Grounding Spatial Language for Video Search

Stefanie Tellex
MIT Media Lab
75 Amherst St, E14-574M
Cambridge, MA 02139
stefie10@media.mit.edu

Thomas Kollar
The Stata Center, MIT CSAIL
32 Vassar St, 32-331
Cambridge, MA 02139
tkollar@mit.edu

George Shaw
MIT Media Lab
75 Amherst St, E14-474
Cambridge, MA 02139
gshaw@media.mit.edu

Nicholas Roy
The Stata Center, MIT CSAIL
32 Vassar St, 32-330
Cambridge, MA 02139
nickroy@mit.edu

Deb Roy
MIT Media Lab
75 Amherst St, E14-574G
Cambridge, MA 02139
dkroy@media.mit.edu

ABSTRACT

The ability to find a video clip that matches a natural language description of an event would enable intuitive search of large databases of surveillance video. We present a mechanism for connecting a spatial language query to a video clip corresponding to the query. The system can retrieve video clips matching millions of potential queries that describe complex events in video such as "people walking from the hallway door, around the island, to the kitchen sink." By breaking down the query into a sequence of independent structured clauses and modeling the meaning of each component of the structure separately, we are able to improve on previous approaches to video retrieval by finding clips that match much longer and more complex queries using a rich set of spatial relations such as "down"and "past." We present a rigorous analysis of the system's performance, based on a large corpus of task-constrained language collected from fourteen subjects. Using this corpus, we show that the system effectively retrieves clips that match natural language descriptions: 58.3% were ranked in the top two of ten in a retrieval task. Furthermore, we show that spatial relations play an important role in the system's performance.

Categories and Subject Descriptors

H.3 [**Information Search and Retrieval**]: Search process

Keywords

video retrieval, spatial language

General Terms

algorithms, experimentation, measurement

ICMI-MLMI'10, November 8–12, 2010, Beijing, China.

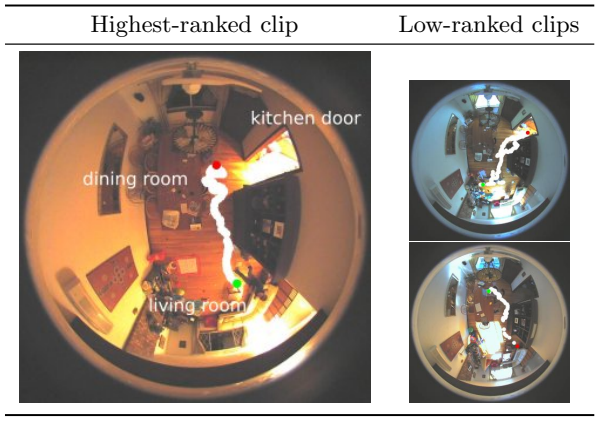

Figure 1: Results from the system for the query "from the couches in the living room to the dining room table." The person's start location is marked with a green dot; the end location is marked with a red dot, and their trajectory is marked in white.

1. INTRODUCTION

In the United States alone, there are an estimated 30 million surveillance cameras installed, which record four billion hours of video per week [15]. Analyzing and understanding the content of so much video data by hand is expensive and time-consuming. To address this problem, many have developed tools for searching video with natural language (e.g., [1, 10, 14]). In the ideal case, the user types a natural language description of the event they wish to find, and the system finds clips that match the description. Such an interface would enable natural and flexible queries, enabling untrained users to find events of interest.

However, open-ended language understanding is a challenging problem, requiring the ability to sense complex events in video and map those events to natural language descriptions. To make progress on this problem we focus on spatial language search of people's motion trajectories which are automatically extracted from video recorded by stationary overhead cameras. The system takes as input a natural language query, a database of surveillance video from a particular environment and the locations of non-moving objects in

the environment. It parses the query into a semantic structure called a *spatial description clause* (SDC) [7]. Using a model for the joint distribution of a natural language query (represented as a sequence of SDCs) and trajectories, it finds $p(query, track)$ for each trajectory in the corpus and returns video clips sorted by this score, performing ranked retrieval. The system can find video clips that match arbitrary spatial language queries, such as "People walking down the hall into the living room," by leveraging the decomposition of the language into SDCs and background knowledge found in large online databases. Video clips for a sample query returned by the system are shown in Figure 1.

The contribution of this paper is an approach for retrieving paths corresponding to millions of different natural language queries using relatively unconstrained spatial language, solving a significantly more difficult and complex problem than previous approaches (e.g., [14]). The system builds on the model described by Kollar et al. [7] for following natural language directions. We present extend our previous work by adding a model for connecting landmark phrases such as "the couch in the living room" that takes into account head nouns and modifiers and add additional spatial relations such as "past" and "down".

We perform a rigorous evaluation of the system's performance using a corpus of 696 natural language descriptions collected from fourteen annotators on a dataset of 100 video clips. Annotators viewed video clips with a person's location marked in each clip and typed a natural language description of this activity. Given a natural language query, as seen in Figure 9, we show that the correct clip will be in the top two clips of ten 58.3% of the time.

2. RELATED WORK

Our system transforms spatial language queries into a sequence of *spatial description clauses*, which we introduced in [7]. This work applies the direction understanding model from our previous work to the problem of video event recognition and retrieval, requiring the ability to work in continuous space with uncertain event boundaries. Furthermore, we present an analysis of the contribution of spatial relations to system performance at this task, showing that spatial relations are more important for this problem than for direction understanding.

Others have developed video retrieval systems, using both linguistic and nonlinguistic interfaces. Tellex and Roy [14] developed a phrase-based retrieval system; the current work moves beyond phrases and takes as input one or more entire sentences as queries. Fleischman et al. [1] built a system that recognizes events in video recorded in the kitchen. Our system also uses classifiers to recognize events, but focuses on events that match natural language descriptions rather than finding higher level patterns of activity.

More generally, Naphade et al. [10] describe the Large-Scale Concept Ontology for Multimedia (LSCOM), an effort to create a taxonomy of concepts that are automatically extractable from video, that are useful for retrieval, and that cover a wide variety of semantic phenomena. Retrieval systems such as the one described by Li et al. [8] automatically detect these concepts in video and map queries to the concepts in order to find relevant clips. This paper describes a complementary effort to recognize fine-grained spatial events in video by finding movement trajectories that match a natural language description of motion.

Ren et al. [11] review video retrieval methods based on matching spatio-temporal information. They describe symbolic query languages for video retrieval, trajectory-matching approaches, and query-by-example systems. Our system uses natural language as the query language: users describe their information need, and the system finds clips that match that description.

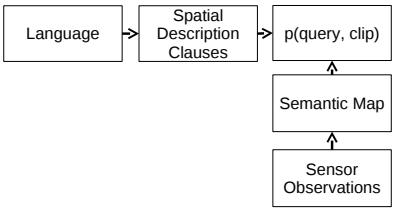

Figure 2: Data flow for the system.

Katz et al. [5] built a natural language interface to a video corpus which can answer questions about video, such as "Show me all cars leaving the garage." Objects are automatically detected and tracked, and the tracks are converted into an intermediate symbolic structure based on Jackendoff [4] that corresponds to events detected in the video. Our work focuses on handling complex spatial descriptions, while they focus on answering questions about events in the video. Harada et al. [2] built a system that finds images that match natural language descriptions such as "a cute one" with color features; our work focuses on spatial language describing trajectories rather than object or landmark descriptions.

Researchers have developed video retrieval interfaces using non-linguistic input modalities which are complementary to linguistic interfaces. Ivanov and Wren [3] describe a user interface to a surveillance system that visualizes information from a network of motion sensors. Users can graphically specify patterns of activation in the sensor network in order to find events such as people entering through a particular door. Yoshitaka et al. [16] describe a query-by-example video retrieval system that allows users to draw an example object trajectory, including position, size, and velocity, and finds video clips that match that trajectory. Natural language text-based queries complement these interfaces in several ways. First, queries expressed as text strings are easily repeatable; in contrast, it is difficult to draw (or tell someone else to draw) the exact same path twice in a pen-based system. Second, language can succinctly express paths such as "towards the sink", which would need to be drawn as many radial lines to be expressed graphically. The combination of a pen-based interface and a natural language interface is more powerful than either interface on its own.

3. GROUNDING NATURAL LANGUAGE

In order to perform video retrieval, the system needs to compute $p(query, clip)$ for each video clip in our database. The system takes as input a database of trajectories which were automatically extracted from surveillance video recorded in a particular environment, and the locations and geometries of observed objects in the environment, such as couches, chairs and bedrooms. Dataflow for the system is shown in Figure 2. Trajectories are represented as a sequence of (x, y) locations corresponding to a person's movement through the environment, as captured by the cameras.

The system also takes as input the locations of landmark objects in the environment. Although automatic object detection could be used in conjunction with overhead cameras, we used manual labels to focus on the natural language understanding component of the system. These explicitly labeled landmarks are used to bootstrap resolution of landmark phrases that appear in queries, enabling the system to

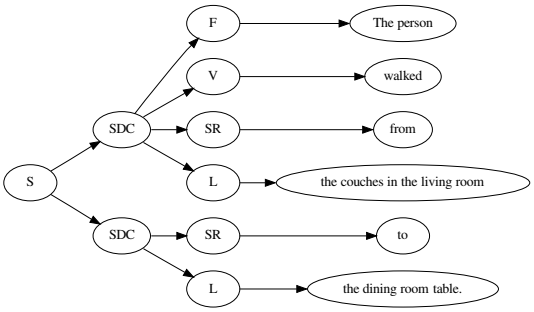

Figure 3: SDCs automatically extracted for the sentence "The person walked from the couches in the living room to the dining room table." (An annotator created this sentence for the clip shown in Figure 1.) An SDC consists of a figure F, a verb V, a spatial relation SR, and a landmark L.

infer the locations of unobserved objects based on directly observed objects.

The system retrieves video clips that match a natural language description using a probabilistic graphical model that maps between natural language and paths in the environment [7]. When performing video retrieval, clips are returned in order according to the joint probability of the query and the clip. Thus, for each video clip in our database, we want to compute: $p(query, clip)$. This distribution is modeled in terms of the fields of a semantic structure extracted from the query called a *spatial description clause* (SDC). An SDC consists of a figure, a verb, a spatial relation, and a landmark, illustrated in Figure 3. The system extracts SDCs automatically using a conditional random field chunker.

To compute the joint distribution of queries and clips, we first rewrite it in terms of the SDCs extracted from the query, and represent the person's movement in the video clip as a trajectory t consisting of (x, y) locations.

$$p(query, clip) = p(SDC_1 \dots SDC_N, t) \tag{1}$$

Assuming SDCs are independent of each other we have:

$$p(query, clip) = \prod_i p(SDC_i, t) \tag{2}$$

We assume the spatial relation and landmark fields of the SDC are referring to a particular object in the environment. However, we do not know which one, so we marginalize over all the possibilities. In general, an SDC_i may apply to only part of a trajectory, especially for longer trajectories. However, we approximate this alignment problem by assuming each SDC applies to the entire trajectory.

$$p(SDC, t) = \sum_O p(SDC, t, o) \tag{3}$$

We can rewrite the inner term in terms of the fields of the SDC (figure f, verb v, spatial relation s and landmark l), and factor it:

$$p(SDC, t, o) = p(f, v, sr, l, t, o) \tag{4}$$
$$= p(f|t, o)p(v|t, o)p(sr|t, o)p(l|t, o)p(t, o) \tag{5}$$

We assume the fields of the SDC are independent of each other and depend only on the ground object and the trajectory. $p(f|t, o)$ is treated as uniform, although it could be learned based on features in the video (e.g., mapping words like "She" to visual features indicating the gender of the person in the video clip). $p(v|t, o)$ captures how well a particular verb describes a trajectory. In our corpus, the verb is most frequently a variation of "go" or "walking" and carries little semantic information, so we treat it as uniform. The following sections describe how the other terms in Equation 5 are modeled.

3.1 Landmarks

In order to ground landmark objects, we need to estimate the probability that a landmark noun phrase l such as "the couches in the living room" could be used to describe a concrete object o with a particular geometry and location in the environment, given a trajectory t. Assuming l is independent of t, we want to estimate:

$$p(l|t, o) = p(l|o)p(o|t)p(t) \tag{6}$$

We assume a uniform prior on trajectories, $p(t)$. $p(o|t)$ is a mask based on whether a particular object is visible from the trajectory. We assume that objects not visible from a particular trajectory are never used as landmark objects in a description.

To model $p(l|o)$, we extract and stem nouns, adjectives, and verbs from the landmark phrase, representing the landmark phrase l as a set of words w_i.

$$p(l|o) = p(w_1 \dots w_M|o) \tag{7}$$

For example, the words extracted from "the couches in the living room" are "couch," "living," and "room." We assume that one of the keywords is referring to the physical landmark object, and the other keywords are descriptors or reference objects. For "the couches in the living room," the grounded keyword is couches; this word directly grounds out as the object being referred to by the landmark phrase, and other extracted words are modifiers. However, the system does not know which keyword is the root and which are modifiers. To address this problem we represent which keyword is the root with an assignment variable $\phi \in 1 \dots M$ which selects one of the words from the landmark phrase as the root and marginalize over possible assignments:

$$p(k_1 \dots k_M|o) = \sum_\phi p(k_1 \dots k_M, \phi|o) \tag{8}$$

Expanding the inner term we have:

$$p(k_1 \dots k_M, \phi|o) = p(k_1 \dots k_M|\phi, o)p(\phi|o) \tag{9}$$
$$= \prod_i p(k_i|\phi, o)p(\phi) \tag{10}$$

The prior on ϕ is independent of the physical object o and depends on grammatical features of the landmark phrase; we treat it as uniform over all the keywords, although a more informed prior would use parse-based features to identify the head noun. The likelihood term can be broken down further depending on the value of ϕ:

$$p(k_i|\phi, o) = \begin{cases} p(k_i \text{ is } o) & \text{if } \phi = i \\ p(k_i \text{ can see } o) & \text{if } \phi \neq i \end{cases} \tag{11}$$

$p(k_i \text{ is } o)$ could be estimated using hypernym/hyponym features from Wordnet and part of speech information, but here we test whether k_i matches the label for o in the semantic map of any object in the environment; if it does, we report whether o has that tag. Otherwise, we back off to co-occurrence statistics learned from tags for over a million

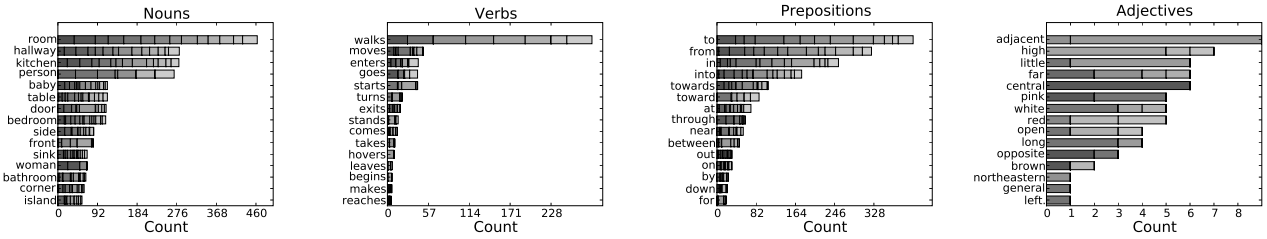

Figure 4: Histograms showing the most frequent words in the corpus for various parts of speech.

images downloaded from the Flickr website following Kollar and Roy [6]. For example, using this corpus, the system can infer which bedroom is "the baby's bedroom" without an explicit label, since only that room contains a crib and a changing table. This model uses visibility information as a proxy for detailed spatial-semantic models of relations such as "in front of" and "in."

3.2 Spatial Relations

To ground spatial relations we need to compute the probability of a spatial relation sr given a trajectory t and an object o in order to model $p(sr|t, o)$ from Equation 5. We train supervised models using labeled examples of spatial prepositions, following Tellex and Roy [14]. Each model takes as input the geometry of a path and a landmark object and outputs a probability that the situation can be described using that spatial relation. Models are trained using a library of features that capture the semantics of spatial prepositions.

3.2.1 Through

The features for "through" compute a set of axes that the figures imposes on the landmark by finding the line that connects the first and last point in the figure, and extending this line until it intersects the landmark.

- **centroidToAxesOrigin** The distance between the origin of the axes and the centroid of the landmark.
- **ratioFigureToAxes** The ratio of the distance between the start and end points of the figure and the axes it imposes on the landmark.

3.2.2 Down

As in "down the hall" or "down the road."

- **standardDeviation**: The standard deviation of the distance between the figure and the ground.
- **figureCenterOfMassToAxesOrigin**: The distance between the center of mass of points in the figure and the axes origin.
- **distAlongGroundBtwnAxes**: The distance along the ground between the start and end of the minor axis.
- **eigenAxesRatio**: The ratio between the eigenvectors of the covariance matrix of the ground when represented as an occupancy grid.

3.2.3 Past

Two of the features for "past" make use of an axes, which is computed by finding the line segment that minimizes the distance between the figure and the landmark.

- **angleFigureToAxes** The angle between the linearized figure and the line perpendicular to the axes.
- **axesLength** The length of the axes.
- **distFigureEndToGround** The distance from the end of the figure to the closest point on the landmark.
- **distFigureEndToGroundCentroid** The distance from the end of the figure to the centroid of the landmark.
- **distFigureStartToGround** The distance from the start of the figure to the centroid of the landmark.
- **distFigureStartToGroundCentroid** The distance from the start of the figure to the closest point on the landmark.

3.2.4 To

Features for "to" include **distFigureEndToGround**, and **distFigureEndToGroundCentroid**, and **minimumDistanceToGround**, the minimum distance between the figure and the landmark.

3.2.5 Towards

The features for "towards" are the same as those for "to," with one addition.

- **displacementFromGround** The difference in distance between the end of the figure to the centroid of the landmark, and the start of the figure to the centroid of the landmark.
- **axesIntersectGround** Whether the extension of a line fit to the points in the figure intersects the landmark object.

3.2.6 From

Features for from include **displacementFromGround** and **axesIntersectGround**.

4. EVALUATION

We evaluate the system's performance in two ways: we use a corpus-based dataset of natural language descriptions of video clip, and we use a much smaller set of hand-designed queries chosen to represent information needs of potential users of the system. An example description from the first corpus is "The woman entered the dining room from the living room." An example query from the second corpus is "People walking into the kitchen."

There is a tradeoff between the two evaluation strategies. In the first evaluation, we have a task-constrained corpus that consists of open-ended language collected from untrained users. We report performance on this corpus using an off-line evaluation metric, which requires much less

"The person walked from the couches in the living room to the dining room table."

"The woman entered the dining room from the living room."

"She walks from the hallway into the dining room and stands by the side of the dining room table that is nearest to the kitchen."

"The person walked from the couch in the living to the dining table in the dining room."

"The person enters the dining room from the living room and goes to the table near the entrance to the kitchen."

"She starts in the living room and walks to in front of the desk."

"The person enters the dining room from the stairway or living room area. She goes to the long side of the table nearest to the kitchen doorway."

"The person walks from the left-bottom side of the dining room table over tot he[*sic*] shelves."

Figure 5: Natural language descriptions created by subjects for the video clip shown in Figure 1. We gave these descriptions as input to the system in order to evaluate its performance.

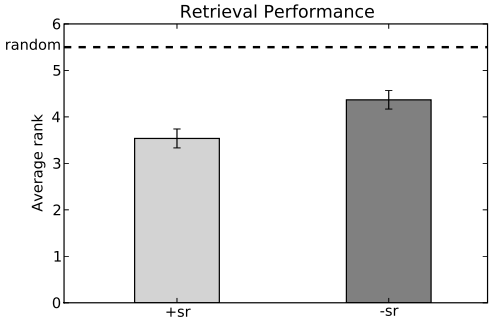

Figure 6: Results with and without models for the semantics of spatial relations, on the entire corpus.

annotation effort than traditional information retrieval metrics such as average precision, enabling us to quickly assess the system's performance on 696 queries in different configurations without performing additional annotation. However this metric and corpus may not accurately reflect retrieval performance when the system is faced with queries from real users. To address this issue we designed the second corpus, a small set of queries based on information needs of potential users. The system searched a database of video clips for matches to these queries, and we report average precision at this task. This metric more accurately reflects retrieval performance, but requires much more annotation, making it impractical to use in our much larger open-ended corpus. Together, the two methodologies show that the system is robust to input from untrained users, and that it can successfully perform ranked retrieval on realistic queries.

4.1 Task Constrained Corpus

To collect a corpus of natural language descriptions paired with video clips, annotators were shown a video clip and

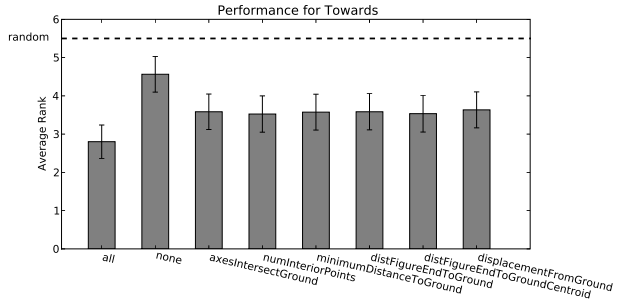

Figure 7: Results with for "towards," showing all features, no features, and each feature individually. Error bars are 90% confidence intervals.

asked to describe the motion of a person in the clip. Video was collected from eleven ceiling-mounted cameras and fourteen microphones were installed in a home as part of an effort to understand a child's language acquisition [12]. Sample frames from the corpus are shown in Figures 1 and 9.

Movement traces, or tracks are generated using a motion-based tracking algorithm [13]. Pixels representing movement in each video frame are collected into dense patches, or particles, with these particles providing probabilistic evidence for the existence of a person. These particle detections allow models to be built up over time. By correlating each such model from frame to frame, we can efficiently and robustly track the movement of people in a scene.

In order to collect a corpus of natural language descriptions of tracks, a larger database of tracks was sampled to extract two datasets of fifty tracks. The tracks were created by sampling 10 random 2.5 second clips, 10 random 5 second clips, 10 random 10 second clips, 10 random 20 second clips, and 10 random 40 second clips from the first five word births. Clips were constrained to end at least two meters from where they started, to ensure that the clip contained at least some motion. (Otherwise, many tracks consisted of a person sitting at a table or on a couch, and never moving.) The first dataset allowed clips to overlap in time in order to collect more than one description from the same person for the same track at different granularities. The second dataset had no overlaps in time to collect a more diverse database. The clips were collected randomly from eight cameras installed in each room of the main floor of the house, but each individual clip was from a single camera.

Fourteen annotators were recruited from the university community to view each clip and describe the activity of a person in the clip. Annotators viewed each clip, with the location of the person being tracked marked by a large green dot on each frame of the clip. They were instructed to describe the motion of the person in the video so that another annotator could draw their trajectory on a floor plan of the house. We showed each annotator a floor plan of the house to familiarize them with the layout and how the scenes from each camera connected. We did not ask them to restrict their language in any way, but rather use whatever language they felt appropriate to describe the person's motion. At times the automatic person tracker made errors. Annotators were instructed to mark tracks where the automatic person tracker made significant errors. They skipped on average 6.5/50 tracks, implying the tracker worked fairly well most of the time.

Sample descriptions from the corpus are shown in Fig-

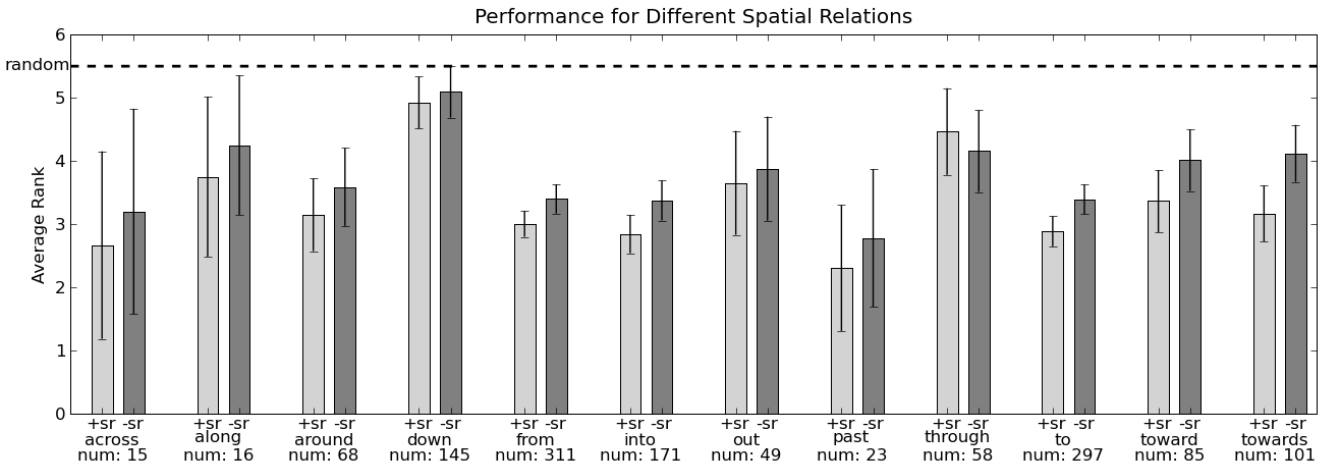

Figure 8: Results with and without models for each spatial relation, using only descriptions in the dataset that contained the particular relation. Error bars are 90% confidence intervals.

ure 5. Annotators' vocabulary was constrained only by the task. They used full sentences, whatever landmark phrases they felt were appropriate and were not instructed to use a particular vocabulary. A histogram of the fifteen most frequent words for different parts of speech appears in Figure 4. Annotators used mostly nouns and spatial relations to specify landmarks with relatively few adjectives.

4.1.1 Results

We report the model's performance in different configurations to analyze the importance of different spatial relations to the system's overall performance. In order to assess the system's performance in different configurations, we developed an evaluation metric based on a ranked retrieval task. For each natural language description in our corpus, we create a dataset of 10 tracks, containing the original track the annotator saw when creating the description and nine other random tracks. The system computes $p(query, track)$, using the description as the query for all ten tracks and sorts the clips by this score. We report the average rank of the original clip in this list over all 696 descriptions in our corpus. If the description is treated as a query, the original clip should have a high rank in this list, since it should match the query better than the other random clips. A system that ranks randomly out of ten would have an average rank of 5.5, marked with a dotted line on the graph. We report 90% confidence intervals in all graphs.

For the first experiment, we compared the system's performance with and without spatial relations. Without spatial relations, it uses only the landmark field to match the video clip to the person's trajectory. The results in Figure 6 indicate that spatial relations significantly improved the performance of the overall system. With spatial relations, 406 (58.3%) of descriptions were ranked one or two; without spatial relations, only 39.9% were ranked one or two. This result shows that spatial relations capture an important part of the semantics of the trajectory descriptions.

Next, Figure 8 shows the performance of the system when run on only those descriptions in the corpus that contain the labeled spatial relations. Only spatial relations for which we have a model, and for which more than 10 examples appeared in the corpus are shown. In almost all cases, the model of the spatial relation decreases the average rank, improving retrieval performance. Although this result is often not significant there is a consistent positive effect for spatial relations; that the overall trend is significant can be seen in Figure 6.

This automatic evaluation metric does not work perfectly. For example, our classifier for "down" performs well, as measured on its cross-validated training set and in the results presented in Table 1. However, in the corpus, "down" was almost always used in the context of "down the hallway." The overall performance of "down" is poor according to this metric because there were many examples of people walking down the hallway in our corpus of video clips; these clips were ranked higher than the original clip because they also matched the natural language description.

Next we investigated the contribution of individual features in the models for the meanings of spatial prepositions to the system's overall performance. We did this comparison on the subset of descriptions that contained that particular spatial relation. The results for "towards" are shown in Figure 7. Here it can be seen that the model using all features outperforms any model trained with a single individual features, showing that information is being fused from multiple features to form the semantics of the spatial relation.

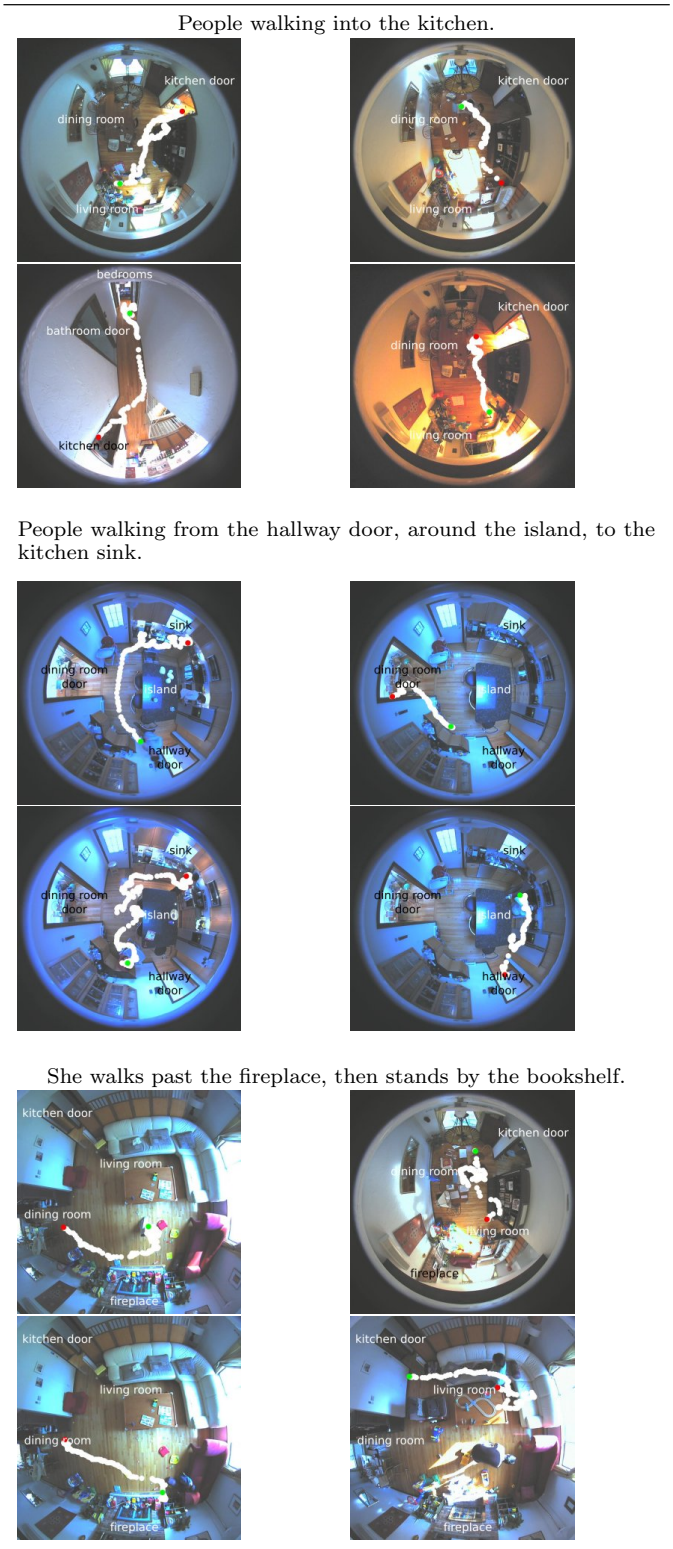

Figure 9: Clips returned for various queries. High-scoring examples are shown on the left; low-scoring examples are shown on the right.

Query	Avg. Precision
People coming out of the bathroom.	0.833
People walking into the baby's bedroom.	0.917
People walking down the hall.	0.967
People walking around the table. in the living room	1.000
People walking into the kitchen.	1.000
People walking out of the kitchen.	0.704
People walking from the hallway door, around the island, to the kitchen sink.	0.583
Mean Average Precision	0.858

Table 1: Average precision for representative queries.

4.2 Performance on Particular Queries

In order to assess the performance of the end-to-end system, we ran it on several representative queries on a dataset of fifty tracks. Tracks were returned in order according to $p(query, track)$. We report performance using average precision, a measure commonly used to report the performance of ranked retrieval systems [9]. It is computed by averaging precision at rank r for all $r \in R$, where R is all relevant documents for a particular query. This metric captures both precision and recall in a single number and reflects how well the system is ranking results. The highest possible value in our corpus is 1, if all relevant documents are ranked before irrelevant documents; the lowest value is 0.02, if there was only one relevant document that was returned last. In order compute this metric, we made relevance judgments for each query: for each of the fifty tracks in our dataset we annotated whether it matched the query or not. Results are reported in Table 1.

We chose queries that seemed to reflect particular information needs. For example, doctors monitoring the health of elders are interested in renal failure, and might be concerned with how frequently they use the restroom or how frequently they enter the kitchen to eat. Social scientists or interior designers might be curious about how people use the space and how to lay it out better.

Average precision is generally quite high. These results indicate that the system is successfully retrieving video clips from a large dataset of trajectories for representative queries that are useful for answering real-world questions.

Finally, Figure 9 shows example of high and low-scoring trajectories returned by the system for various queries. This concretely shows that the system is correctly making fine distinctions in the semantics of various queries.

5. CONCLUSION

In this paper we have presented a system that can map between natural language descriptions of motion and video clips. The system has been evaluated on a large corpus of queries, and we have shown that spatial relations contribute

significantly to the system's overall performance.

In the future, we wish to develop systems that retrieve clips based on a general natural language description of a person's activity. A key challenge lies in developing models for verbs. In our current corpus, people used words such as "enters," "hovers," and "takes." Although these could all be modeled spatially using the feature set we have developed for spatial relations, this modeling step requires collecting a separate corpus of positive and negative examples for each verb. To address this problem, we are developing models that can learn the meaning of words with less supervision, using only a corpus of descriptions paired with trajectories.

Moving beyond spatial verbs, when we asked an annotator to describe a person's general activity rather than movement, she used phrases such as "preparing food," "cleaning up," and "opening the refrigerator." As computer vision and tracking improve, a richer variety of information will be available to query engines and models of language understanding need to be developed to exploit it. In order to handle these types of language, we need to apply richer models of linguistic structure, and richer probabilistic models that capture the relationship between plans, goals, and actions.

Despite these limitations, our evaluation has shown that the system understands not just a few carefully chosen queries, but rather can take as input millions of potential trajectory-based queries and robustly find matching trajectories. We analyze the system's performance, measuring the contribution of different spatial relations to the system's overall performance, as well as the components of models for various spatial relations. Our system breaks down a complex spatial language query into components, models each components, and then finds video clips in a large corpus corresponding to the natural language description.

6. ACKNOWLEDGMENTS

We are grateful for the support of the Office of Naval Research, which supported Thomas Kollar and Stefanie Tellex under MURI N00014-07-1-0749.

References

[1] M. Fleischman, P. DeCamp, and D. Roy. Mining temporal patterns of movement for video content classification. In *Proceedings of the 8th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2006.

[2] S. Harada, Y. Itoh, and H. Nakatani. Interactive image retrieval by natural language. *Optical Engineering*, 36 (12):3281–3287, Dec. 1997.

[3] Y. A. Ivanov and C. R. Wren. Toward spatial queries for spatial surveillance tasks. In *Pervasive: Workshop Pervasive Technology Applied Real-World Experiences with RFID and Sensor Networks (PTA)*, 2006.

[4] R. S. Jackendoff. *Semantics and Cognition*, pages 161–187. MIT Press, 1983.

[5] B. Katz, J. Lin, C. Stauffer, and E. Grimson. Answering questions about moving objects in surveillance videos. In M. Maybury, editor, *New Directions in Question Answering*, pages 113–124. Springer, 2004.

[6] T. Kollar and N. Roy. Utilizing object-object and object-scene context when planning to find things. In *IEEE International Conference on Robotics and Automation*, 2009.

[7] T. Kollar, S. Tellex, D. Roy, and N. Roy. Toward understanding natural language directions. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, 2010.

[8] X. Li, D. Wang, J. Li, and B. Zhang. Video search in concept subspace: a text-like paradigm. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 603–610, Amsterdam, The Netherlands, 2007. ACM.

[9] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999.

[10] M. Naphade, J. Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *Multimedia, IEEE*, 13(3):86–91, 2006.

[11] W. Ren, S. Singh, M. Singh, and Y. Zhu. State-of-the-art on spatio-temporal information-based video retrieval. *Pattern Recognition*, 42(2):267–282, Feb. 2009.

[12] D. Roy, R. Patel, P. DeCamp, R. Kubat, M. Fleischman, B. Roy, N. Mavridis, S. Tellex, A. Salata, J. Guinness, M. Levit, and P. Gorniak. The Human Speechome Project. In *Proceedings of the 28th Annual Cognitive Science Conference*, pages 192–196, 2006.

[13] G. Shaw. Efficient multiple objectr tracking using motion features. Technical report, MIT Media Lab, 2010.

[14] S. Tellex and D. Roy. Towards surveillance video search by natural language query. In *Conference on Image and Video Retrieval (CIVIR-2009)*, 2009.

[15] J. Vlahos. Welcome to the panopticon. *Popular Mechanics*, 185(1):64, 2008.

[16] A. Yoshitaka, Y. Hosoda, M. Yoshimitsu, M. Hirakawa, and T. Ichikawa. Violone: Video retrieval by motion example. *Journal of Visual Languages and Computing*, 7:423–443, 1996.