and the success rate of the recall task are examined. The main aim of the recall task is to test task-related viewing patterns, thus more analysis is provided in the gaze bias section below.

Gaze Bias Analysis. Several aspects of the recorded gaze bias were analyzed: to which path alternative each fixation belonged; the average fixation duration; the time-course of the fixations; and the location of the fixation.

3 Results

3.1 Behavioral data

Spatial Tasks. Results from the behavioral data shed light on the hypotheses.

Average response time. The average response time was 2.61 secs (range: 0.694–19.01 secs). There was a marked difference in response times for the two tasks. Average response time for the undirected task was 3.0 secs (range: 0.694–13.49 secs), compared to 2.22 secs (range: 0.709–19.01 secs) for the directed task. This provides evidence that the introduction of a specific task facilitated the wayfinding decision.

Connectivity. Results show that participants tended to choose the more connected street. The null hypothesis is that half of the decisions made by each participant would follow the more connected street. In all four measures of connectivity however, the number of decisions were above the expected value of 14 (table 1). More decisions could be modeled using measures of Integration than Choice; and for both Integration and Choice more decisions could be modeled using global measures over local measures.

Table 1. Average number of decisions per participant made in both path search tasks that can be modeled according to the space syntax measures of Integration and Choice at global and local scales

	Integration r=n	**Integration r=100m**	**ln Choice r=n**	**ln Choice r=100m**
Av. no of decisions	21.53	19.93	19.67	15.13
p value	<0.01	<0.01	<0.01	<0.02

The results show a difference in choices made according to the spatial task. Participants responded to the first half of stimuli according to the undirected path search task, thus results are tested against the expected value of 7. All connectivity measures recorded an average number of decisions above that expected value (table 2). Fewer decisions could be modeled according to Choice than Integration. The greatest number of decisions in the undirected path search task could be modeled using the local Integration measure.

Table 2. Average number of decisions per participant made in the undirected spatial task that can be modeled according to the space syntax measures of Integration and Choice at global and local scales

	Integration	**Integration**	**ln Choice**	**ln Choice**
	r=n	**r=100m**	**r=n**	**r=100m**
Av. no of decisions	9.333	10.0	8.7333	7.4667
p value	< 0.01	< 0.01	<0.01	0.209

More decisions could be modeled using the space syntax measures in the directed than the undirected path search task (table 3); in particular both global measures reflected a larger number of decisions. Results suggest that the global integration measure is particularly effective when examining wayfinding in a directed path search task.

Table 3. Average number of decisions per participant made in the directed path search task that can be modeled according to the space syntax measures of Integration and Choice at global and local scales

	Integration	**Integration**	**ln Choice**	**ln Choice**
	r=n	**r=100m**	**r=n**	**r=100m**
Av. no. of decisions	12.2	9.9333	10.9333	7.666
p value	<0.01	<0.01	<0.01	<0.05

Controls. Testing whether there was any Left/Right bias, the results show that participants chose the same path alternative, regardless of whether it was shown on the left or right hand side. On average, each participant chose the same path alternative in the mirrored image 11.53 times (p<0.01), tested against an expected value of 7. This level of consistency suggests that participants were not making random choices.

Attractors affected the decisions made; of the 120 decisions made in the presence of attractors, 110 (91.67%) followed them. This was especially marked in the case of people (96.67%), and high also for traffic (90%). Although the sample size for the controls is small, the data seems to relate a clear picture on the role of light and attractors in the experiment.

Results suggest that participants' choices were not determined by lighting conditions. Of the 60 decisions made in these conditions, 52 (86.67%) were not affected by light conditions.

Recall Task. Results show a high success rate in the recall task. Of the 75 decisions made in the recall section of the experiment, 67 were correct in their response. That is, with 89% accuracy, participants successfully recalled whether a stimulus had been previously shown in the experiment. Average response time was lower in the recall task compared to the spatial tasks. On average, a successful recall response was recorded at 1.68 secs (range: 0.703–7.91 secs), which is a far quicker response time. This reflects the fact that fact that the stimulus had already been shown.

3.2 Eye Tracking Data

Spatial Tasks. Results from the eye tracking study offer a greater level of detail in the analysis of the choices made.

On average, there were 4.29 fixations per participant per stimulus. There was no significant tendency to look at the eventually chosen path first (14.53 decisions per participant; p=0.228 in a one-sample t-test for a number of decisions above the expected value of 14). This is because there is substantial evidence to support the hypothesis that participants tended to look left first, irrespective of the stimulus shown; on average each participant look left first 20 times (p<0.01). A possible explanation for this could be that British traffic customs encourage pedestrians to look left first before crossing the street. Of more relevance is that the final fixation was often directed towards the chosen path choice; this occurred on average 18.93 times per participant (p<0.01). In addition, subjects tended to spend more time looking at the path they eventually chose as opposed to the discarded path alternative (0.87 secs vs. 0.67 secs). Further evidence to support the hypothesis that subjects' gaze behavior tended to be directed towards the path alterative they eventually chose, can be found in the higher number of fixations to that side (2.39 vs. 1.88) as well as longer fixations (0.35 secs vs. 0.3 secs).

In between the initial and final fixations, subjects crossed the center line 1.91 times on average. This shows that subjects were evaluating the path alternatives. It also suggests that the participants were able to respond to the study as, on average, they did not contemplate the two path alternatives for a lengthy period of time. Further evidence to support this statement comes from the average fixation duration (0.36 secs) and average number of fixations (4.29), which appear low when comparing with other studies. This, however, is to be expected in an eye tracking study using photographs as stimuli and strengthens the hypothesis that subjects were responding to the study as though they were physically located in the environment itself. Indeed the post-study interview recorded many comments to the same effect, confirming the benefit of working from real-world stimuli.

Thus a general trend for participants' viewing behavior for each stimulus of the study can be described as looking left initially, crossing the center line almost twice per stimulus, and viewing the eventually chosen path last, having spent more time in the eventually chosen half.

Initial analysis using isovist-related techniques suggest that measures of spatial complexity provide an adequate model of the gaze bias. Promising measures are i) the number of vertical edges in a scene, ii) the depth of view and iii) the amount of visible floorspace. Further research will be able to provide a model of the gaze bias data deriving from the geometric properties of the environment. Significantly, the location of fixations in the control scenes does not alter the general viewing pattern. This suggests that low-level stimulus properties are not the determining factor in the gaze bias data.

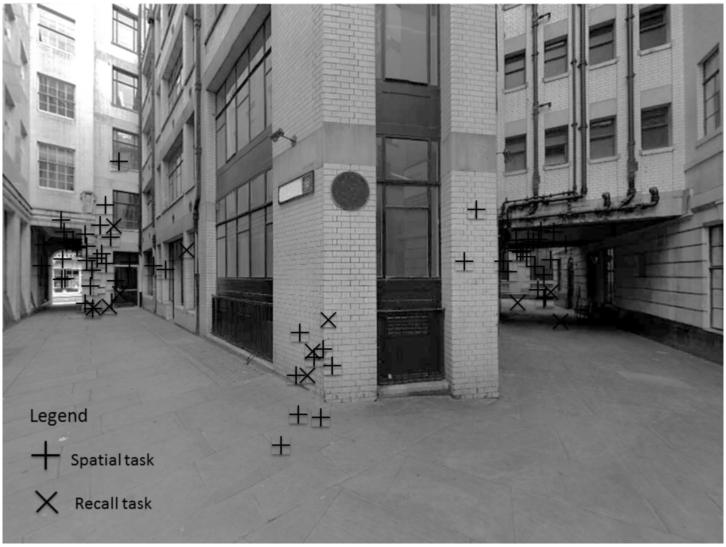


Fig. 7. Example fixation data for one stimulus, comparing the fixations for the spatial and recall tasks

Recall Task. The average fixation duration was 0.29 secs, shorter than in the spatial tasks (0.36 secs), reflecting the fact that the recall task was (necessarily) subsequent to the spatial tasks. As in the spatial tasks, there was a tendency for the first fixation to be directed towards the left path alternative; 62% of initial fixations were towards the left. On average, the time-course pattern for the recall task involved crossing the center line 1.61 times, evidence that compared to the spatial task there was less need to double-check the first fixated path alternative.

Initial analyses of the fixation locations show that the change of task did not affect where participants directed their attention. As seen in the example stimulus above, there is no great variation of the direction of attention during the recall task compared to the spatial tasks. This suggests that task-related influences are not the main driver of fixation location.

4 Discussion

Results from the study promote the use of real world photographs for wayfinding studies using eye tracking. This study analyzed one type of wayfinding behavior, path search behavior, with no information pertaining to route or survey knowledge.

One aim of the study was to test the role of spatial configuration during wayfinding. This was tested using a distinct form of analysis: the number of decisions made that favored the more connected street. Connectivity was measured using the space syntax model of street connectivity; specifically global and local (100m) Integration and Choice were used. Two thirds of the behavioral responses in the spatial tasks corresponded with space syntax measures of spatial configuration. This is an important finding as it lends weight to the relevance of the space syntax model for decisions made by the individual. The space syntax model is based solely on urban grid connectivity; the path search tasks used in the study did not provide any information relating to route or survey knowledge. Thus the study is well-placed to address the role of topological connectivity in individual spatial decision-making. Further work would provide more detailed understanding of the phenomenon; for example, the discrepancy in topological connectivity between the two path alternatives could be the basis from which the stimuli set is based.

The paper discusses the merit of the space syntax measures of Integration and Choice at local and global scales; results in this study provide empirical data confirming the findings of previous studies (see [34] and [6]). Specifically, global integration proves to be a relevant measure; this is crucial finding because it suggests that we understand the global structure of a space from a local viewpoint, a tenet of space syntax theory underlying the concept of intelligibility. Thus this paper provides initial empirical evidence in support of the theory that the intelligibility of an environment stems from the relation between the local and global scales. Further empirical evidence is required. One avenue to investigate this would be to test the relationship between both depth of view and visual connectivity with global integration, variables identified in a previous study [34], in the fixation data.

The study identified the effect of a specific task on path search behavior. The introduction of the taxi rank task i) re-enforced the number of decisions made in the undirected path search scenario that favored the more connected street and ii) required less time for a response to be made. It is suggested therefore that future studies need not introduce a specific task when looking at path search behavior; the research could focus on the most innate form of wayfinding would be examined, without the subjectivity involved when interpreting a task.

Controls tested for a number of conditions. Any Left/Right bias was tested by including a mirrored version of each stimulus in the final set; on average participants chose the same path alternative in 82% of cases, regardless of whether that path alternative was shown on the left or right hand side. The role of attractors was controlled for by having people and traffic present in some of the stimuli. Results confirm those of previous research where wayfinding behavior is affected by attractors; this was especially marked when people were present.

Results from the gaze bias data suggest that participants' viewing behavior cannot be sufficiently explained by looking either at the low-level properties or at task-related influences. In the control conditions for light, there was no substantial variation in viewing behavior between the different conditions. This shows that lighting conditions cannot be used alone to explain the viewing patterns and suggests that a model of fixation data based solely on low-level stimulus properties would not be adequate. No substantial change in viewing behavior was seen in the change of task. Participants tended to view the stimuli in a similar fashion whether they were asked to execute a spatial decision-making task or a recall task. This shows that task-related influences cannot be used alone to explain viewing behavior. Taken together, these two characteristics expose the relevance of visuo-spatial measures to model the fixation data. It is hoped that future research will contribute to a predictive model for wayfinding behavior based on the structural properties of the environment.

Acknowledgements. I am grateful to Jan M. Wiener, Christoph Hölscher, Ruth Conroy Dalton and Alan Penn for their guidance; to Muki Haklay for use of the eye tracker; and to the EPSRC for their generous Studentship.

References

1. Montello, D.: Spatial Cognition. In: Smelser, N.J., Baltes, P.B. (eds.) International Encyclopedia of the Social & Behavioral Sciences, pp. 14771–14775. Pergamon (2001)
2. Weisman, J.: Evaluating architectural legibility. Wayfinding in the built environment. Environment and Behavior 13, 189–204 (1981)
3. Peponis, J., Zimring, C., Choi, Y.K.: Finding the Building in Wayfinding. Environment and Behavior 22, 555–590 (1990)
4. Haq, S., Zimring, C.: Just down the road a piece - The development of topological knowledge of building layouts. Environment and Behavior 35, 132–160 (2003)
5. Hölscher, C., Brösamle, M., Vrachliotis, G.: Challenges in multilevel wayfinding: a case study with the space syntax technique. Environment and Planning B: Planning and Design 39, 63–82 (2012)
6. Emo, B., Hölscher, C., Wiener, J.M., Dalton, R.C.: Wayfinding and spatial configuration: evidence from street corners. In: Greene, M., Reyes, G., Castro, A. (eds.) Proceedings of the Eight International Space Syntax Symposium, Santiago de Chile (2012)
7. Siegel, A., White, S.: The development of spatial representations of large-scale environments. In: Reese, H.W. (ed.) Advances in Child Development and Behavior, vol. 10, pp. 9–55. Academic Press, New York (1975)
8. Gärling, T., Böök, A.: Cognitive mapping of large-scale environments: the interrelation of action plans, acquisition and orientation. Environment and Behavior 16, 3–30 (1983)
9. Hölscher, C., Meilinger, T., Vrachliotis, G., Brösamle, M., Knauff, M.: Up and down the staircase: wayfinding strategies in multi-level buildings. Journal of Environmental Psychology 26, 284–299 (2006)
10. Appleyard, D.: Styles and methods for structuring a city. Environment and Behavior 2, 100–117 (1970)
11. Zacharias, J.: Path choice and visual stimuli: stimuli of human actiivity and architecture. Journal of Environmental Psychology 21, 341–352 (2001)
12. Dalton, R.C., Troffa, R., Zacharias, J., Hölscher, C.: Visual information in the built environment and its effect on wayfinding and explorative behavior. In: Bonaiuto, M., Bonnes, M., Nenci, A., Carrus, G. (eds.) Urban Diversities – Environmental and Social Issues, pp. 6–76. Hogrefe (2011)

13. Hegarty, M., Richardson, A.E., Montello, D.R., Lovelace, K., Subbiah, I.: Development of a self-report measure of environmental spatial ability. Intelligence 30, 425–447 (2002)
14. Wiener, J.M., Büchner, S., Hölscher, C.: Taxonomy of Human Wayfinding Tasks: A Knowledge-Based Approach. Spatial Cognition & Computation 9, 152–165 (2009)
15. Bafna, S.: Space Syntax: A Brief Introduction to Its Logic and Analytical Techniques. Environment and Behavior 35, 17–29 (2003)
16. Harary, J.: Graph Theory. Perseus Books (1969)
17. Hillier, B., Iida, S.: Network and Psychological Effects in Urban Movement. In: Cohn, A.G., Mark, D.M. (eds.) COSIT 2005. LNCS, vol. 3693, pp. 475–490. Springer, Heidelberg (2005)
18. Penn, A.: Space syntax and spatial cognition – or why the axial line. Environment and Behavior 35, 30–65 (2003)
19. Dalton, R.C., Hölscher, C., Turner, A.: Understanding space: the nascent synthesis of cognition and the syntax of spatial morphologies. Environment and Planning B: Planning and Design 39, 7–11 (2012)
20. Wiener, J.M., Hölscher, C., Büchener, S., Konieczny, L.: Gaze behaviour during Space Perception and Spatial Decision Making. Psychological Research (in press)
21. Pinelo, J.: Towards a Spatial Congruence Theory. How spatial cognition can inform urban planning and design. PhD Thesis, University College London (2010)
22. Henderson, J.: Human gaze control during real-world scene perception. Trends in Cognitive Sciences 7, 498–504 (2003)
23. Itti, L., Koch, C.: Computational Modelling of Visual Attention. Nature Reviews Neuroscience 2, 194–203 (2001)
24. Torralba, A., Oliva, A., Castelhano, M.S., Henderson, J.M.: Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological Review 113, 766–786 (2006)
25. Yarbus, A.: Eye movements and vision. Plenum, New York (1967)
26. O'Neill, M.: Evaluation of a conceptual model of architecural legibility. Environment and Behavior 23, 553–574 (1991)
27. Hardy, A.C.: Landscape and Human Perception. In: Murray, A.C. (ed.) Methods of Landscape Analysis. Landscape Research Group, London (1967)
28. Tandy, C.: The Isovist Method of Landscape Survey. In: Murray, A.C. (ed.) Methods of Landscape Analysis. Landscape Research Group, London (1967)
29. Benedikt, M.L.: To take hold of space: isovists and isovist fields. Environment and Planning B: Planning and Design 6, 47–65 (1979)
30. Turner, A., Doxa, M., Sulivan, D.O., Penn, A.: From isovists to visibility graphs: a methodology for the analysis of architectural space. Environment and Planning B: Planning and Design 28, 103–121 (2001)
31. Franz, G., Wiener, J.M.: From space syntax to space semantics: a behaviorally and perceptually oriented methodology for the efficient description of the geometry and topology of environments. Environment and Planning B: Planning and Design 35, 575–592 (2008)
32. Wiener, J.M., Franz, G., Rossmanith, N., Reichelt, A., Mallot, H.A., Bülthoff, H.H.: Isovist analysis captures properties of space relevant for locomotion and experience. Perception 36, 1066–1083 (2007)
33. Meilinger, T., Franz, G., Bülthoff, H.H.: From isovists via mental representations to behaviour: first steps towards closing the causal chain. Environment and Planning B: Planning and Design 39, 48–62 (2009)
34. Emo, B.: The Visual Properties of Spatial Configuration. In: Dara-Abrams, D., Conroy Dalton, R., Hölscher, C., Turner, A. (eds.) Environmental Modeling: Using Space Syntax in Spatial Cognition Research. Proceedings of the Workshop at Spatial Cognition 2010, Mt. Hood, Oregon. SFB/TR 8 Report No. 026-12/ (2010)

Relocating Multiple Objects during Spatial Belief Revision

Leandra Bucher and Jelica Nejasmic

Experimental Psychology and Cognitive Science, Justus Liebig University Giessen, Germany
{leandra.bucher,jelica.nejasmic}@psychol.uni-giessen.de

Abstract. Reasoners need to revise their beliefs about the state of the world when confronted with contradicting evidence. In the spatial context, belief revision is assumed to be accomplished by variation of initially constructed spatial mental models. The revision process includes decisions about which part of a model to retain and which one to modify. Usually, there are several alternatives for model variation that re-establish consistency within belief sets. Frequently, these alternatives are logically equivalent. Nevertheless, human reasoners show clear preferences for certain alternatives. The assumption is that the preferences result from the application of principles that are cognitively more economic compared to others. In two experiments, we investigate how the number of objects involved in model variation processes affects preferences in model variations during spatial belief revision. We discuss whether the results can be explained in terms of cognitive economy.

Keywords: Spatial reasoning, relational reasoning, mental models, model variation, belief revision.

1 Introduction

Multiple sources provide us with information about the state of the world. Sources differ in many ways, for instance concerning reliability, familiarity, and trustworthiness. Additionally, the context, information is concerned with, is more or less familiar, information itself more or less plausible, probable, important, and so on. Some pieces of information simply confirm what we already know or belief to know about the world and some increase our knowledge base. However, some conflict with what we believe. That entails that these pieces of information are not addable to sets of already existing beliefs in a monotonic way but cause the need for giving up existing beliefs to re-establish consistency within these sets [1]. Belief sets need to be *updated* when reliable information indicate that the world has changed. They need *revision,* when information surface which reliably indicate that some of our existing beliefs are not maintainable because they are obviously not true [1-3]. Here is an example:

Imagine you have received the description of a route to your friend´s house that includes the statements:

(1) “The filling station is on the left hand side, opposite the bakery.”
(2) “At the same side as the bakery you find a supermarket.”

C. Stachniss, K. Schill, and D. Uttal (Eds.): Spatial Cognition 2012, LNAI 7463, pp. 478–491, 2012.

Viewing it from the front, you would expect the arrangement of the buildings to look like

(3)	Filling station	Bakery
		Supermarket

In order to recheck the description you consult Google Earth. However, unfortunately, Google Earth only provides you with a picture of a very small excerpt of the scene, you actually wanted to look at; and rather confusing, what you see could be described as:

(4) "Supermarket in front of filling station"

You take the picture you look at and thus the statement (4) as fact because it is putatively more reliable than the description you have received. However, this implies that you need to revise your initial assumptions about the relations of the buildings. To do so, you are required to retract at least either information provided by statement (1) or (2) and/or to modify the spatial arrangement of the buildings (3) you had in mind. For the latter, there are two alternatives to go with both, taking into account the fact and preserving as much of the initial information as possible:

(5)	Filling station	Bakery
	Supermarket	

or

(6)	Bakery
	Filling station
	Supermarket

Arrangements (5) and (6) are modifications of the initial arrangement. Arrangement (5) is obtained by relocation of the "Supermarket" within the initial arrangement; it preserves the information conveyed by statement (1). Arrangement (6) is obtained by relocation of the "filling station"; it preserves information from statement (2). Both arrangements involve the same amount of changes in terms of relocated objects (one, respectively). Also, the same amount of information from the descriptions is preserved/retracted in both alternatives. Thus, the amount of changes does not help with the choice of which assumption about the arrangement of the buildings should be preferred over the other. From a logical point of view, arrangements (5) and (6) are equivalent variations of the initial arrangement. Thus, logic does not help with the choice of how to modify the initial arrangement, either. Both, the amount of changes and logic indeed would leave reasoners undecided or confused. However, studies that looked at spatial reasoning suggest that human reasoners indeed clearly prefer certain alternatives over others [4-9].

The present paper is concerned with the factors that lead to preferences in spatial reasoning during the revision of spatial beliefs. The general questions are: how are spatial relations processed during revision? And, what guides the revision process? Subsequently to summarizing theoretical assumptions and empirical findings about spatial relational reasoning and spatial belief revision, we present two experiments that investigated revision of spatial beliefs with two-dimensional arrangements of four objects. Alternatives to revise these arrangements differed in the number of objects (one vs. two) involved in the revision process. The specific question was: What role does the number of objects, relocated in order to modify an initial arrangement, play during revision? Does it provide a guiding factor in the course of revision in the sense that reasoners prefer to keep the changes in terms of relocated objects little? In that case, reasoners should prefer relocating single and avoid relocating multiple objects.

1.1 Relational Reasoning with Spatial Mental Models

Many studies show that during reasoning with spatial relations, the objects and relations are represented in spatial mental models. Relational reasoning is describable in distinct phases: in the first phase, reasoners *construct* spatial mental models that reflect the information provided by the premises. This model allows for the *inspection* in search of information that is not explicitly given in the premises. Inspection phase is followed by a *variation* phase. During this phase, reasoners vary preferred mental models in order to find alternative interpretations of the premises. However, this model variation takes place only if it is required by the specific problem. If this is not the case people only rarely search for counterexamples that refute a putative conclusion. This often leads them into errors and logically invalid inferences [4-5]. Vast empirical evidence corroborates the notion that construction and inspection of spatial mental models provide the basis of relational reasoning [5], [10-19].

The question, the present work focuses on is how reasoners deal with spatial information that partly conflicts with information they have received beforehand and that run counter their beliefs about the spatial arrangements of objects. To take contradicting information into account with the aim to re-establish consistency within a certain belief set, reasoners need to detect inconsistencies between a former description and a new piece of information. Johnson-Laird, Legrenzi, & Girotto (2004) and Johnson-Laird, Girotto, & Legrenzi (2004) describe the ability of reasoners to detect inconsistencies in terms of model inspection. They conclude that, in general, reasoners detect conflicts quite reliably [20-21]. Inconsistency detection implies the detection of a conflict between previously given information with a new piece of information. Given the new piece of information is an incontrovertible fact, reasoners need to decide which piece(s) of information, initial beliefs are based upon conflict with this fact. Frequently, there are multiple alternatives for re-establishing consistency within belief sets and logic does not provide criteria for the decision which information to retract and which to retain in the course of revision.

1.2 Variation of Spatial Mental Models during Revision

It is assumed that reasoners base their revisions of beliefs about the relations of objects in space on a variation of initially constructed spatial models [4-10], [12]. The revision phase is deemed as a distinct phase, following construction and subsequent inspection (implicating inconsistency detection) of a spatial mental model [10].

The process of model variation itself has been shown to be guided by specific factors. With verbal descriptions expressed in a binary relations with r(X,Y), it has been shown that it relies on cues provided by conflicting information [10], [22]. A binary relation holds semantic cues that result from the sematic distinction of X as the “to-be-located object” (LO) in contrast to Y as the “reference object” (RO). The asymmetry of the two arguments (LO, RO) specifies the location of the LO relative to the known location of the RO [23]. Several empirical findings corroborate this assumption, for instance, findings in studies on the integration of new spatial information provided by binary relations into already existing spatial models [24-27]. In his theory, Logan (1994, 1995) proposes that attention is turned to a certain region by linguistic cues provided by these relations [28-29]. Attention moves from a statement´s RO to the region the LO can be expected. The findings converge to that effect that reasoners consider an RO´s position as fixed while the LO is flexible and locatable relative to the RO´s position.

For the variations of horizontal linear arrangements, the following finding concerning reasoners´ preferences is characteristic [8-10]:

Initial arrangement	A B C	
Contrafact	C is left of A,	*with C as the relation´s LO*
Preferred varied arrangement:	C A B	

Note that the logical equivalent (non-preferred) alternative for variation of the initial arrangement by relocating the contrafact´s RO (here: A) results in the revised arrangement:

B C A.

Basically, the position of objects as rearranged during the variation process of spatial belief revision is guided by information provided by the contrafact.

1.3 The Number of Objects and Relational Complexity

Here we present two experiments which investigated revision processes with relations more complex compared to relations of three objects Verbal descriptions were presented that described the arrangements of four objects in a two-dimensional layout, such that the first three objects were aligned horizontally and the fourth object was related to one of the outer positioned objects of the horizontal arrangement. Participants´ task was to construct the arrangement from the description and present the solution either by drawing (experiment 1) or by choosing the correct arrangement from two presented arrangements (experiment 2). Subsequently, participants were asked to judge whether a statement (binary relation about the two objects positioned at the outmost locations of the horizontal arrangement) that was introduced as “fact” about the arrangement at hand was consistent with previous information or not. In case of inconsistency detections, participants were asked to revise their assumptions about the relations of objects among each other by taking into account the fact´s spatial information. In experiment 1, participants were asked to sketch relations of the objects by drawing. In experiment 2, they were asked to choose between two alternative arrangements.

Our hypotheses concerning fundamentals underlying the revision process are:

- revision is accomplished by *variations* of initial arrangements.
- variations incorporate the fact´s information while as much of the initial information as possible is preserved.

Experiment 1 tested these assumptions in a situation allowing participants to sketch the objects´ relations unconfined and to generate their drawings freely. Given, reasoners vary initial models, the following factors might influence variation: semantic cues provided by contrafactual information or the number of objects that need to be relocated in order to regain consistency. When the spatial information provided by contrafactual statements is vital, rather than factors adherent to arrangements, the basic principle applied in variation should rely on semantic cues provided by the contrafact. The application of this principle would lead to the following observation:

- model variation is preferably based on the relocation of the fact´s LO relative to its RO as compared to the relocation of the fact´s RO relative to its LO.

When the number of objects relocated during revision might provide a guiding factor for variation, in the sense that reasoners prefer keeping the changes (number of relocated objects) little, this would lead to the following observation:

- model variation preferably involves the relocation of one object compared to two objects.

Reasoners´ preferences were examined in both experiments. Additionally, experiment 2 intended to provide a closer look at processing times related to variation. The assumption is:

- variation processes that involve multiple objects take longer compared to single objects.

The number of items, chunks, or units of information has been suggested to increase complexity [30]. With relational processing the pivotal role for determining complexity is played by relational complexity [31-32]. Accordingly, increased complexity of ternary or quaternary relations compared to binary relations results from interactions of the components adherent to a problem, e.g. the number of objects, relations etc. The more complex a problem the more processing capacity is needed or, differently phrased, the more cognitive resources are demanded. Processing capacity means, the amount of information stored to be processed later [33]. It is often referred to as "working memory capacity" with the working memory maintaining the information which is processed by the central executive. Allocation of cognitive resources and thus the demand for processing a certain task can be measured, e.g. by the decrement of performance of more difficult tasks that require more resources compared to less difficult tasks that require less. [31], [34]. Solutions of or decisions on problems should take longer with increased demands on cognitive resources as caused by increased complexity. The question is whether cognitive economy - reflected by the application of parsimonious principles in reasoning - can explain reasoners´ preferences.

2 Experiment 1: Drawing of Spatial Arrangements

2.1 Participants

24 participants (6 male; age: *M* = 23.92; *SD* = 5.00), all students (among them 2 students of psychology) from the University of Giessen, gave written informed consent to participation. Subjects were tested in small groups (n = 4-7) and paid at a rate of 8 Euro per hour.

2.2 Materials, Procedure, and Design

Verbal descriptions of two-dimensional spatial arrangements of four small, equal-sized, disyllabic-termed objects, belonging to either one out of two categories (fruits or tools) were presented. The descriptions consisted of three statements (premises), presented in a sequential manner with display duration of 10s each. The premises contained the relations "left of" and "right of" (1^{st} and 2^{nd} premises) and the relations "above" and "below" (3^{rd} premise). The occurrence of the relations ("left of" and right of" in 1^{st} and 2^{nd} and "above" and "below" in 3^{rd} premises) was counter-balanced across the experimental problems, such that each combination occurred equally often. An example description is provided below:

1st premise:	"Apple left of mango.
2nd premise:	"Pear right of mango."
3rd premise:	"Kiwi below pear".
Resulting in the arrangement:	Apple Mango Pear
	Kiwi

In half of the problems the fourth object was attached to the object at the outer left position in the linear arrangement, in the other half it was attached to the outer right positioned object.

The description was followed by the prompt "Please sketch the arrangement of the objects." with display duration of 20s, allowing the participants to sketch the arrangement. The prompt "Please turn the page." with a duration of 3s and a blank slide with the duration of 2s was shown before a fourth statement (fact) was presented for 10s. The fourth statement provided information about the relation between the two of the objects positioned at the two outer positions of the initially described arrangement (in the example above: Apple, pear). Facts (presented in red letters to contrast them from the initially presented premises which were black) were either consistent (in half of the problems) or inconsistent (in the other half) with the initial arrangement. Fact relations were "left of" (in half of the problems) and "right of" (in the other half). See an example below:

Consistent fact:	"The apple is left of the pear."
Inconsistent fact:	"The pear is left of the apple."

In half of the problems the fact´s semantic structure implied that the fourth object of the initial arrangement was attached to the object that was the to-be-located object (LO) of the fact. In the other half, the fourth object of the initial arrangement was attached to the object that occurred as the reference object (RO) in the fact. See an example below:

```
Initial arrangement:    Apple Mango Pear
                                    Kiwi
```

Inconsistent fact (1): "Pear left of apple.", with "pear" as the relation´s LO
Inconsistent fact (2): "Apple right of pear.", with "pear" as the relation´s RO

Note that the kiwi (as the fourth object) is attached to the pear which occurs as the LO of the inconsistent fact (1) and as the RO of the inconsistent fact (2).

The fact was followed by the prompt "Please sketch the arrangement of the objects." with display duration of 20s and subsequently, the prompt "Please turn the page." with duration of 3s. There was a blank slide presented for 2s before the next problem was shown. 32 experimental problems were presented, preceded by four practice problems (not analyzed).

Descriptions were provided using Microsoft PowerPoint (Version 2007) running in the windows environment XP on a standard personal computer. PowerPoint slides were presented on a big screen via video projector. Participants used a pencil to individually write down the constructed arrangements into specially prepared booklets.

Participants were instructed to draw the object arrangement according to the description. The instruction included the hint that they could not be entirely sure whether the information about the object arrangement was true. And that they would be presented with a fourth statement which would provide information about the arrangement at hand that has to be taken as a fact and being incontrovertible as such. Participants´ task was to follow the prompts, i.e. sketching the object arrangements into the booklet and turning the page, when instructed by the presented slides likewise.

Drawings of each participant were analyzed after the experimental session. Percentage values for correctly drawn arrangements were calculated. Of special interest were the drawings, generated after the information of the fact was taken into account. There were two alternatives for revising the initial arrangement in order to take into account the fact information while preserving as much of the initially provided information (from the premises): The alternatives can be described as follows:

1. The initial arrangement is revised by relocation of the fact´s LO;
2. The initial arrangement is revised by relocation of the fact´s RO

In half of the problems, the LO, in the other half the RO was attached by the fourth object. The question was whether the object attached to either the LO or the RO would affect the choice of a certain revision alternative. From a formal logical point of view, all alternatives were equivalent.

2.3 Results

Mean percentage rate of correctly drawn arrangements in the first step (construction phase) was 85.94% (*SD* = 17.31). Erroneous trials were excluded from further analysis. Mean percentage rate of correctly drawn arrangements in the second step (revision phase) was 84.23% (*SD* = 20.36). Erroneous trials were excluded from further analysis.

For the revision phase an ANOVA with the factors object (LO,RO) × object number (1,2) was conducted. There was a significant main effect of object [$F(1,23) = 130.32$; $p < .001$; $\eta^2_{part} = .850$]. Main effect of object number and the interaction were non-significant (*p*s > .15).

T-tests revealed that the drawings of revised arrangements were based significantly more often on relocations of LOs (*M* = 88.34%, *SD* = 16.45) compared to ROs (*M* = 11.66%, *SD* = 16.45). See also figure 1.

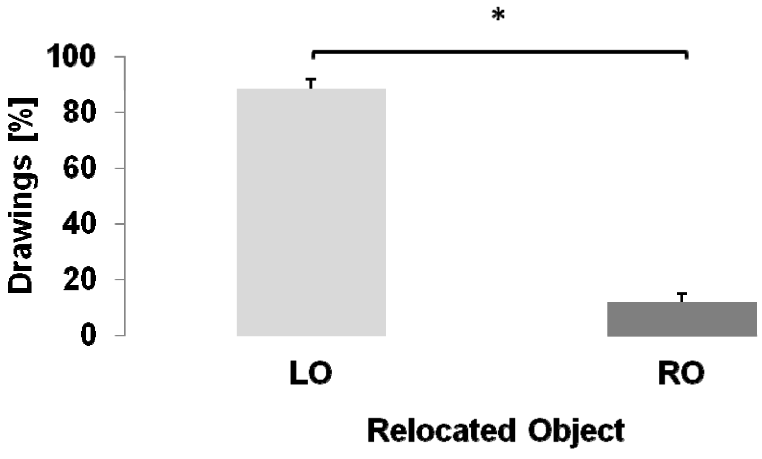


Fig. 1. Percentage rates for drawings of revised arrangements based on the relocation of to be located (LO) and reference objects (RO) are depicted. Error bars indicate standard errors.

Results indicate that revision was accomplished by varying initial arrangements and that variations were guided by semantic cues provided by the contrafact. Drawings revealed clearly reasoners´s preference for relocating facts´ LOs relative to ROs. LOs as single objects (in half of the problems) were comparably often subject to relocation compared to LOs attached with an additional object (in the other half of the problems). Attached objects were always relocated together with the relocated object, i.e. participants treated the pair of objects as a "chunk". There was no exception in none of the drawings. This finding implicates that information initially provided during the construction phase was preserved best possible.

With experiment 2, we assessed variation processes in terms of cognitive economy in a two-alternative forced-choice task.

3 Experiment 2: Choosing between Spatial Arrangements

3.1 Participants

23 participants (9 male; age: *M* = 23.22 *SD* = 3.06) all students (among them 8 students of psychology) from the University of Giessen, gave written informed consent to participation. Participants were tested individually and paid at a rate of 8 Euro per hour.

3.2 Materials, Procedure, and Design

All problems followed a tripartite structure with a layout description, inconsistency detection, and a revision part.

Comparable with experiment 1, in the layout description part, three premises (presented sequentially in a self-paced manner, only one visible on the screen at a time) described a two-dimensional layout of four objects. Objects and relations presented in the descriptions were comparable to those of experiment 1. Subsequently to premise presentation the correct spatial arrangement of the objects and an incorrect arrangement (correct arrangement inverted), see example below, were presented:

Correct arrangement:	incorrect arrangement:
Apple Mango Pear	Pear Mango Apple
Kiwi	Kiwi

Participants were instructed to choose the correct object arrangement (resulting from the description), presented on the left and right side of the computer screen, indicating their choice by pressing a left or right response button with the left or right hand, accordingly. Left and right locations for correct and incorrect arrangements were counter-balanced across the experiment. The number of correct decisions and corresponding decision times were recorded.

In the following inconsistency detection part a fourth statement (fact) was provided. Consistent (in half of the problems) and inconsistent facts (in the other half) resembled the facts presented in experiment 1. The participants were instructed to decide whether the fact was consistent or inconsistent, indicating their decision by pressing the respective response button (the side of the response-buttons “yes” and “no” were counter-balanced across participants) with the left or right hand, accordingly. Successful inconsistency detection and corresponding detection times were recorded.

The third part, the revision part followed only if the participant recognized a fact as inconsistent with the initial description. Participants were then instructed to revise their assumption about the objects´ relations by taking into account the inconsistent fact´s spatial information. They were presented with two object arrangements, on the left and the right side of the computer monitor. Participants were asked to choose that arrangement which matches their assumption, indicating their choice by pressing the left or right response button, respectively.

In fact, both arrangements were consistent with the presented “fact”. However, the arrangements differed with respect to the initial arrangement based on the relocation of the inconsistent fact´s LO or the fact´s RO. Again, as in experiment 1, in half of the problems, the fact´s LO and in the other half the fact´s RO was identical with the object attached by the fourth object in the initial arrangement.

Presentation locations of the arrangements were counter-balanced across the experiment. Revised arrangements chosen and corresponding revision times were recorded.

32 experimental, preceded by 6 practice trials (not analyzed) were presented in a random order. All stimuli were generated and presented using Superlab 4.0 (Cedrus Corporation, San Pedro, CA, 1999) with an RB-530 response box running on a standard personal computer connected to a 19’’-monitor.

3.3 Results

Overall, the participants’ performance was very high. The correct arrangements of objects (in the first phase) was chosen in 91.03% (*SD* = 9.76) of the trials within 3.27s (*SD* = 1.04). Erroneous problems were excluded from further analysis. In the second step (inconsistency detection) the performance was also high. Facts´ inconsistency with initial information from the premises was recognized correctly by the participants in 95.42 % (*SD* = 8.79) of all problems and took 3.03 s; (*SD* =.88) on average (erroneous trials were excluded from further analysis).

To examine the principle applied by reasonsers during the revision phase (the third phase), ANOVAs with the factors object (LO,RO) × object number (1,2) were calculated separately for percentage rates and revision times. ANOVA for the percentage rate revealed a main effect of object [$F(1,22) = 34.44$; $p < .001$; $\eta^2_{part} = .610$]. The main effect of object number and the interaction were non-significant ($p > .30$). Arrangements revised based on relocations of contrafacts´ LOs ($M = 70.41\%$; $SD = 16.68$) were chosen significantly more often compared to ROs ($M = 29.59\%$; $SD = 16.68$; $t(22) = 5.86$; $p < .001$).

ANOVA for the revision times revealed a marginally significant main effect of object number [$F(1,14) = 3.97$; $p = .066$; $\eta^2_{part} = .221$]. The main effect of object and the interaction object × object number were non-significant ($p > 30$). T-tests comparing decision time means for relocating one (M = 6.21s; SD = 3.41) compared to two objects (M = 7.62s; SD = 4.23) were marginally significantly lower ($t(23) = 2.07$; p = .051). Figure 2 provides an overview.

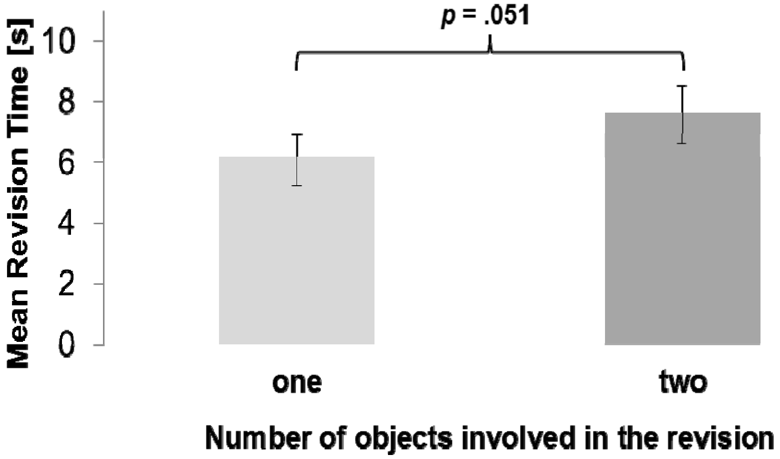


Fig. 2. Mean revision times for revisions involving one and two objects are depicted. Error bars indicate standard errors.

4 Discussion

Reasoning about spatial relations encompasses multiple reasoning abilities. For instance, reasoners are able to assume how spatial arrangements look like when provided with verbal descriptions about relations of objects. Further, they are able to infer information that is not explicitly provided by the descriptions [5-7]. The reasoning processes underlying these abilities can be described as distinct phases that involve the construction, inspection, and variation of spatial mental models [5-6]. Quite often new information run counter previously provided information. Inferences allow the detection of inconsistencies with conflicting information [20-21]. In case, contradicting information can be considered as reliable or incontrovertible, it needs to be taken into account. Accordingly, reasoners are required to retract initial information and revise current beliefs in order to re-establish consistency within belief sets [1-3].

The ability of belief revision is an important one but little is known about its process. In the spatial context, belief revision is assumed to be accomplished by variation of initially constructed spatial mental models. Variation itself is assumed to incorporate fact´s information while conserving as much of the initial information as possible [10], [22]. Usually, there are several alternatives to vary initially constructed models that allow re-establishing consistency within belief sets; and the alternatives are frequently logically equivalent. Thus, logic does not help reasoners with their decision which alternative to prefer. Nevertheless, human reasoners prefer certain alternatives above others [4-5], [7-9], [12]. Studies on revision of spatial beliefs that examined determinants of preferences for variations of simple one-dimensional horizontal arrangements of three objects, showed that semantic cues of contrafactual information - verbally provided as binary relational statements - guide variations and lead preferably to variations based on the relocation of to be located objects (LOs) relative to reference objects (ROs) [10], [22].

With the current work, we focused on the influence of the number of objects involved in model variation processes. From an economic point of view, changes to initially constructed models in the course of revision should be kept as minimal as possible. The assumption was that – given factors adherent to arrangements at hand influence revision - the number of objects attached to objects influence variation of initially constructed models. Variations involving the relocation of only one object should be preferably performed compared to variations involving the relocation of two objects. In two experiments, we examined whether reasoners follow this economic principle. However, in accordance with previous studies that show that reasoners focus on cues provided by inconsistent spatial information, we found preferences for variations that were based on the relocations of LOs relative to ROs in both experiments, when participants generated their revised arrangements freely (experiment 1) as well as when they choose from two alternative revised arrangements (experiment 2). Experiment 1 specifically corroborates fundamental assumptions about the revision process. Without exception, the participants' drawings indicated that revision was accomplished by variation of initially constructed models. As much of the initial information as possible was preserved.

With experiment 2, we looked deeper into the processing demand of variations as it emerges from relocating one compared to two objects. We assumed that higher complexity due to an increased number of objects increases processing difficulty, entailing increased processing demand. Variation that involved two, compared to one object, demanded higher cognitive resources. This was reflected by longer decision times. Results suggest that cognitively more economic principles are not necessarily guiding reasoning processes. LOs were preferably subject to relocation, regardless of whether their relocation implicated to relocate one (LO only) or two (LO and attached object) objects.

To summarize, our experiments show that variation is based on cues provided by contrafactual spatial information. Although complexity did not affect the basic principle, it prolonged the revision process.

Given the importance of relational reasoning and the revision of beliefs for everyday life and the fact that every-day life´s problems are rather complex, an important question remains: what exactly causes the additional demand on cognitive resources due to complexity during relational reasoning and in particular during the revision of spatial beliefs? Generally, the number of items, chunks, or units of information has been suggested to increase complexity [30]. Relational load however is deemed not to be ascribable to item load per se [31-32]. It remains to specify what exactly increases complexity: the number of objects, relations, or dimensions; and how do these factors interact and impact or finally even guide relational reasoning processes?

Acknowledgement. We thank Stefan Czoschke, Anja Gatzsche, Lilly Fehr, and Isabell Tapia-Leon for assistance with the experiments. Correspondence should be addressed to Leandra Bucher, University Giessen, Experimental Psychology and Cognitive Science, Otto-Behaghel-Str. 10F, 35394 Giessen, Germany (e-mail: Leandra.Bucher@psychol.uni-giessen.de).

References

1. Gärdenfors, P. (ed.): Belief revision. Cambridge University Press, Cambridge (1992)
2. Elio, R., Pellitier, F.J.: Belief change as propositional update. Cognitive Science 2, 419–460 (1997)
3. Wolf, A.G., Knauff, M.: The strategy behind belief revision: A matter of judging probability or the use of mental models. In: Love, B.C., McRae, K., Sloutsky, V.M. (eds.) Proceedings of the 30th Annual Conference of the Cognitive Science Society, pp. 831–836. Cognitive Science Society, Austin (2008)
4. Jahn, G., Knauff, M., Johnson-Laird, P.N.: Preferred mental models in reasoning about spatial relations. Memory & Cognition 35, 2075–2087 (2007)
5. Knauff, M., Rauh, R., Schlieder, C.: Preferred mental models in qualitative spatial reasoning: a cognitive assessment of Allen's calculus. In: Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society, pp. 200–205. Lawrence Erlbaum, Mahwah (1995)
6. Knauff, M., Rauh, R., Schlieder, C., Strube, G.: Continuity effect and figural bias in spatial relational inference. In: Proceedings of the Twentieth Annual Conference of the Cognitive Science Society, pp. 573–578. Lawrence Erlbaum Associates, Mahwah (1998)
7. Ragni, M., Knauff, M., Nebel, B.: A computational model for spatial reasoning with mental models. In: Bara, B., Barsalou, B., Bucciarelli, M. (eds.) Proceedings of the 27th Annual Cognitive Science Conference, pp. 1064–1070. Lawrence Erlbaum Associates, Mahwah (2005)
8. Krumnack, A., Bucher, L., Nejasmic, J., Knauff, M.: Spatial reasoning as verbal reasoning. In: Ohlsson, S., Catrambone, R. (eds.) Proceedings of the 32nd Annual Conference of the Cognitive Science Society, pp. 1002–1007. Cognitive Science Society, Austin (2010)
9. Krumnack, A., Bucher, L., Nejasmic, J., Nebel, B., Knauff, M.: A model for relational reasoning as verbal reasoning. Cognitive Systems Research 11, 377–392 (2011)
10. Bucher, L., Krumnack, A., Nejasmic, J., Knauff, M.: Cognitive processes underlying spatial belief revision. In: Carlson, L., Hölscher, C., Shipley, T. (eds.) Proceedings of the 33rd Annual Conference of the Cognitive Science Society, pp. 3477–3482. Cognitive Science Society, Austin (2011)
11. Johnson-Laird, P.N., Byrne, R.M.J.: Deduction. Erlbaum, Hove (1991)
12. Rauh, R., Hagen, C., Kuss, T., Knauff, M., Schlieder, C., Strube, G.: Preferred and alternative mental models in spatial reasoning. Spatial Cognition and Computation 5, 239–269 (2005)
13. Knauff, M., Johnson-Laird, P.N.: Visual imagery can impede reasoning. Memory & Cognition 30, 363–371 (2002)
14. Maybery, M.T., Bain, J.D., Halford, G.S.: Information-processing demands of transitive inference. Journal of Experimental Psychology 12, 600–613 (1986)
15. Schaeken, W., Johnson-Laird, P.N., d'Ydewalle, G.: Mental models and temporal reasoning. Cognition 60, 205–234 (1996a)

16. Carreiras, M., Santamaria, C.: Reasoning about relations: spatial and non-spatial problems. Thinking and Reasoning 3, 191–208 (1997); Clark, H.H.: Linguistic processes in deductive reasoning. Psychological Review 76, 387–404 (1969a)
17. Vandierendonck, A., De Vooght, G.: Working memory constraints on linear reasoning with spatial and temporal contents. Quarterly Journal of Experimental Psychology 50A, 803–820 (1997)
18. Schaeken, W., Girotto, V., Johnson Laird, P.N.: The effect of an irrelevant premise on temporal and spatial reasoning. Kognitionswissenschaft 7, 27–32 (1998)
19. Schaeken, W., Johnson-Laird, P.N.: Strategies in temporal reasoning. Thinking and Reasoning 6, 193–219 (2000)
20. Johnson-Laird, P.N., Girotto, V., Legrenzi, P.: Reasoning from inconsistency to consistency. Reasoning from inconsistency to consistency. Psychological Review 111, 640–661 (2004)
21. Johnson-Laird, P.N., Legrenzi, P., Girotto, V.: How we detect logical inconsistencies. Current Directions in Psychological Science 13, 41–45 (2004)
22. Krumnack, A., Bucher, L., Nejasmic, J., Knauff, M.: Efficiency and minimal change in spatial belief revision. In: Carlson, L., Hölscher, C., Shipley, T. (eds.) Proceedings of the 33rd Annual Conference of the Cognitive Science Society, pp. 2270–2275. Cognitive Science Society, Austin (2011)
23. Miller, G.A., Johnson-Laird, P.N.: Language and perception. Havard University Press, Cambridge (1976)
24. Nejasmic, J., Krumnack, A., Bucher, L., Knauff, M.: Cognitive processes underlying the continuity effect in spatial reasoning. In: Carlson, L., Hölscher, C., Shipley, T. (eds.) Proceedings of the 33rd Annual Conference of the Cognitive Science Society, pp. 1127–1132. Cognitive Science Society, Austin (2011)
25. Oberauer, K., Wilhelm, O.: Effects of directionality in deductive reasoning: I. The comprehension of single relational premises. Journal of Experimental Psychology: Learning, Memory, & Cognition 26, 1702–1712 (2000)
26. Oberauer, K., Hörnig, R., Weidenfeld, A., Wilhelm, O.: Effects of directionality in deductive reasoning: I. Premise integration and conclusion evaluation. Quaterly Journal of Experimental Psychology 58, 1225–1247 (2005)
27. Hörnig, R., Oberauer, K., Weidenfeld, A.: Two principles of premise integration in spatial reasoning. Memory & Cognition 33, 131–139 (2005)
28. Logan, G.D.: Spatial attention and the apprehension of spatial relations. Journal of Experimental Psychology: Human Perception and Performance 20, 1015–1036 (1994)
29. Logan, G.D.: Linguistic and conceptual control of visual spatial attention. Cognitive Psychology 28, 103–174 (1995)
30. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review 63, 81–97 (1956)
31. Halford, G.S., Wilson, W.H., Phillips, S.: Processing capacity defined by relational complexity: implications for comparative, developmental, and cognitive psychology. Behavioral and Brain Sciences 21, 803–865 (1998)
32. Phillips, S., Niki, K.: Separating relational from item load effects in paired recognition: temporoparietal and middle frontal gyral activity with increased associates, but not items during encoding and retention. Neuroimage 17, 1031–1055 (2002)
33. Hitch, G.: Developing the concept of working memory. In: Claxton, G. (ed.) Cognitive Psychology: New Directions. Routledge and Kegan Paul, London (1980)
34. Navon, D., Gopher, D.: Task difficulty, resources, and dual-task performance. In: Nickerson, R.S. (ed.) Attention and Performance VIII, pp. 297–315. Erlbaum, Hillsdale (1980)

Author Index