A STUDY ON HOW HUMANS DESCRIBE RELATIVE POSITIONS OF IMAGE OBJECTS

Xin Wang Pascal Matsakis Lana Trick[2] Blair Nonnecke Melanie Veltman

Department of Computing and Information Science,
[2]Department of Psychology
University of Guelph, Guelph, Ontario Canada, N1G 2W1
Fax: (519) 837-0323 E-mail: xin@uoguelph.ca Email: matsakis@cis.uoguelph.ca Email ltrick@uoguelph.ca
Email: nonnecke@uoguelph.ca Email: mveltman@uoguelph.ca

ABSTRACT

Information describing the layout of objects in space is commonly conveyed through the use of linguistic terms denoting spatial relations that hold between the objects. Though progress has been made in the understanding and modelling of many individual relations, a better understanding of how human subjects use spatial relations together in natural language to is required. This paper outlines the design and completion of an experiment resulting in the collection of 1920 spoken descriptions from 32 human subjects; they describe the relative positions of a variety of objects within an image space. We investigate the spatial relations that the subjects express in their descriptions, and the terms through which they do so, in an effort to determine variations and commonalities. Analysis of the descriptions determines that common elements of spatial perception do indeed exist between subjects, and that the subjects are quite consistent with each other in the use of spatial relations.

Keywords
Spatial relations, natural language, spatial cognition, human information processing

INTRODUCTION

Spatial information is understood and conveyed through the use of spatial relations, which describe how one object in a scene or an image is located in relation to some other object. Spatial relations have been studied in a number of different disciplines, including computer science, geographic information science, cognitive science and linguistics. Within these disciplines, spatial relations are generally considered to fit into one of three categories: topological, including relations like *OVERLAP* and *SEPARATE*; directional, including relations like *ABOVE*, *BELOW*, *RIGHT* and *LEFT*; and distance, including relations like *NEAR* and *FAR*. In natural language, relations such as these are referred to using a variety of different terms and phrases. Most spatial terms, including ones we are very familiar with, like *near* or *beside*, possess semantics which are far more nuanced than might be expected at first glance.

The perception of spatial relations is determined by many factors, including the point of observation and any intrinsic axes of image objects (Herskovits 1986); and any additional objects or associated context (Regier 1992). Additionally, the dimensionality of the space and objects in question determines the relations that can be used – in a 3D space, relations like *BEHIND* and *IN FRONT OF* may be appropriate. In addition to these factors, individual differences like gender (Linn & Petersen 1985) and handedness (Halpern 1986; Mark *et al.* 1995) may affect how one forms mental models of spatial phenomena and assigns meanings to spatial concepts. Anthropologist Hall (1966) found that a subject's experience of space, and hence perception of spatial relations, are affected by culture. This was confirmed by Montello (1995), in his critical discussion of the significance of cultural differences in spatial cognition. Based on these factors, two human subjects may perceive the same concept quite differently, and hence describe it differently (Mark *et al.* 1994; Mark *et al.* 1995; Worboys 2001). As early as 60 years ago, Whorf and Sapir (1940) proposed that language influences or constrains the way in which people think; this work is known as the Sapir-Whorf hypothesis. Although it is not clear whether or how such effects apply to spatial relations, there are significant distinctions in the use of spatial terms in different languages. A number of cognitive and linguistic scientists have informally described how spatial relations are expressed in natural language. Talmy (1983) suggested that in linguistic descriptions, the spatial disposition, i.e., the site and orientation, of one object, referred to as the *figure* (or *argument*), is always characterized in terms of one or more other objects selected from the remainder of the scene, referred to as the *ground* (or *reference*). The ground objects are used as a fixed reference from which the position of the figure is described. Talmy also pointed out that any natural language has only a limited number of words available for describing the spatial relations in an infinite number of spatial layouts, and each of these words actually represents a family of layouts that all share certain abstract characteristics.

Additionally, some research into the computational modelling of spatial relations has taken human perception into account. The most common method of capturing human perception is as follows: To begin, subjects are presented with a small number (usually less than 100) of images, referred to as *configurations*, containing basic shapes. Subjects are then given a set of spatial relations (usually less than 10), and for each configuration, they are required to either answer a series of yes or no questions (i.e., whether a given relation describes the configuration (Robinson 1990)), or to rate a list of spatial relations based on how well they describe the configuration individually (Gapp 1995; Wang & Keller 1997, 1999; Zhan 2002). The weaknesses in these methods are obvious. Firstly, the data used to train the system is collected by having all subjects describe a small number of configurations using the same relations, and consequently, only a small number of relations and configurations are applicable. Secondly, as pointed out by Landau (1996), the English lexicon of spatial prepositions numbers above eighty members, not considering terms used to describe compound spatial relations, or uncommon relations. Hence, it would be implausible for a given subject to test all of these relations for each configuration, yet any practices of limiting spatial relations to a given smaller set may bias the subject and subsequently, the results. Additionally, spatial relations are not independent from each other, rather, a variety of relations occur and interact in a given natural language description. We are more interested in which relations users refer to when describing an image, and how they use them together (i.e., in the context of image retrieval).

The goal of this research is to design and carry out an experiment to better understand how human subjects naturally describe the relative positions of objects in a series of

images. The primary research question is: when describing different *configurations*, what spatial relations do people refer to, and how (i.e., using what terms)? In order to design the experiment so that we capture human perception of a number of different relations in a variety of circumstances, we must collect descriptions of a large number of varied configurations, desirably more than 1000. Under this condition, obtaining enough information from one subject would prove a cumbersome task, and we investigate the use of data collected from multiple subjects. Even though cognitive studies show that individual differences in spatial perception exist among different people, we hypothesize that common elements of perception may be found among different subjects (if this were not the case, humans would not be able communicate spatial concepts to one another). Thus, two more questions arise: How significant are the variations between descriptions given by different subjects? What are the common elements of perception? To ascertain commonalities and variations in spatial perception, we are interested in measuring the consistency of use of spatial relations in descriptions given by numerous subjects; for any given configuration, our focus is on the presence of spatial relations in subjects' descriptions.

The remainder of this paper is organized as follows: Section 2 describes the experiment design. Section 3 outlines the collection of the descriptions and the extraction of relations from these descriptions. Section 4 discusses the results of analysis and Section 5 concludes the paper with discussion of limitations and future work.

EXPERIMENT DESIGN

Assume we want to design a computer system capable of providing linguistic descriptions of relative positions of image objects the way a given person would. A system like this would be of use in a number of practical applications. In many robot vision scenarios, the robot's understanding of its environment will include some representation of spatial information, and the robot must then communicate this to a human user. Similarly, some Content Based Image Retrieval (CBIR) systems use spatial information in indexing and natural language in searching. In both of these scenarios, training a system to provide accurate, human-like descriptions will increase the quality of the interaction. As mentioned previously, obtaining enough information for system training from one subject is not reasonable, and we investigate the concept of a prototypical perception, based on common elements found in descriptions from different subjects. If we can determine a prototypical perception, and train the system using this data, a later point in time, the system could be fine-tuned to one individual perception through training with one user.

Hence, the goal of this research is to determine any common elements of perception, and the significance of any variations between key elements of descriptions provided by different subjects. We aim to validate the existence of a prototypical perception, and gain a better understanding of what the elements of this perception are (i.e., the spatial relations and corresponding linguistic terms people commonly refer to). We also create and demonstrate an appropriate method for collecting natural language descriptions that could be used to train a system that would be able to learn and generate descriptions according to a prototypical perception.

Generating configurations

In an effort to simplify the problem and eliminate the influence of factors such as point of observation and dimension, two and only two 2D image objects are considered. The two objects are abstract shapes, with no intrinsic axes or context associated with them. An object pool containing 25 different shapes was created, and is illustrated in Figure 1. The shapes represent regular and non regular convex shapes (O1 to O10), and simple and more complex concave shapes (O11 to O25).

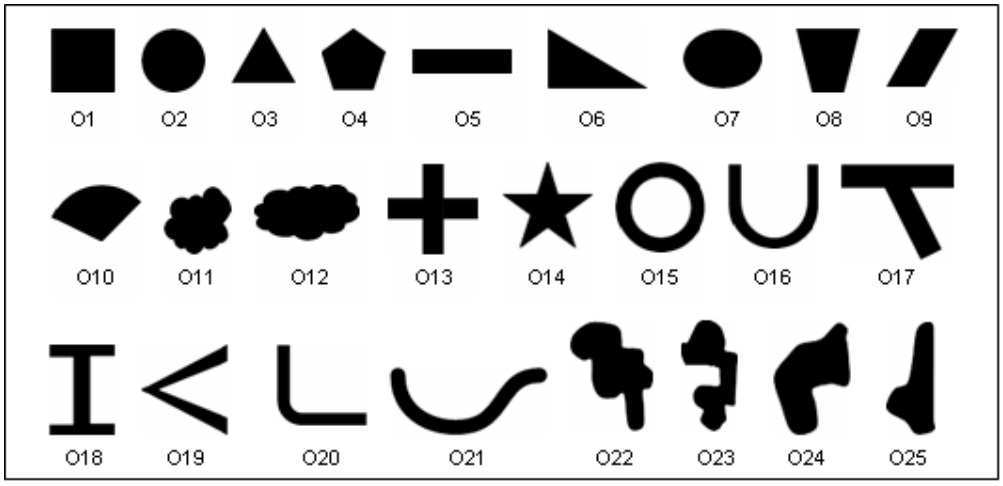

Figure 1. The Object Pool

Using these 25 shapes, configurations were then generated using the following method: First, we randomly draw two objects from the object pool, with replacement. Secondly, each object is zoomed by a random zooming factor μ between 30% and 300%, and rotated by a random angle θ. Note that the values of μ and θ may be different for the two objects. One object, selected randomly, is then coloured grey and the other is coloured black. Finally, the transformed objects are randomly placed inside the image space, a 500 x 500 white background, 20 times to create a set of 20 configurations containing the same objects. If there is an intersection between the two objects, it is coloured a dark grey (between the grey shade and the black). The above method is repeated to create 68 such configuration sets, and thus we have in total 1360 configurations for experimentation.

Describing Task

Because we are interested in natural language descriptions (i.e., we do not want to constrain the subjects by providing a check list of terms or relations), we must properly design the experiment and communicate the describing task to the subjects in such a manner as to ensure that they focus their descriptions on relevant spatial information. To achieve this, each subject is tasked with describing configurations in the context of a game that loosely emulates image retrieval, as illustrated in Figure 2. The game is described to the subject using the following scenario: “Imagine that you are playing a game with a friend. You have a set of configurations and your friend has another set. For each configuration in your set, there are one or more similar configurations in your friend's set. Now, imagine that you are on the phone with your friend. He/she cannot see your set, and you cannot see his/hers. Please describe the configuration shown on the computer screen so that your friend is able to find similar configuration(s) in his/her set." Note that terms like

size, shape, relative position, and absolute position are actually never pronounced, and no verbal example is given (only visual examples), thereby minimizing potential biases.

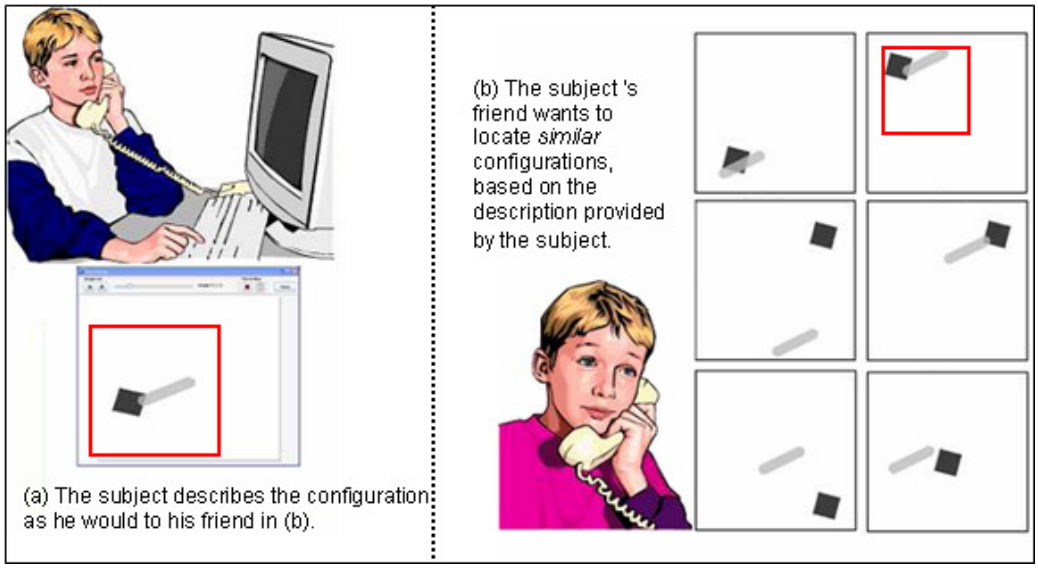

Figure 2. The Image Retrieval Game

Through learning the concept of the game, the subject is made aware of the following important characteristics of their task:

1. The objects are of the same size, shape, and colour in all configurations within a configuration set; with this in mind, the subject will presumably avoid describing the objects features, and instead will describe where the objects are.
2. For two configurations to be considered similar, the objects can be anywhere in the image, as long as their relative positions are the same, i.e., as illustrated in Figure 2, the objects may be shifted together in the space. Knowing this, the subject will presumably avoid providing information about the absolute positions of the objects, in favour of describing relative positions.
3. The relative positions of the objects in similar configurations may not be identical, just similar; with this in mind, the subject will presumably avoid excessively long or detailed descriptions.

COLLECTING DESCRIPTIONS AND EXTRACTING SPATIAL INFORMATION

Collecting Descriptions

Approval to conduct the experiment for data collection was granted by the Research and Ethics Board at the University of Guelph in April of 2007. 32 participants, ranging in age from 18 to 45 years, took part in the experiment. To eliminate influence of language and cultural factors, it was required that subjects' first language be Canadian English.

After being introduced to the concept of the image retrieval game, each of these subjects described a total of 60 configurations from six different sets. 40 of the configurations were from 2 sets unique to each subject; these 1280 descriptions (32 participants x 40 configurations) provide sufficient information for system training, and allow us to obtain some statistics on spatial relations used and terms used to refer to them. Additionally, each

subject was tasked with describing the first 5 configurations in 4 configuration sets common to all subjects. The 640 descriptions of the common configurations provide information for the study of consistencies and variations among subject's descriptions. These allow us to validate the existence of a prototypical perception, and get an idea of what elements are involved in this prototypical perception.

Extracting Target Information

In total, 1920 spoken descriptions were collected. To capture from these spoken descriptions the information we are interested in, manual processing is required. In order to reliably extract and encode this information for analysis, constraining procedures must exist at this point. The extraction task was modelled based on the following observations, made in a preliminary review of the descriptions.

Not surprisingly, we found the descriptions to be in a variety of forms and grammar structures, and some of the information provided to be irrelevant or unusable. We specify *target information* as information about the position of one object (the argument) relative to another (the referent). The following description is an example of what is considered target information: "The grey object is to the right and below the black object." In this description, the reference object is the black object, and the relative position is described by the terms *right* and *below*, which denote spatial relations. Non-target information includes the following:

- Information about the shapes, sizes and orientations of objects. For example: "There are two objects. They are both star-like shapes. But the grey object is smaller than the black object." Because the goal of the description task is to describe the relative position of the objects, all of the information provided in this description is irrelevant.
- Information about the absolute positions of objects. For example: "The grey image is on the right side, towards the top of the page." The information provided by this description, about the grey object's absolute position, is irrelevant to the task of finding one or more *similar* configurations.
- Information related to or dependent on other configurations. For example: "The gap [between the objects] is much smaller than in the previous image." This information is relevant, but it is not exploitable; within the context of this work, one and only one configuration is considered at a time.
- Information that is confusing or involves the use of abstract concepts. For example: "If a vertical line rejoins at the lower edge point of the dark object, it will pass through the lower center point of the light object." The information provided by this description may be relevant, but it is not usable, because spatial relations are not explicitly involved, nor can they be reliably implied.

Many different terms were used to describe the same relation, i.e., *north* and *higher* both refer to the spatial relation *ABOVE*. We counted more than 50 distinct terms in the initial review of the descriptions, not including grammatical variations (e.g., *intersect, intersecting*), negative expressions (e.g., *not near, no overlap*), or linguistic hedges (e.g., *barely*, *almost*). A preliminary list of the most commonly used terms, denoting 19 different relations was generated. These *prelisted relations*, and their associated terms and categorizations, are provided in Table 1.

Table 1. The Prelisted Relations

DIRECTIONAL RELATIONS			
RIGHT	***LEFT***	***ABOVE***	***BELOW***
right *east*	*left* *west*	*above* *north* *(on)top(of)* *overtop* *upper(up)* *higher*	*below* *south* *bottom* *down* *lower* *underneath* *under*

TOPOLOGICAL RELATIONS						
SEPARATE	**OVERLAP**	**SURROUNDED**	**IN THE MIDDLE**	**TOUCH**	**IN**	**OUT**
separate *apart* *disjoint* *not intersecting* *not overlapping* *not overlaying* *not being covered* *(uncovered)*	*overlap* *overlay* *intersect* *cover* *(on)top(of)* *overtop* *underneath* *bottom*	*surrounded by* *circled by* *enclosed by*	*(in the) middle* *(in the) center* **BETWEEN** *between*	*(barely,almost…)touch* *(barely,almost…)tangent* *(barely,almost…)meet*	*in* *within* *inside* *contained*	*out* *outside*

DISTANCE RELATIONS					
NEAR	**FAR**	**NOT NEAR**	**NOT NEAR AND NOT FAR**	**BESIDE**	**MEASUREMENT**
near *close* *small (tiny…) gap* *small(tiny…) space* *small(tiny…) distance*	*far* *big(wide…) gap* *big(wide…) space* *big(wide…) distance*	*not near* *not close* **NOT FAR** *not far*	*not near and not far* *not close and not far* *median(moderate…) gap* *median(moderate…) space* *median(moderate…) distance*	*beside* *next to* *side by side*	*inch* *cm* *mm* *size of object* *size of space*

We found that in some descriptions, spatial relations are implied. Consider the following description: “The grey object is overlapping the bottom left part of the black object." Although the subject has not said that the grey object is *BELOW* or to the *LEFT* of the black object, one might accurately deduce that this is the case, based on the description of the overlapping regions. Similarly, it was also observed that subjects at times referred to spatial relations between object parts (“The grey object lies to the left of *the upper half of* the black object"), as opposed to considering the objects in general (“The grey object lies to the left of the black object").

Based on these observations, an interface was developed to allow a Research Assistant (RA) to extract target information by answering the following questions while listening to each description:

Q1. Does the description contain any target information?

Q2. Which object is the reference and which one is the argument?

Q3. Does the description involve any of the prelisted spatial relations? What terms are used to describe the relation?

Q4. Is the relation referred to explicitly or implicitly?

Q5. Is the relation between parts of the objects, or the entire objects?

Q6. Does the description involve relations that are not in the provided list? For each non-prelisted relation: Is the relation referred to explicitly or implicitly? Is the relation between parts of the objects, or the entire objects?

To assist the RA in his/her task, the interface provides the prelisted relations and associated terms illustrated in Table 1. However, these lists of relations and terms are by no means exhaustive, and the RA is trained and encouraged to extend them. The interface allows for the audio description to be paused and repeated, to allow the RA to correctly perform his/her task. Viewing the configuration is possible, but is strongly discouraged - the RA is instructed that this option should be accessed only when he/she feels that further clarification on a description is required. This encourages reliable extraction of the descriptions provided, and minimizes potential biases.

Although most descriptions maintain a consistent reference object, there are some in which it is not stated explicitly which object is the reference (e.g., “The black and grey objects are intersecting each other"), or the objects are used alternately as the referent (e.g., “The black object is below the grey object. The grey object is close to the black object"). For consistency, in both of these cases, the RA is instructed to select the *black* object as the referent by default. In the second case, the RA must also enter what he/she deems to be the *semantic inverse* (Freeman, 1975) of any relations in which the grey object is used as the reference object. In the example above, the RA will choose the relations *ABOVE* (the semantic inverse of *BELOW*) and *CLOSE*. Although this inversion step requires some extra effort on the part of the RA, it is required for comparison of descriptions of the same configuration provided by different subjects.

Although the information collected pertaining to Q4, Q5, and Q6 is not involved in the current analysis, we do plan to use it in future research. Also, in the framework of this work, linguistic hedges are ignored, and so is the order in which spatial relations are referred to in the description. For example, no distinction is made between “The grey object is *to the right* of the black object" and “The grey object is *mostly to the right* of the black object". Also, no distinction is made between “The grey object is *to the right and below* the black object" and “The grey object is *below and to the right* of the black object." In an effort to ensure reliability in the results, two RAs were assigned to process each of the 1920 descriptions independently, resulting in two independent data sets.

DATA ANALYSIS

Agreement on spatial information extraction

Each RA found that 1914 of the 1920 spoken descriptions contain target information (Q1). The computed average number of spatial relations provided in each of these descriptions is 3.26 in the first RA’s data set, and 3.2 in that of the second RA. The minimum number of relations provided in a description is 1 in both data sets, and the maximum number of relations is 4 in the first RA’s data set, and 6 in that of the second RA. The most frequently used prelisted relations, along with the two most commonly used terms for each one, are illustrated in Table 2. The relation *RIGHT* was used in 718 descriptions (37.5% of the 1914 descriptions) according to the first RA, and the second RA extracted *RIGHT* from 719 descriptions (37.6% of the 1914). Both RAs found that of the descriptions that included reference to the relation *RIGHT*, the term *right* was used in 97%, and the term *east* was used in 3%. Clearly, from Table 2, the two RAs reach strong agreement that directional relations dominate over topological and distance relations in terms of frequency of use. Many of the discrepancies that exist in this table can be explained by the implicit attribute

of some relations provided in descriptions, i.e., for a given description, one RA may have deduced some relations from the description that the other RA did not.

In 1790 (93.5%) of the 1914 descriptions containing target information, the RAs agreed on which object is the reference object(Q2), and the two objects are seldom used alternately as the referent (2.3%). Since it is difficult to judge whether the RAs agreed on spatial information extraction (Q3 to Q6) when they have failed to agree on the referent, we based further analysis about the agreement between the data sets on 1790=1914-124 descriptions. These 1790 descriptions, along with the 19 listed spatial relations, resulted in a total of 34010 answers to question Q3. The RAs gave congruent positive answers (i.e., the relation was included in the description) in 4949 cases (14.6%), and congruent negative answers (i.e., the relation was not found in the description) in 28075 cases (82.6%); thus agreement was achieved on the answer to question Q3 in 97.2% of the cases.

Table 2. The Frequency of Use of the Prelisted Relations

Relation	RA1 Frequency (%) Term (%)	RA2 Frequency (%) Term (%)
RIGHT	718 (37.5)	719 (37.6)
	right (97) *east* (3)	*right* (97) *east* (3)
LEFT	819 (42.8)	791 (41.3)
	left (97) *west* (3)	*left* (97) *west* (3)
ABOVE	759 (39.7)	766 (40.0)
	above (58) *top* (14)	*above* (55) *top* (15)
BELOW	733 (38.3)	712 (37.2)
	below (55) *lower* (23)	*below* (54) *lower* (23)
SEPARATE	907 (47.4)	726 (37.9)
	separate (43) *space between* (18)	*space between* (31) *separate* (25)
OVERLAP	640 (33.4)	619 (32.3)
	overlap (76) *top* (7)	*overlap* (75) *top* (7)
MEASUREMENT	516 (27.0)	582 (30.4)
	size of object (43) *inch* (35)	*size of object* (44) *inch* (31)

Agreement between descriptions

In measuring the agreement between the relations extracted from each description, we focus our analysis on only the prelisted relations. We also consider knowing whether or not a relation is involved in the description more important than knowing the specific attributes (the term used, whether the relation was referred to explicitly or implicitly, and whether the relation was between whole objects or object parts) of that relation. For each congruent description, we say that two RAs reach agreement on spatial relations only if the set of spatial relations extracted by the first RA, S1, and that by the second RA, S2, satisfy the following condition: $S1 \subseteq S2$ or $S2 \subseteq S1$ ($S1 \neq \emptyset$ and $S2 \neq \emptyset$ because we are considering congruent descriptions). In total, 1617 (90.3%) of the 1790 congruent descriptions are considered to have such agreement, and therefore are regarded to be *reliable*. For each reliable description, we can then merge the information extracted from both RAs, and let $S = S1 \cap S2$ be the merged set of spatial relations associated with the description. The actual term used for each relation is discarded, and if the relation is implicitly referred to according to S1 or S2, then it is considered implicitly referred to in S; this is because

explicit is the default, and the RA must intentionally input that the relation is implicit, so we consider that it is likely not done in error. For the same reason, if the relation is between object parts according to S1 or S2, it is considered between object parts in S.

Inter-subject variations

Inter-subject variations reflect how different subjects describe configurations, and are measured based on the descriptions of the 20 configurations that are common to all subjects, illustrated in Figure 3. Although all 32 participants provided descriptions for these common configurations, some are not considered to be reliable. The results presented here are therefore based on the remaining 563 reliable descriptions. The number of reliable descriptions for each of the 20 configurations varies from 23 to 30, with the average number 28.

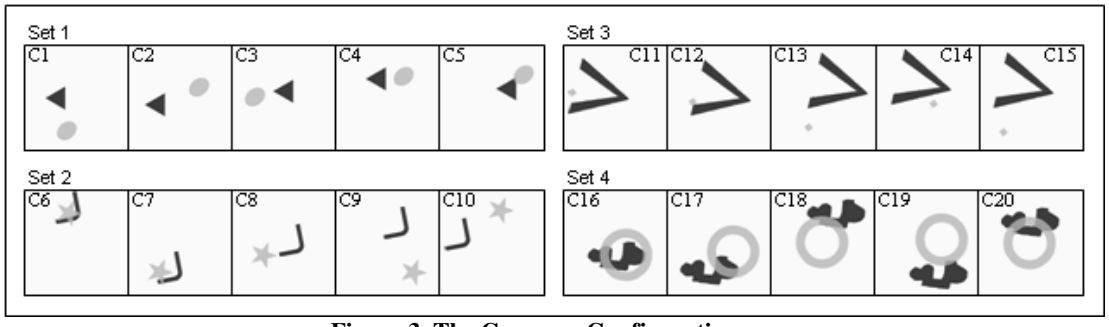

Figure 3. The Common Configurations

We first analyzed the data looking for tendencies in the subjects' use of different types of spatial relations, e.g., directional, topological and distance relations. The number and category of the specific relations provided in each of the 563 merged strings were counted. From the number of *N* reliable descriptions for a given configuration, we count the number of times spatial relations in a given category are used. We then normalize these values to percentages to determine the frequency of use. A high percentage (close to 100%) means that nearly every subject included this type of relation in his/her description of the configuration. A low percentage (close to 0%) denotes that very few of the subjects involved this type of relation. In both of these cases, the subjects have similar tendencies. A median percentage (close to 50%), however, indicates the subjects have very different tendencies, since about half of them involve this type of relation and about half of them do not.

We found that subjects' tendencies to involve different types of spatial relations vary from configuration to configuration and from set to set. For example, subjects' tendencies to involve directional relations, while they are similarly strong for configurations in Sets 1 and 3 (the average frequencies for directional relations are 99% and 97% respectively) are weak for C6 and C16 (only 49% and 39% of the descriptions for these configurations involved directional relations). Not surprisingly, for these two particular configurations, 93% and 100% of the descriptions involved topological relations. The subjects seldom involve distance relations when describing configurations in Set 4, and furthermore, none of the subjects mention distance relations when describing C16. It seems that subjects have similar tendencies to involve directional relations throughout describing different sets of configurations, however, they have very different tendencies to involve topological and distance relations when describing, especially, the first three sets of configurations. One

possible reason is that the first three sets are simpler than the fourth, causing some subjects to only use one or two types of spatial relations when describing them.

To expand further on these observations, the consistency between subjects in their use of specific relations, and categories of relations, is computed for each of the 20 configurations. To do so, we consider two subjects, A and B, who have both provided a reliable description for a given configuration, where A's description provides *P* relations, and B's description provides *Q* relations. The consistency between the relations provided by A and B is defined as c(A,B) = *R/min(P,Q)*, where *R* is the number of relations common to both descriptions. The value of c(A,B) reaches 1 when the set of relations used by one subject entirely includes that used by the other. The consistency in the use of a given category of relations is computed in a similar fashion: c_{TYPE} (A,B) = R_{TYPE} / min(P_{TYPE},Q_{TYPE}). In cases where one subject does not provide any relations of a certain type, min(P_{TYPE},Q_{TYPE}) = 0, and c_{TYPE}(A,B) is set to 1, because one subject's failure to include a certain type of relation that was provided by the other does not constitute disagreement. The consistency between subjects in describing the same configuration is measured as the average of the consistencies between all possible pairs of subjects. For instance, the consistency between three subjects A, B and C in their use of topological relations is [c_{TOP} (*A;B*) + c_{TOP} (*A;C*) + c_{TOP} (*B;C*)] / 3. For *n* subjects, the consistency is the average of *n*(*n*-1)/2 values. According to this formula, we calculate the consistency between all of the subjects in their use of directional relations (c_{DIR}), topological relations (c_{TOP}), distance relations (c_{DIS}), and all relations together (*c*).

Overall, the subjects involve directional relations consistently, i.e., consistency >90%, in merged descriptions of configurations C1 through C5, C8 through C10, and C13 through 15; for 9 of these 11 configurations, consistency of directional relations is 100%. Topological relations were consistently involved only in the merged descriptions of configurations in Set 4, and distance relations were consistently involved only in descriptions of C6. In cases where only a small number of subjects provide inconsistent relations of a given type, the type consistency will remain quite high, and a measure of the consistency between only the informative descriptions (descriptions actually involving the type of relation) is more appropriate. When only the informative descriptions were considered, the average consistencies in the use of directional and topological relations did not change much, but the average consistencies in the use of distance relations decreased significantly. Table 3 illustrates the average consistencies between the subjects in their use of different types of relations in descriptions of the common configuration sets, and how these consistencies vary when only informative descriptions are considered.

Table 3. Consistency of Use of Types of Relations

	DIRECTIONAL All (Informative)	**TOPOLOGICAL** All (Informative)	**DISTANCE** All (Informative)	**ALL**
Set 1	0.97 (0.97)	0.99 (0.99)	0.89 (0.71)	0.87
Set 2	0.95 (0.88)	~1.0 (~1.0)	0.88 (0.66)	0.85
Set 3	0.90 (0.89)	0.92 (0.81)	0.90 (0.46)	0.70
Set 4	0.99 (0.95)	0.93 (0.93)	1.0 (0.92)	0.91
All Sets	0.95 (0.92)	0.96 (0.93)	0.92 (0.70)	0.83

Next, we investigate if the subjects involve the same spatial relations when describing the configurations in the common set of 20. Note that there are 19 prelisted relations, but because none of descriptions of the common configurations make use of the relations *BESIDE* and *NOT FAR*, we only count occurrences of 17 relations. As can been seen from Table 4, which illustrates only the frequencies and consistencies of the most commonly

used relations, the subjects have similar tendencies to involve most spatial relations, except for the relations *SEPARATE* and *MEASUREMENT*. For example, 59% of descriptions of C1 make use of the relation *SEPARATE*, and the consistency between subjects for this relation in describing C1 is 0.17.

Table 4. Frequency and Consistency of Relations in Descriptions of Common Configurations

	RIGHT		LEFT		ABOVE		BELOW		SEPARATE		OVERLAP		MEASUREMENT	
	%	C	%	C	%	C	%	C	%	C	%	C	%	C
C1	62	*0.24*	3	*0.93*	0	*1*	100	*1*	59	*0.17*	0	*1*	48	*0.03*
C2	97	*0.93*	3	*0.93*	90	*0.8*	0	*1*	57	*0.13*	0	*1*	47	*0.07*
C3	4	*0.92*	92	*0.85*	0	*1*	46	*0.08*	54	*0.08*	0	*1*	42	*0.15*
C4	100	*1*	0	*1*	7	*0.86*	0	*1*	46	*0.07*	0	*1*	39	*0.21*
C5	82	*0.64*	0	*1*	96	*0.93*	0	*1*	0	*1*	96	*0.93*	14	*0.71*
Set 1		***0.75***		***0.94***		***0.92***		***0.82***		***0.29***		***0.99***		***0.23***
C6	7	*0.86*	34	*0.31*	14	*0.72*	14	*0.72*	0	*1*	93	*0.86*	0	*1*
C7	0	*1*	90	*0.79*	10	*0.79*	14	*0.72*	0	*1*	93	*0.86*	7	*0.86*
C8	0	*1*	93	*0.87*	0	*1*	43	*0.13*	70	*0.4*	0	*1*	33	*0.33*
C9	80	*0.6*	3	*0.93*	0	*1*	100	*1*	60	*0.2*	0	*1*	43	*0.13*
C10	97	*0.93*	3	*0.93*	93	*0.87*	0	*1*	57	*0.13*	0	*1*	43	*0.13*
Set 2		***0.88***		***0.77***		***0.88***		***0.71***		***0.55***		***0.94***		***0.49***
C11	0	*1*	76	*0.52*	24	*0.52*	31	*0.38*	24	*0.52*	0	*1*	21	*0.59*
C12	0	*1*	83	*0.67*	10	*0.8*	47	*0.07*	3	*0.93*	17	*0.67*	3	*0.93*
C13	4	*0.93*	63	*0.26*	0	*1*	100	*1*	52	*0.04*	0	*1*	41	*0.19*
C14	57	*0.13*	10	*0.8*	0	*1*	100	*1*	43	*0.13*	0	*1*	43	*0.13*
C15	0	*1*	77	*0.54*	4	*0.92*	92	*0.85*	35	*0.31*	0	*1*	46	*0.08*
Set 3		***0.81***		***0.56***		***0.85***		***0.66***		***0.39***		***0.93***		***0.38***
C16	35	*0.3*	0	*1*	4	*0.91*	4	*0.91*	0	*1*	96	*0.91*	0	*1*
C17	85	*0.7*	0	*1*	81	*0.63*	0	*1*	0	*1*	100	*1*	15	*0.7*
C18	0	*1*	56	*0.11*	0	*1*	93	*0.85*	0	*1*	96	*0.93*	11	*0.78*
C19	3	*0.93*	7	*0.87*	93	*0.87*	0	*1*	0	*1*	93	*0.87*	17	*0.67*
C20	0	*1*	4	*0.92*	0	*1*	84	*0.68*	0	*1*	96	*0.92*	16	*0.68*
Set 4		***0.79***		***0.78***		***0.88***		***0.89***		***1***		***0.93***		***0.77***
All		***0.81***		***0.76***		***0.88***		***0.77***		***0.56***		***0.95***		***0.47***

Overall, we can conclude that the subjects are quite consistent with each other in their use of spatial relations, especially when the fact that some people may use more relations than others is taken into account.

CONCLUSIONS

In this work we have presented a method of capturing natural language descriptions of relative positions of image objects that reduces bias in the descriptions. The fruitful results of spatial information extraction and analysis provide a general idea of the most common terms and relations used in natural language descriptions of spatial relations between objects in images. In addition to determining these elements of a 'prototypical' perception, we feel that the results provide a solid foundation for further study of inter-subject variations. Although the configuration sets assigned to the subjects provided a wide variety of scenarios for the study of common elements of perception, having each subject describe the same 60 configurations (3 sets of 20) would provide further information for the study of inter-subject variations. The design of the experiment described in this paper is not flawless and some of the choices we made may be considered questionable. We present the following limitations and considerations:

The use of a single reference object for all relations provided in a description is a clear limitation. In descriptions that provide relations with mixed reference objects, where the RA was instructed to select the black object as the reference object by default, and enter the semantic inverse of relations in which the grey object was used by the subject as the reference, the relations and terms entered by the RAs may not reflect the description exactly. This instruction may have also resulted in disagreement between the two data sets as to which object was the reference; cases where the RAs do not agree on which object is the reference are disregarded in further analysis.

Because the instances of spatial relations are not uniformly distributed, the numbers of positive examples for different spatial relations are disproportionate. Additionally, because non-prelisted relations were omitted from the measurement and merging procedures, information about occurrences of rarely used relations is limited. Furthermore, we did not use all of the information collected in the experiment, and we feel the additional consideration of each new piece of information could assist in achieving a closer approximation of how humans describe images.

ACKNOWLEDGMENTS

The authors wish to express their gratitude for support from the Natural Science and Engineering Research Council of Canada (NSERC), grant 262117.

REFERENCES

Freeman, J. (1975) “The Modeling of Spatial Relations”, *Computer Graphics and Image Processing*, (4),156-171.

Gapp, K. P. (1995) “Angle, Distance, Shape, and their Relationship to Projective Relations", *Proceedings of the17th Annual Conference of the Cognitive Science Society*, San Diego, CA, 112-117.

Hall, E. T. (1966) *The Hidden Dimension*, New York: Doubleday.

Halpern, D. F. (1986) *Sex Differences in Cognitive Abilities*, Hillsdale, NJ: Lawrence Erlbaum Associates Press.

Herskovits, A. (1986) *Language and Spatial Cognition: A Interdisciplinary Study of the Prepositions in English*. Cambridge, England: Cambridge University Press.

Landau, B. (1996) “Multiple Geometric Representations of Objects in Language and Language Learners", in P. Bloom, M. Peterson, L. Nadel, and M. Garrett (eds), *Language and Space*, Cambridge: MIT Press, 317-363.

Linn, M. C. and Petersen, A. C (1985) “Emergence and Characterization of Gender Differences in Spatial Ability: A Meta-Analysis", *Child Development*, 56(6), 1479-1498.

Mark, D. M., Comas, D., Egenhofer, M. J., Freundschuh, S. M., Gould, M. D. and Nunes, J (1995) “Evaluating and Refining Computational Models of Spatial Relations Through Cross-Linguistic Human-Subjects Testing", in Frank, A. U. and Kuhn, W. (eds), *Spatial Information Theory: A Theoretical Basis for GIS*, number 998, Springer-Verlag, Lecture Notes in Computer Sciences: Berlin, 553-568.

Mark, D. M. and Egenhofer, M. J. (1994) “Modeling Spatial Relations Between Lines and Regions: Combining Formal Mathematical Models and Human Subjects Testing", *Cartography and Geographic Information Systems*, 21(4),195-212.

Montello, D. R. (1995) “How Significant Are Cultural Differences in Spatial Cognition?", in Frank, A. U. and Kuhn, W. (eds), *Spatial Information Theory: A Theoretical Basis for GIS*, Springer-Verlag, Lecture Notes in Computer Sciences: Berlin, 485-500.

Regier, T. P. (1992) *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization"*, PhD thesis, University of California at Berkeley, USA.

Robinson, V. B. (1990) “Interactive Machine Acquisition of a Fuzzy Spatial Relation", *Computers and Geosciences*, 16(6), 857-872.

Talmy, L. (1983) “How Language Structures Space", in Pick, H. and Acredolo, L. (eds), *Spatial Orientation: Theory, Research, and Application,* New York: Plenum Press, 225-282.

Wang, X. and Keller, J. M. (1999) “Fuzzy Surroundedness", *Fuzzy Sets and Systems*, 101(1), 5-20.

Wang, X. and Keller, J. M. (1997) “Human-based Spatial Relationship Generalization through Neural Fuzzy Approaches", *Proceedings of the Sixth IEEE International Congress on Fuzzy Systems*, Barcelona, Spain, 1173-1178.

Whorf, B. L. (1940) “Science and Linguistics", *Technological Review*, 42(6), 229-231, 247-248.

Worboys, M. F. (2001) “Nearness Relations in Environmental Space", *International Journal of Geographical Information Science*, 15(7), 633-651.

Zhan, F. B (2002) “A Fuzzy Set Model of Approximate Linguistic Terms in Descriptions of Binary Topological Relations Between Simple Regions", in Matsakis, P. and Sztandera, L.M. (eds), *Applying soft computing in defining spatial relations*, Physica-Verlag, Heidelberg, Germany, 179-202.