Spatial relationships and fuzzy methods: Experimentation and modeling

James Lee Ward (robotcycle@gmail.com)
Army Research Office, Research Triangle Park, NC

Robert St. Amant (stamant@ncsu.edu)
Computer Science Department, North Carolina State University, Raleigh, NC

MaryAnne Fields (mary.a.fields22.civ@mail.mil)
U.S. Army Research Laboratory, 28000 Powder Mill Rd. Adelphi, MD

Abstract

This paper describes an experiment and fuzzy set models in the domain of linguistic labels for simple spatial relationships: for example, that one object is "in front of" or "to the right of" another. Input to the models was generated by robot sensors (camera and distance sensors), from a viewer perspective on different configurations of two objects. Performance of the models is is qualitatively similar to human judgments; performance is also quantitatively similar to that of a model working from an environmental bird's-eye view. Such models are part of a robot's interpretation of the context of human activity.

Keywords: spatial relationships; fuzzy sets; cognitive robotics

Introduction

As we attempt to create a new generation of automated helpers to solve problems in the military, elder assistance, transportation, and other areas, we increasingly find that we need robots that can interact naturally with humans and that can move through environments designed for humans. A critical challenge for such robots is the use of context.

Context can help a robot to impose structure on information available to it, in a top-down manner. Some kind of contextual information can be provided by background knowledge or experience of common human activities. For example, "writing a paper" may be associated with scenes such as a "computer lab" or a library, or with object clusters such as a table and chair (Fields, Lennon, Martin, & Lebiere, 2017).

Context is also provided by information about human behavior and performance, which can be exploited in research that integrates cognitive modeling and robotics. Our lab has begun to explore the combination of language comprehension models and information-based search; some of our current research deals with spatial relationships.

Consider the diagram of two objects in Figure 1, one labeled L (for "landmark") and the other T (for "trajector"). The landmark sets the context for the relationship, while the trajector occupies a position—a place in the relationship—with respect to the landmark. If this diagram were a bird's-eye view of a room, a person in the position of observer O_1 would probably say that "T is to the right of L." Would the person also say that "L is in front of T" or that "T is in back of [or behind] L"? Spatial relationships that can be easily diagrammed may be ambiguous when described in language.

The ways that people conceptualize space (and action) have long been a subject of study in psychology. Applying research findings to robot behavior is a more recent development. There are clear advantages in human-robot interaction for a robot that incorporates the ability to take as input, generate as output, or reason about expressions of spatial relationships (Trafton & Harrison, 2011; Guadarrama et al., 2013; Tellex et al., 2011). The work of Regier and Carlson (2001) and others hints at another possibility: a model of human interpretations of spatial relationships may provide information to a robot about what is of interest to individuals or to people in general. For a simple example, people typically attend to what is in front of them; in a classroom full of desks and chairs, it is straightforward to infer the general area a teacher will occupy. We even find spatial directions used in metaphorical language concerning attention: "it's right in front of you" indicates that you should notice whatever it is.

To explore such issues, it will be useful to have a reliable way for a robot to associate spatial relationships with labels such as *left*, *right*, *in front of*, and *in back of*, in the same way that humans do. While there are obvious, canonical examples of such relationships, not all fall crisply into one category or another. Further, robots must deal with noisy sensors and motor movements, which might plausibly interfere with their categorizations of objects in the environment.

In the remainder of this paper we give a brief overview of work on linguistic labels for spatial relationships. We describe an experiment in which participants made judgments about spatial relationships between two objects. We then describe three *a priori* models, from the fuzzy systems literature (Keller & Wang, 1995), that allow a robot to make the same viewer-perspective judgments about the spatial relationships. We compare their performance to the human data and find qualitatively similar model predictions.

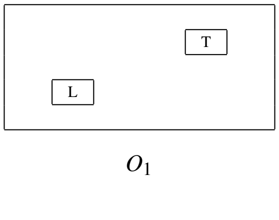

Figure 1: Object/observer relationships

Related work

The literature related to representation, reasoning, and communication concerning spatial relationships is enormous. Mechanisms underlying spatial representation and reasoning have been explored in some depth in the cognitive modeling literature (Harrison & Schunn, 2003; Gunzelmann & Lyon, 2006; Trafton & Harrison, 2011); cognitive models also encompass the language of spatial relationships (Ball, 2015). The research described in this paper is much narrower, however, focusing on spatial relationships that can be expressed in linguistic terms (e.g., *left*, *right*, *above*, *below*, *in front of*, and *behind* or *in back of*) and the extent to which they can be grounded in the perception (of a human or a robot agent).

Regier and Carlson (2001)'s Attentional Vector-Sum (AVS) model is widely accepted as the best model of how a set of spatial expressions can be grounded in perception. Informally, for the *above* relationship, an attentional beam is conceptualized as extending from a trajector object to a landmark object. Attention is strongest at the point on the landmark directly below the trajector, and weaker at other points on the landmark as distance from this point increases; the drop-off is a free parameter in the model. A distribution of vectors is identified, originating at different points on the landmark and directed toward the trajector, the magnitude of each determined by attention. The sum of these vectors is compared with a vertical line, and the deviation determines the extent to which the trajector is above the landmark.

Regier and Carlson (2001) compare the AVS model with others, including a Bounding Box (BB) model and a Proximal and Center-of-Mass (PC) model, both of which it outperforms. For the BB model and the *above* relationship, "a trajector object is above a landmark object if it is higher than the highest point of the landmark and between its rightmost and leftmost points" (Regier & Carlson, 2001). For the PC model, consider a vector from the center of a landmark to the center of a trajector. As this vector deviates from the vertical (roughly, 68° to 72°) ratings of the *above* relationship decrease linearly; further increases cause a much faster drop off, to zero at 90° or greater. Proximity comes into play with a line segment connecting the landmark and trajector at their minimum distance; to the extent that this segment is aligned with the center-of-mass vector, the *above* relationship holds.

Judgments about *above* generalize to comparable relationships, including *left*, *right*, *in front of*, and so forth (Regier & Carlson, 2001). For example, if we interpret the bottom of the box in Figure 1 as being a horizontal surface, then we could ask, "Is *T* above *L*?" and use the same PC model, reinterpreted, to answer the question.

The AVS model and others have been used in computer vision and robotics research, though they typically require some adaptation. In most experiments on labeling spatial relationships, a scene is presented in which the relationship of interest is visible in a plane normal to the participant's line of sight. For example, consider rating the relationships *left*, *right*, *in front of*, and *in back of* for two objects on the floor in a room. A bird's-eye view, along a normal to the plane of the floor, would be a typical presentation, as in Tellex et al. (2011, Figure 4), which we will call an *orthogonal* view. An observer—such as a robot—inside the room with the objects, however, would face a related but slightly different problem. This viewer perspective, in which judgments are required for relationships that are aligned with viewer's line of sight, are part of the experiment described in the next section.

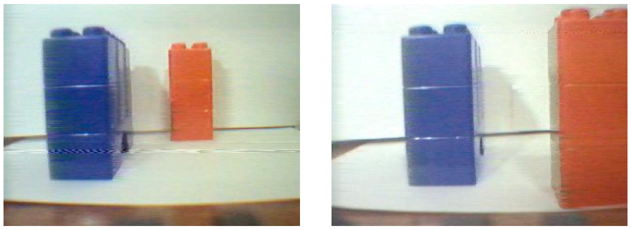

Figure 2: Object configurations (red on the right)

Experiment

This experiment was intended to benchmark performance in answering questions about spatial relationships. Because our eventual goal is a robot that can reproduce human judgment of specific spatial relationships, by reference to a model, the raw data was provided by camera images from a robot: a LEGO Mindstorms NXT robot with customized sensors, including a still camera.

Two stacks of blocks were used for the experiment, as shown in Figure 2. The stack of square red blocks was 10 cm on each side. The oblong stack of blue blocks was 10 cm wide and 30 cm long. Both stacks were about 21 cm tall.

These two stacks were placed in different configurations as follows. The blue stack was initially placed with its narrow side facing the robot, and the red stack was placed to the right and a few cm behind the blue stack, comparable to a configuration in Gapp (1995), in Figure 1b. The red stack was then advanced in increments of 10 cm in a straight line towards the robot. The advances were performed six times until the red stack was about 10 cm in front of the blue stack. After the six advances the red stack was returned to its starting position and the blue stack was rotated by a one-eighth turn. The process was repeated. This continued until the blue stack was at a right angle from its original position.

The robot, approximately 60 cm distant from the centroid of the two stacks in the starting configuration, followed the procedure in Figure 3 after each change in the configuration. Thirty images were collected in total, at six different locations of the red stack and five different rotations of the blue stack. Figure 2 shows two images, the starting configuration on the left and after five steps into the procedure on the right.

The images[1] recorded during this procedure formed the basis of a survey. Twelve participants completed the survey,

[1] Images were used instead of a real environment for consistency across experiment participants.

```
Record compass reading
Record camera image
For target color in {RED, BLUE}
    Identify object of target color
    For target in {LEFT, CENTER, RIGHT}
        CALC: Calculate rotation to center on target
            Rotate
            If within threshold
                Record compass reading
                Record distance reading
            else go to step CALC
```

Figure 3: Measurement procedure

eight men and four women, ranging in age from 28 to 71. The participants received no compensation for participation and were not observed during the task. The sequence of images was randomized; all participants saw the same ordering. For each image, four statements were evaluated by participants, on a scale of 0 to 10; for analysis, all values were linearly transformed to a unit scale.

1. The red blocks are to the right of the blue blocks.
2. The red blocks are in front of the blue blocks.
3. The red blocks are in back of the blue blocks.
4. The blue blocks are to the left of the red blocks.

In other words, we have two independent variables in this experiment. The variable Distance of the red stack to the robot provides for different participant ratings concerning whether the red stack is in front of, in back of, and even to the right or left of the blue stack, in each location. The variable Angle, for the rotation of the blue stack, provides a different cross-section to the viewer as well as a different angle with respect to the red stack.

Note that the experiment excludes the most "obvious" configurations for Front and Back ratings—for example, with one block directly in front of the other, from the position of the camera. There is a sense in which the experiment tests "edge cases" for spatial relationship judgments.

The three plots in Figure 4 show survey ratings for the *right*, *front*, and *back* questions. Values for *left* are not shown, being almost identical to *right*. Each group of six connected dots shows the mean values, scaled from 0.0 to 1.0, the six locations of the red stack, at decreasing Distance from the robot's camera. Five groups are shown, with a graphical icon for each Angle value of the blue stack. Within each group, the sequence shows the red stack moving foward in steps.

No significant effect on Right ratings or on Left ratings was found. The mean values of Left and Right were above 0.9, over all trials, the median equal to 1.0. While a slight inverted U-shaped pattern is visible, we did not analyze the data further.

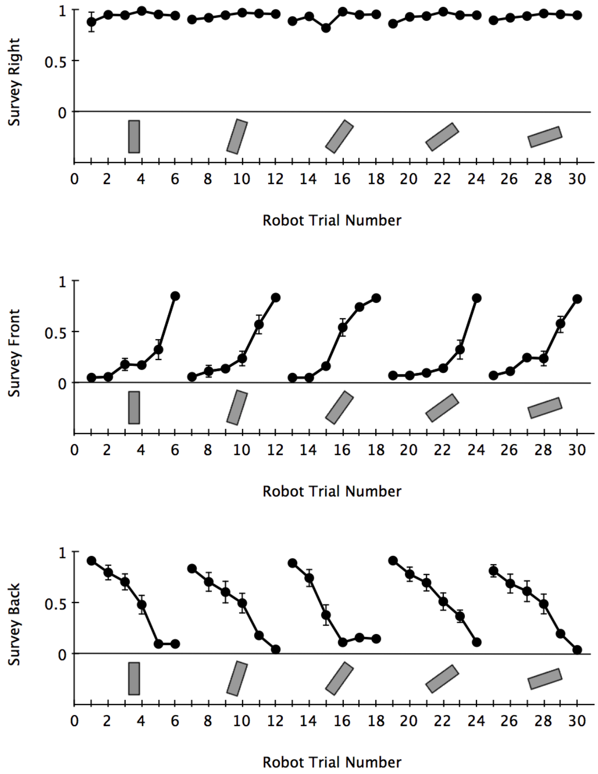

Figure 4: Mean ratings of *right*, *front*, and *back*, with standard error bars; gray blocks show rotation angle of blue stack

An analysis of variance showed a significant overall effect of Distance and Angle on Front ratings, as expected ($F(5,4) = 25.790, p < 0.01$). Distance alone had a significant effect on Front ($F(5,4) = 28.036, p < 0.01$), but Angle did not ($F(5,4) = 0.034$, n.s.); the interaction between Distance and Angle was significant ($F(5,4) = 2.677, p < 0.01$).

Similarly, ANOVA showed a significant overall effect of Distance and Angle on Back ratings ($F(5,4) = 18.973, p < 0.01$). Distance alone had a significant effect on Back ($F(5,4) = 27.505, p < 0.01$), but Angle did not ($F(5,4) = 0.433$, n.s.); the interaction between Distance and Angle was significant ($F(5,4) = 1.812, p < 0.05$).

The general patterns are as expected: experiment participants were able to make plausible judgments about the spatial relationships between the two stacks in different configurations, basing their judgments on the information provided by the robot's camera. There was no influence of the angle of the blue stack, acting as a landmark in the experiment, though the rotations acted to change the "overlap" with the trajector; the lack of an effect is possibly due to the front view and the limited depth information available in the images. Different shapes are noticeable in the Front and Back ratings, with the Front ratings showing a slightly more pronounced curvature with change in Distance.

Modeling

Models such as AVS, PC, and BB have been adapted for use in computer vision and robotics (Guadarrama et al., 2013; Tellex et al., 2011), and new models have been developed Matsakis and Wendling (1999). In such work, however, the models are generally not evaluated directly or compared with human performance, and the models may not take a parameterized form with explicitly identified features, with Gapp (1995) being an exception.[2]

Our work adopts vector-based methods to model the ratings described in the previous section. Regier and Carlson's PC model is a possible candidate, but its four free parameters make it difficult to adapt—specifically, changing to a viewer perspective to evaluate relationships parallel to the viewer's line of sight. Instead, we evaluate three simpler models from the computer vision literature due to Keller and Wang (1995).

These are fuzzy set models, which can deal with membership grades in categorization. In a standard "crisp set" formulation, a predicate is true or false; with a fuzzy set, a membership grade can be a value from 0 (not a member of the set) to 1 (a member of the set). Thus, for example, in Figure 1, T might have a membership grade of 0.8 for the categorization "to the right of L;" it would be greater if it were closer to a horizontal line extending through L. Fuzzy methods can do more than assign a grade for a given label; with several labels, they can be used for categorization, even in cases where a given configuration fall into more than one category.

Keller and Wang's *Centroid* method uses Equation 1 as the membership function for the *right* function; analogous functions are defined for *left*, *front*, and *back*. Let the centroid of L be the origin in a Cartesian coordinate system; let θ be the angle of a vector $\overrightarrow{LT}$ through the centroid of T. The function μ_{right} maps θ to a value between 0 and 1, representing the degree to which T is to the right of L:

$$\mu_{right}(\theta) = \begin{cases} 1 & |\theta| < a\frac{\pi}{2} \\ 0 & |\theta| > \frac{\pi}{2} \\ \dfrac{\pi/2 - |\theta|}{\pi/2(1-a)} & o.w. \end{cases} \tag{1}$$

In words, T is maximally to the right of L when θ is within a small range above or below 0 radians ($a\frac{\pi}{2}$, where a is a free parameter, which we set to 0.05 as a default), along the implicit x-axis in the landmark-based coordinate system. μ_{right} decreases linearly as θ increases or decreases, reaching zero when the center of L falls on or below zero on the x-axis.

Computing the centroids of objects is straightforward with an orthogonal view, but this is more difficult for some relationships from a viewer perspective. As described in the measurement procedure in Figure 3, the robot identified three points on each object in the scene, its left edge, center, and right edge; distances were measured to these points and a centroid computed from the values. This gives the centroid estimation for both objects a strong forward bias, though this is partly alleviated for higher Angle values in cases where the blue stack has been rotated.

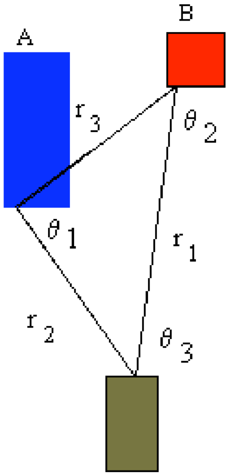

Figure 5: Distance and compass readings, bird's-eye view

The angle θ between the two stacks must be provided as input to μ_{right} but cannot be measured directly; θ is computed, based on the triangle in Figure 5. θ_3, r_1, and r_2 are measured directly by the robot, while θ_1, θ_2, and r_3 are computed:

$$\theta_1 = \cos^{-1}\left(\frac{r_2{}^2 + r_3{}^2 - r_1{}^2}{2r_2r_3}\right); \theta_2 = \pi - \theta_3 - \theta_1, \tag{2}$$

$$r_3 = \sqrt{r_1{}^2 + r_2{}^2 - 2r_1r_2\cos\theta_3}. \tag{3}$$

With measured or computed values for the triangle's sides and angles, plus the assumption that the robot's camera is midway between the two stacks, computing θ for different directions is straightforward trigonometry.

Keller and Wang's second method is *Angle Aggregation*, which samples points from the landmark and trajector objects, computes angles for each pair of points, and aggregates the angles (by a generalized mean operator) into a single value for θ. With three sampled points on each object in the experiment, a total of nine pairs of angles are considered.

The third method is the compatibility method, which we implemented in variant form as a *Histogram/Composition* method. Angles between the two objects are computed as with Angle Aggregation, but instead of computing the mean of the samples, the samples are binned into a histogram, with normalized frequencies for the bins being treated as fuzzy set. This set is composed with the fuzzy sets for *right*, *left*, *front*, and *back*, to compute the membership grade in each category.

[2] Indirect evaluation is carried out, however. Guadarrama et al. (2013) evaluate overall measures of success for experiments with a robot that incorporates a combined PC and BB model to interact with configurations of multiple objects; Tellex et al. (2011) similarly with an AVS model.

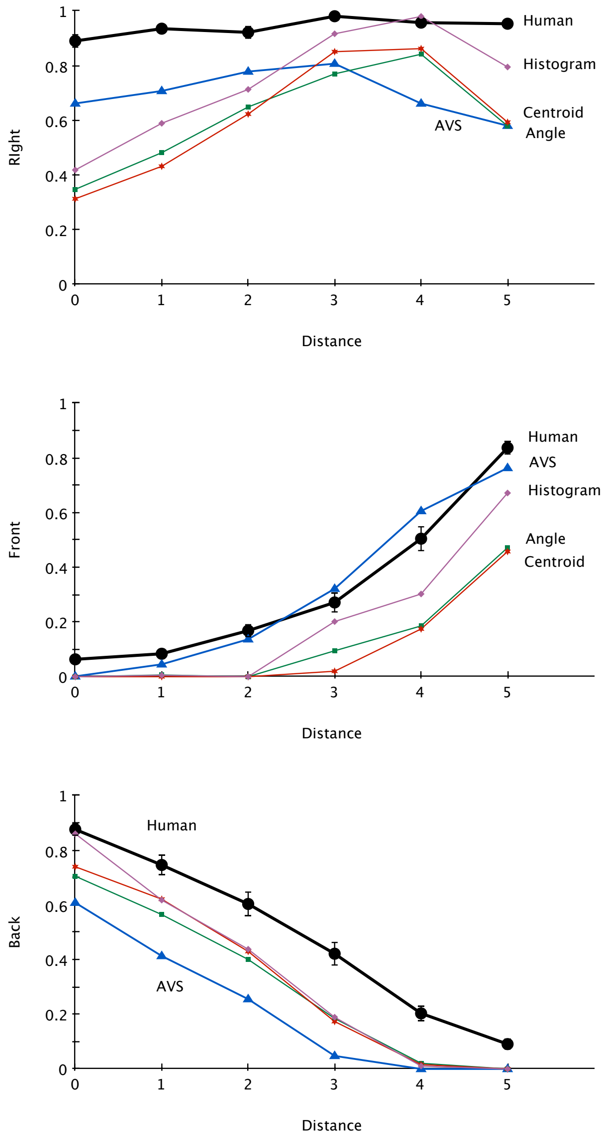

Figure 6: Model predictions for *right*, *front*, and *back*

	Right	Front	Back
AVS	0.891, 0.068 $R^2 = 0.056$	0.046, 0.880 $R^2 = 0.813$	0.285, 0.925 $R^2 = 0.629$
Angle	0.888, 0.083 $R^2 = 0.202$	0.139, 1.449 $R^2 = 0.812$	0.194, 0.946 $R^2 = 0.835$
Centroid	0.887, 0.085 $R^2 = 0.282$	0.164, 1.453 $R^2 = 0.792$	0.193, 0.900 $R^2 = 0.850$
Histogram	0.902, 0.051 $R^2 = 0.138$	0.151, 0.867 $R^2 = 0.698$	0.208, 0.797 $R^2 = 0.817$

Table 1: Model fit statistics

Model predictions are shown in Figure 6 for the survey data for *right*, *front*, and *back*; as before, *left* is elided, being almost identical to *right*. The parameters of the models were not tuned to fit the data.[3] In these plots, Distance values reflect the ordering in which they were considered by the robot, an inversion of the actual distance from the camera: 0 corresponds to the red stack being farthest away from the robot's camera, i.e., behind the blue stack, with a Distance value 1 being closest to the camera and in front of the blue stack. Predictions are aggregated by Distance value, in that Angle had no effect on the spatial judgments.

The black lines show mean ratings from the survey, with error bars showing the standard error with respect to all participant ratings per Distance value. The other colored lines represent predictions of the Centroid, Angle Aggregation, and Histogram/Composition fuzzy models. The fuzzy models are not separately labeled in the Back plot; their values are almost indistinguishable. For reference, predictions of the AVS model are shown as well. The AVS predictions were based on a virtually constructed orthogonal diagram of each configuration and are given for comparison to a method with access to a perspective not available to the other models.

Qualitatively, the fuzzy models are consistent with human ratings, though they are more "conservative" in the sense of assigning lower membership grades than the human participants. This is in part due to the forward bias in the *front* predictions—recall that reference points were identified with the robot's distance sensors, which detect the distance to the front of each object.

The under-predictions of the AVS model for the mean Back rating surprised us; we initially suspected design or implementation errors in the modeling code. Exploration of the participants' data led us to a different conclusion, however.

Leaving out Distance values of 5, the remaining conditions are approximately symmetrical, front-to-back, in geometrical terms. For example, at Distance 0, the front edge of the red stack is aligned with the back edge of the blue stack, while at Distance 4, the back of the red is aligned with the front of the blue, as if reflected in a mirror (though rotation introduces minor asymmetry). The AVS model produces consistent predictions for such symmetrical configurations.

Participant ratings did not show the same consistency between Front and Back ratings. For the symmetrical configurations, Back ratings were higher than Front ratings by about 0.2 on the unit scale. Further experimentation would be needed to verify this bias. It is generally held that models of spatial relationships such as these generalize across orientations; our results suggest that the viewer's perspective may be a factor in the magnitude of ratings.

Table 1 shows how well the models fit the survey ratings, following the approach of Regier and Carlson. Each entry is the y-intercept and slope of a regression that uses the model's output to predict ratings; an R^2 value is below each such entry.

[3] For Keller and Wang's models, $a = 0.05$. The AVS model used parameters given by Regier and Carlson (2001, Table 1): $\lambda = 1.0$ (attentional field width); y-intercept $= 1.007$ and slope $= -0.006$ (alignment function); gain $= 0.131$ (top sigmoid).

For example, we see that AVS gives the best fit for Front: a regression line with a y-intercept of 0.046 and a slope of 0.880 gives the best prediction of survey ratings, with $R^2 = 0.813$. The better the model, the closer the intercept is to zero and the closer the slope is to 1.

Table 1 shows that of the fuzzy models, the Histogram model has the best balance of intercept and slope for the *front* relationship, though a lower R^2. Angle Aggregation is the best for the *back* relationship, though all the fuzzy models are similar. All models perform poorly for *right* and *left* relationships, which is inevitable—Distance and Angle have no predictive value on the participants' ratings.

We tested the sensitivity of the modeling predictions by virtually reconstructing each configuration and running the models on the "theoretical" values for $\theta_1 \ldots \theta_3$ and $r_1 \ldots r_3$. We found no marked difference between the predictions based on the sensor data and the theoretical data; data is not presented for reasons of space.

To summarize, the fuzzy models' predictions tend to underestimate ratings in all categories of spatial relationships that we tested. The predictions for *front* and *back* are of the same shape as the human ratings with respect to relative distance from the viewer, however, which suggests that an additional constant or linear factor could improve their performance. Poor performance on *right* and *left* is partly due to this underestimate; this does not account for the models' systematic dependence on Distance, however. The AVS model was used for comparison, based on a virtual bird's-eye view. With access to depth information about the stacks, the AVS model outperformed the fuzzy models for the *front* relationship but was considerably worse than the fuzzy models for the *back* relationship, due to asymmetrical ratings by the experiment participants. We leave these issues for future work.

Conclusion

The long-term thrust of this area of research is to give robots a language of spatial relationships that are consistent with human understanding. This can facilitate human-robot interaction and potentially improve a robot's ability to interpret human activity or designed environments. Some of the work cited in this paper goes much farther toward this goal than we have here, and another direction for future work is to determine how best to integrate our results with theirs.

Our results are nevertheless informative. One of the challenges in determining spatial relationships is the uncertainty of data in dealing with an egocentric view; another is in noisy sensor data. A step toward this goal is to identify a technique that gives results in line with human understanding. This paper compared three fuzzy methods with human survey data to find if any of the techniques performed acceptably against human perception. These techniques were developed for judgments about orthogonal presentations and performed approximately as expected from a viewer perspective.

Only four primitive spatial relationships were used in this work; many more would need to be addressed in an effective vocabulary: near, far, surrounding, inside, outside, and so forth. Another direction for future research is to determine the minimum amount of information the robot must sense in the environment before being able to make accurate predictions about the spatial relationships. Our work used three points on each object. With more points some of our models described might have produced better predictions. The work presented here compares relatively straightforward methods of determining spatial relationships given the current scene available. But if the robot moves a new scene is presented and any information from previous scenes is not incorporated into the current calculations of spatial relationships. This could play a part in determining spatial relationships with human robot interaction. A final interesting question is whether people use only available information in the picture to determine the spatial relationships between objects or whether they incorporate background knowledge or previous experience.

Acknowledgments

We would like to thank two anonymous reviewers for suggestions that improved our understanding of our results. This work was partly supported by the Robotics Collaborative Technology Alliance of the U.S Army Research Laboratory.

References

Ball, J. T. (2015). *Toward a logical description of Double-R grammar.* (www.doublertheory.com/double-r-grammar/logical-description.pdf)

Fields, M., Lennon, C., Martin, M., & Lebiere, C. (2017). Priming for autonomous cognitive systems. In *Proc. SPIE 10194.*

Gapp, K.-P. (1995). An empirically validated model for computing spatial relations. In *Proc. KI* (pp. 245–256).

Guadarrama, S., Riano, L., Golland, D., Go, D., Jia, Y., Klein, D., ... Darrell, T. (2013). Grounding spatial relations for human-robot interaction. In *Proc. IROS* (pp. 1640–1647).

Gunzelmann, G., & Lyon, D. R. (2006). Mechanisms for human spatial competence. In *Proc. Spatial Cognition* (pp. 288–307).

Harrison, A. M., & Schunn, C. D. (2003). ACT-R/S: Look ma, no "cognitive-map"! In *Proc. ICCM* (pp. 129–134).

Keller, J. M., & Wang, X. (1995). Comparison of spatial relation definitions in computer vision. In *Proc. Uncertainty Modelling and Analysis* (p. 679).

Matsakis, P., & Wendling, L. (1999). A new way to represent the relative position between areal objects. *IEEE Transactions On PAMI*, 634-643.

Regier, T., & Carlson, L. A. (2001). Grounding spatial language in perception: an empirical and computational investigation. *JEP: General*, *130*(2), 273.

Tellex, S. A., Kollar, T. F., Dickerson, S. R., Walter, M. R., Banerjee, A., Teller, S., & Roy, N. (2011). Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of AAAI.*

Trafton, G., & Harrison, A. (2011). Embodied spatial cognition. *Topics in Cognitive Science*, *3*(4), 686–706.