Grounding Spatial Prepositions for Video Search

Stefanie Tellex
MIT Media Lab
20 Ames St. E15-486
Cambridge, MA, 02139
stefie10@media.mit.edu

Deb Roy
MIT Media Lab
20 Ames St. E15-488
Cambridge, MA 02139
dkroy@media.mit.edu

ABSTRACT

Spatial language video retrieval is an important real-world problem that forms a test bed for evaluating semantic structures for natural language descriptions of motion on naturalistic data. Video search by natural language query requires that linguistic input be converted into structures that operate on video in order to find clips that match a query. This paper describes a framework for grounding the meaning of spatial prepositions in video. We present a library of features that can be used to automatically classify a video clip based on whether it matches a natural language query. To evaluate these features, we collected a corpus of natural language descriptions about the motion of people in video clips. We characterize the language used in the corpus, and use it to train and test models for the meanings of the spatial prepositions "to," "across," "through," "out," "along," "towards," and "around." The classifiers can be used to build a spatial language video retrieval system that finds clips matching queries such as "across the kitchen."

Categories and Subject Descriptors

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

Keywords

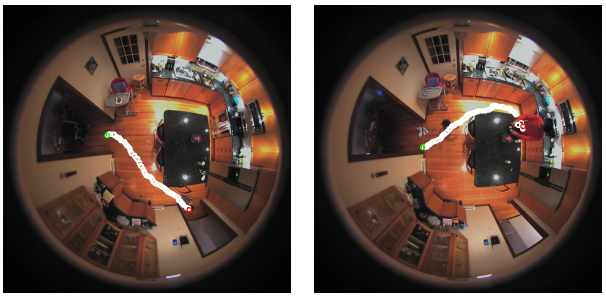
video retrieval, spatial language

General Terms

algorithms, experimentation, measurement

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[22]. However, analyzing and understanding the content of video data remains a challenging problem. To address aspects of this problem, we are developing interfaces that allow people to use natural language queries to naturally and flexibly find what they are looking for in video collections.



(a) (b)

Figure 1: Frames from two clips returned for the query "across the kitchen."

Our system is a multi-modal user interface that finds video clips in surveillance video containing people moving in ways that match natural language queries such as "to the refrigerator" and "across the kitchen." A core problem in building a natural language query system for surveillance video is encoding robust visually-grounded models of the meaning of spatial prepositions. In our approach, the meanings of spatial prepositions are modeled by visual classifiers that take spatial paths as input. These classifiers are trained using labeled path examples. Continuous geometric paths of people in video are converted into a set of features motivated by theories of human spatial language [7, 11, 20]. In this paper we focus on spatial prepositions that describe motion, such as "around," "across," and "to."

In order to train and evaluate our models, we collected a corpus of natural language descriptions of video clips. Our video corpus consists of data recorded from a fish-eye camera installed in the ceiling of a home[19]. Sample frames from this corpus, retrieved by the system for the query "across the kitchen," are shown in Figure 1. To associate video clips with a natural language description, annotators were asked to write a short phrase describing the motion of a person in the clip. Using this corpus of descriptions paired with video clips, we trained models of the meanings of some spatial prepositions, and explored their semantics by analyzing which features are most important for good classification performance.

Although our data consists of only video from a home, it is representative of a much larger class of domains. Air-

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.

ports, retailers, and many other organizations are amassing millions of hours of video from statically placed surveillance cameras. Beyond video, our spatial-semantic models may be applied to other kinds of space-time data, from searching GPS logs to understanding natural language directions.

Previous work in video surveillance has focused on tracking objects in video (e.g., [5, 25]), automatically recognizing unusual events in video such as unattended luggage in public areas or unusual behavior in a home (e.g., [1, 8]), and integrated retrieval interfaces (e.g., [6, 23]). Our work points towards a method of bridging the semantic gap in surveillance video retrieval by enabling users to type a natural language description of the activity in the video, and find clips that match that description.

2. RELATED WORK

Our system transforms spatial language queries into a function/argument structure based on the theories of Jackendoff [7], Landau and Jackendoff [11], and Talmy [20]. Following Talmy [20], we refer to the person being described by a sentence such as "The person is going to the sink" as the *figure*, and the noun phrase argument as the *ground*. The features in our classifiers are inspired from their work.

Others have implemented and tested models of spatial semantics. Regier [17] built a system that assigns labels such as "through" to a movie showing a figure moving relative to a ground object. His system learned the meanings of many spatial prepositions across different languages, and tests the models on schematic videos. Our system uses some of the same features, but tests our model using annotations of real video. Kelleher and Costello [10] built a model for the meanings of static spatial prepositions, which primarily denote the location of an object. Their model takes into account the influence of distractor objects in the generation and understanding of phrases such as "the ball near the red box." They used their model to enable a robot to engage in visually situated dialog about a table-top environment. Our work focuses on dynamic spatial prepositions describing paths, and evaluates our models by applying them to a corpus of natural language descriptions of movement.

Fleischman et al. [3] built a system that recognizes events in video recorded in the kitchen. Their system learns hierarchical patterns of motion in the video, creating a lexicon of patterns. The system uses the lexicon to create feature vectors from video events, which are used to train a classifier that can recognize events in the video such as "making coffee." 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. [15] 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 Li et al. [12] automatically detect these concepts in video, and map queries to the concepts in order to find relevant clips. In contrast to our work, LSCOM focuses on open-class coarse-grained semantic events for retrieval from corpora of broadcast news, including movement categories such as "Exiting_A_Vehicle" and "People_Marching." This paper describes a complementary effort to recognize fine-grained spatial events in video by finding movement trajectories that match a natural language description.

Ren et al. [18] 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 work points towards a system that uses a subset of natural language as a query language: users describe their information need, and the system finds clips that match that description.

Katz et al. [9] 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 [7] that corresponds to events detected in the video. Our work focuses on handling complex spatial prepositions such as "across" while they focus on understanding a range of questions involving geometrically simpler prepositions. Harada et al. [4] built a system that finds images that match natural language descriptions such as "a cute one" with color features.

Researchers have developed video retrieval interfaces using non-linguistic input modalities which are complementary to linguistic interfaces. Ivanov and Wren [6] 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. [24] 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. SYSTEM ARCHITECTURE

Our system finds video clips that match natural language queries such as "across the kitchen" and "towards the door." When a user enters a natural language query, the system first parses it, then uses a classifier to find matching video clips in the data. Prepositions are treated as functions which take an ordered list of points (the figure) and a polygon (the ground). The functions return a boolean value indicating whether these geometries match the preposition. For example, for the query "to the sink," the figure is the trajectory of a person in a video clip, and the ground is a polygon representing the sink. Figure 2 shows how the system processes a natural language query and retrieves matching video clips.

The functions are instantiated as binary classifiers which are learned from a corpus of natural language descriptions of video clips. A combinatory categorial grammar parser [14] extracts the function/argument structure from the query and resolves referring expressions in the query. Unique objects that rarely move such as "the kitchen" or "the coffee maker" are resolved to pre-annotated regions in the video. Some noun phrases such as "the door" or "the counter" can not be automatically resolved in this way because there are more than one in the room being recorded. For example, in

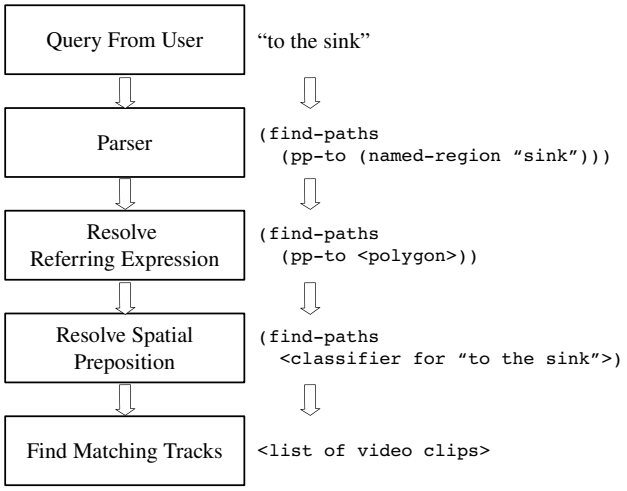


Figure 2: Data flow as the system processes a query.

Figure 3: A frame from a clip in our corpus. Descriptions for this clip included "to the counter," "along the east side of the island," "from the refrigerator", "to the cabinet," and "across the kitchen."

Figure 3, if an annotator labeled the clip "past the counter," the proper ground object is the east counter; on the other hand, if the label was "to the counter," it probably refers to the north counter. To sidestep this issue, we manually labeled ambiguous objects with the proper ground. We plan to use this corpus to train and evaluate a system to automatically resolve referring expressions such as "the door."

Other noun phrases referred to parts of objects, most frequently to parts of the island. For example, one annotator labeled the clip shown in Figure 3 as "along the east side of the island," selecting just part of the island for annotation. In this case, "the island" is resolved to a polygon as described above, and the semantics for "side" encode a rule that automatically selects an appropriate geometry for the ground object.

We chose to use binary classifiers to model the meaning of spatial prepositions for two reasons. First, a binary classifier can be directly used for video retrieval: the classifier simply returns all clips in the corpus that match a particular query. Second, binary classifiers naturally capture the fact that a given situation can have more than one correct description. For example, the clip shown in Figure 3 could be described as "along the island" or "past the island." A binary classifier for each spatial preposition captures this duality.

The second argument to the classifier, the figure, is an ordered list of points extracted automatically from the video, corresponding to the motion of a person through the room. The system searches over a database of *person tracks*. People are tracked using a motion-based tracker implemented using the SwisTrack open source tracking pipeline [13]. When a person moves in the video, the tracker detects the location of the motion, and either creates a new track, or adds the detected point to an existing track. When a person stops moving, the track is ended. These boundaries are often, but not always, reasonable places to start and stop video playback, since they correspond to the start and stop of a person's motion.

Once the query has been converted to a path and a polygon, these structures are converted to a feature vector. The features are inspired by work in spatial semantics, and are different for each preposition. The specific features used in our system are described in the next section. The feature vector is given as input to a classifier which outputs whether the clip matches the natural language query.

3.1 Spatial Prepositions

We are focusing on dynamic spatial prepositions such as "across" and "to," which describe movement through space. The dynamic spatial prepositions described here were chosen because they occurred frequently in our corpus of natural language descriptions for video clips. Each preposition is instantiated as a classifier trained using labeled feature vectors extracted from a schematic representation of the video clip. The features used are described below. Many of them are quite simple to compute, and may seem obvious. A key contribution of our work is to analyze which features work best for real-world retrieval tasks, and to create a methodology for analyzing what geometric and contextual features need to be taken into account for good retrieval performance. A second contribution of our work is to build up a library of features that can be assembled to quickly create a classifier for a new spatial preposition.

Many features involve distances or averages of distances between two points. In order to make the model scale invariant, all distance values are normalized by dividing the distance by the size of the diagonal of the bounding box of the figure and the ground together or in some cases the figure alone.

3.1.1 To

Features for "to" focus on the endpoint of the figure with respect to the ground.

distFigureEndToGround The distance between the endpoint of the figure and the closest point on the border of the ground. If the figure ends inside the ground, the value is snapped to zero.

distFigureEndToGroundCentroid The distance between the end of the figure and the centroid of the ground.

endpointsInGroundBoundingBox Whether the end of the figure intersects the bounding box of the ground.

minimumDistanceToGround The minimum distance between the figure and the ground.

numInteriorPoints The number of points in the figure which are inside the ground, when the figure is divided into a sequence of 100 equally spaced points.

3.1.2 Across

An important underlying concept inherent in the meaning of many spatial prepositions is the idea of coordinate axes. "Across" has been defined as a spatial relation that takes a linear figure and planar ground, and requires the figure to be perpendicular to the major axis of the ground. [11, 20]. However this definition does not specify how to find the major axis of the ground. In many contexts, there is no single set of axes: for example, there are many paths across a square room. The system solves this problem by finding the unique axes that the figure imposes on the ground, and then quantifying how well those axes match the ground. These axes are computed by finding the line that connects the first and last point in the figure, and extending this line until it intersects the ground. The origin of the axes is the midpoint of this line segment, and the endpoints are the two points where the axes intersect the ground. Once the axes are known, the system computes features that capture how well the figure follows the axes, and how well the axes fit the ground. The features used by a decision tree learner to train a classifier for "across" are listed below.

averageDistance The average distance between the figure and the axes it imposes on the ground.

centroidToAxesOrigin The normalized distance between the origin of the axes and the centroid of the ground.

distAlongGroundBtwnAxes The distance along the perimeter of the ground between the endpoints of the axes, normalized by the perimeter of the ground. The minimum of two possible values.

figureCenterOfMassToGroundCentroid The normalized distance between the center of mass of the figure and the origin of the axes.

ratioFigureToAxes The ratio of the distance between the start and end points of the figure and the axes it imposes on the ground.

standardDeviation The standard deviation of the normalized distance between the figure and the axes.

3.1.3 Through and Out

"Through" and "out" use many of the same features as "across," including the notion of an axes. "Through" adds one feature, **peakDistance**.

peakDistance The maximum distance between the figure and the axes it imposes on the ground.

3.1.4 Along

"Along" is a spatial relation in which the figure and ground are both conceptualized as linear: the figure must be coaxial to the ground, or parallel with the ground's major axis [11, 20]. The system does a preliminary segmentation of the track by sliding a window 75% of the figure's length along the figure, and only uses the part of the figure that minimizes the average distance to the ground. In this way, the model reduces noise from the beginning or end of the path if the person starts far away from the ground object but quickly approaches it.

angleBetweenLinearizedObjects The angle between figure and ground when each is modeled by a best-fit line.

averageDistance The average distance between the figure and the ground. The algorithm steps along the figure at a fixed resolution, and for each point computes the distance to the closest point on the ground.

distEndGroundBoundary The distance between the end of the figure and the closest point on the ground.

distStartGroundBoundary The distance between the start of the figure and the closest point on the ground.

peakDistance The maximum distance between the figure and the ground.

visibleProportionFigureGround The fraction of the figure which is visible from the ground, taking into account obstacles in the environment.

3.1.5 Around

"Around" uses a subset of features from "along", plus **averageDistStartGroundDistEndGround**, which is the average of **distStartGroundBoundary** and **distEndGroundBoundary**.

4. CORPUS COLLECTION

In order to train and evaluate our models, we collected a corpus of natural language descriptions of video clips. This corpus gives insight into the types of descriptions humans use when labeling video, and provides a source of data for training and evaluating the system. We plan to make this corpus publicly available. Check the authors' home page for more information. Our aim in collecting the corpus was to pair a person's movement in a video clip with a prepositional phrase describing the movement, so that a system could use these mappings to train classifiers which model the meanings of spatial prepositions. In order to do this, we showed annotators short clips with the location of a person marked in each frame of the clip. The interface displayed the partial sentence "The person is going" with a text entry box at the end. We instructed annotators to complete the sentence with a single prepositional phrase that described the person's motion. Annotators were asked to skip the clip if there was a tracking failure, or if they could not write a description for the clip. Table 1 shows the number of tracks skipped by annotators. The video was overlayed with labels marking the location of non-moving objects such as the refrigerator, doors, and cabinets. Annotators were asked to try to use those labels in their descriptions, but were not required to use them. Our corpus contains data from five different annotators.

Annotators were shown video clips from two days of video. To focus on prepositions describing movement, we showed annotators only tracks that were longer than four seconds, where the distance between the first and last points in the track was larger than 200 pixels. Clips were shown in random order drawn from the two days of data, and each clip appeared in the data set three times in order to collect multiple descriptions from the same annotator for the same clip.

After annotation, each clip in our data set had up to fifteen descriptions associated with it, with an average of 10.7 Figure 3 shows a frame from a clip in our corpus, together with some descriptions. Figure 4 shows a histogram of the frequency of the descriptions that appeared in the corpus, color coded by annotator, while Figure 5 shows the distribution of prepositions in our corpus. From the histograms,

Corpus Size	
tracks left blank	971
grounding and parsing failures	393
parse successes	7051
total	8415

Table 1: The size of the corpus, together with numbers of tracks excluded for various reasons.

it seems that annotators tend to reuse descriptions, rather than inventing an entirely new one for each track. Despite this tendency, a diverse set of spatial prepositions appears in our corpus

Figure 6 shows the distribution of ground objects used in the corpus. Ambiguous ground objects such as "the door" and "the counter," which appeared more than once in the kitchen are resolved through manual annotations. Descriptions which resolved to more than one ground object were excluded from the evaluation. Examples of descriptions rejected for this reason include "from one counter to the other," "back and forth," and "through both doors."

All descriptions that the system successfully parsed and had associated ground objects were included in the evaluation. Table 1 shows the number of tracks that annotators skipped, and the number of parse failures for the tracks used in the evaluation.

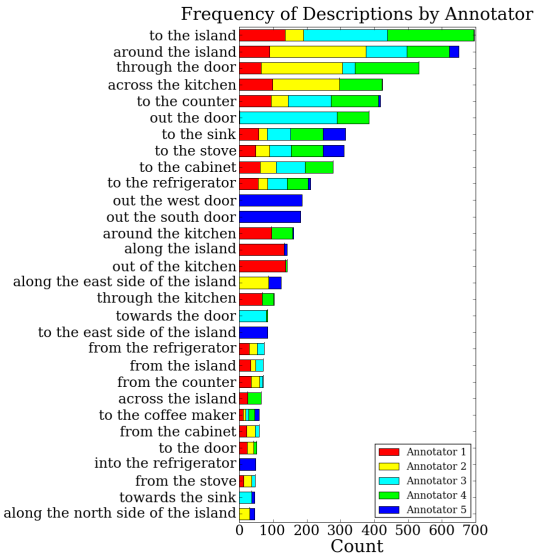


Figure 4: Histogram of the frequencies of various descriptions in the corpus.

5. EVALUATION

We used our corpus to train classifiers for spatial prepositions and evaluate the performance of the classifiers. In order to train classifiers for each preposition, each description was converted into a training example. If the description

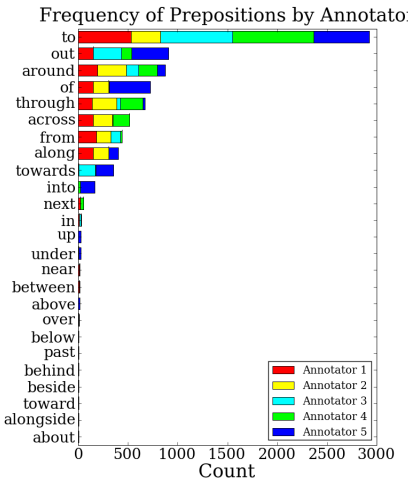


Figure 5: Histogram of the prepositions in our corpus.

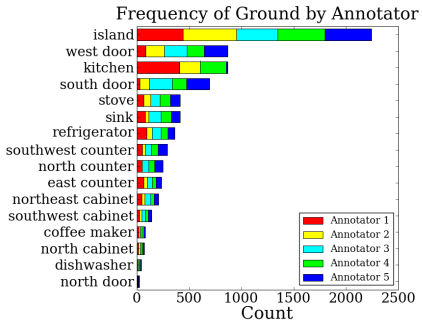


Figure 6: Histogram of ground objects used to label tracks in our corpus. Each ground corresponds to a specific object in the camera's visual field; the mapping was determined from human annotations.

used a preposition, it was treated as a positive training example for that preposition, and if the description used some other preposition, it was treated as a negative training example, following Regier [17]. For example, the description "around the island" paired with a video clip was treated as a positive example of "around," and a negative example of "to." This heuristic is imperfect: a track that is "along the island" may also be "around the island." In some of these ambiguous cases, we excluded similar spatial prepositions from the training and test sets. For example, for "to," we excluded examples labeled with "towards," "out," "through," and "into" because a track labeled "out the door" was often, in our judgement, a good example of "to the door." Data was separated into training and testing by track: all descriptions associated with a track appeared either in the training set or the test set. 80% of the tracks were used as training data, and the rest as test data.

To visualize classifier performance we report ROC curves. All results use the Naive Bayes classifier in the Orange Data

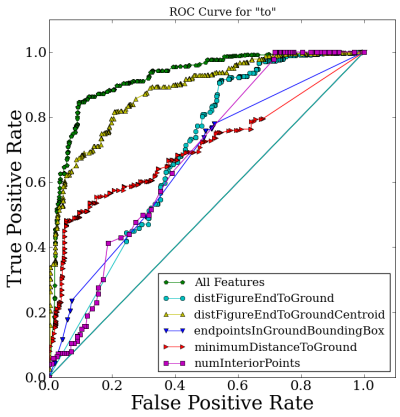


Figure 7: Performance of classifiers for "to," with examples containing "out," "through," "towards," and "into" excluded.

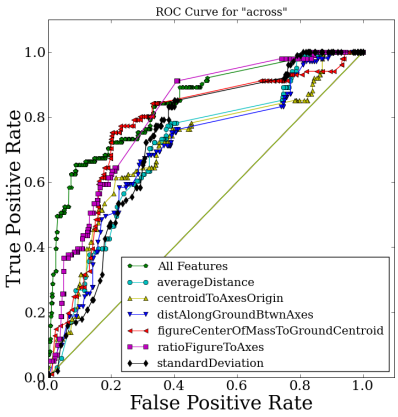


Figure 8: Performance of classifiers for "across."

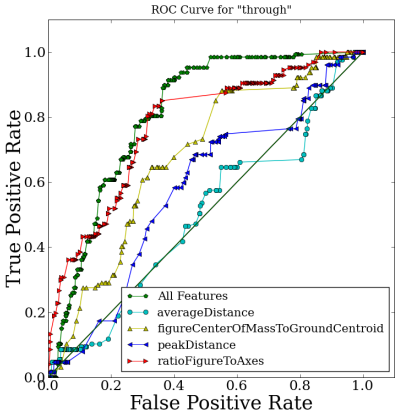


Figure 9: Performance of classifiers for "through."

Mining library [2]. We measured the performance of a classifier trained using all features, as well as one trained on each feature in isolation, to see how well each feature works on its own. It is possible a classifier would perform even better with a subset of the features. We have reported performance broken down in this way for some spatial prepositions elsewhere [21], but we chose not to do it here for brevity. Although our aim is to use our models to support video retrieval, this evaluation does not directly measure retrieval performance, but rather the effectiveness of our classifiers at capturing the semantics of spatial prepositions that might be used in natural language queries.

Figure 7 shows the performance of various binary classifiers for the spatial preposition "to." The classifier trained using all the features clearly performs the best. An alternative interface to search for people going "to the sink" is to manually specify a region of interest, for example by drawing a bounding box. The two features, **numInteriorPoints** and **endpointsInGroundBoundingBox**, capture this heuristic, and perform quite poorly on their own. This result implies that a user searching for people going "to the sink" would be better served by an explicit model of the meaning of "to," implemented in terms of a combination of features, than they would be by a query interface in which they drew a bounding box around a region of interest.

In an earlier paper [21], we analyzed "across" based on binary annotations, in which annotators marked whether a video clip matched a query such as "across the kitchen." There we found that the feature **ratioFigureToAxes** was critical to good performance, and other features performed poorly on their own. In this corpus, the feature **figureCenterOfMassToGroundCentroid** is also effective on its own. Possibly difference is due to the different tasks in the two papers: it is possible the methodology of collecting natural language descriptions for clips yields fewer borderline "across" examples, changing which features work the best.

"Through" and "out" use a subset of the features used by the "across" classifier. For "out", the feature **ratioFigureToAxes** performed the best. This feature captures the degree to which the figure moves from one point on the boundary of the ground to another point. Both of these spatial prepositions are somewhat problematic in this domain because our tracks do not extend beyond a single camera. When an annotator wrote "through the door," the system saw a track that extended to the door and then stopped. We are currently exploring the usage of these words in a corpus of natural language directions, which has a more satisfying representation, and developing person trackers that work between different cameras to create longer tracks.

The results for "along" are shown in Figure 11. We report the performance of a classifier trained on all features, and on all features except **visibleProportionFigureGround**. This feature is the only feature (so far) which requires additional context from the environment besides the geometry of the figure and ground: it must know about obstacles in the environment which can prevent the figure from being visible from the ground. We added this feature because a classifier for "along" trained on a corpus of explicitly labeled positive and negative examples of "along the right side of the island" sometimes returned tracks that were along the left side of the island. We hoped that adding features that referred to obstacles in the environment would alleviate this problem,

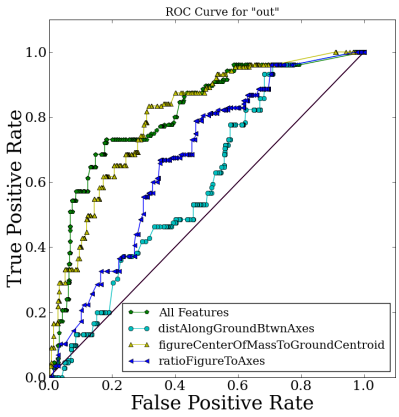


Figure 10: Performance of classifiers for "out," with examples containing "towards," "through," and "to" excluded.

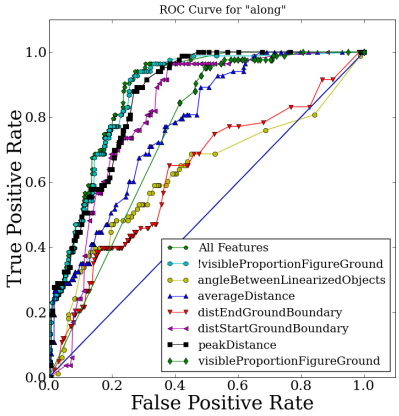


Figure 11: Performance of classifiers for "along." !visibleProportionFigureGround is a classifier trained on all features except visibleProportionFigureGround.

but so far we have not found an effect.

Figure 12 shows the performance of the classifier for "around." As it turned out, the most common use of around by far was "around the island," so although the system performs well, it probably does not generalize well to other examples. Interestingly, the feature **distStartToGround** performs much better than **distEndToGround**, despite the bias in our corpus for the spatial preposition "to" compared to "from," and despite evidence that people pay more attention to the goal of a spatial motion event[16].

Overall our results are promising. We have identified a set of features that can successfully classify examples in our corpus. Although we have not yet evaluated the classifiers in an end-to-end retrieval context, the performance on this task is encouraging because it indicates that the features are capturing important aspects of the semantics of spatial prepositions. A major remaining issue is that in a retrieval context, even a low false positive rate can yield a poor F-

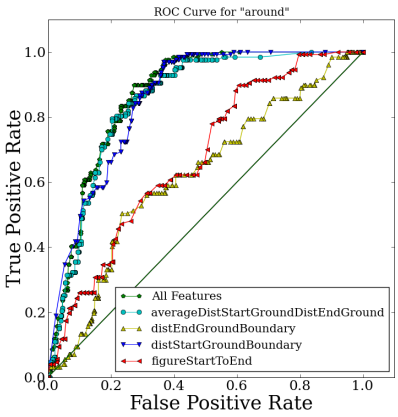


Figure 12: Performance of classifiers for "around."

scores if there are many more negative examples in the corpus than positive examples. Despite this issue, our models for the meanings of spatial prepositions are a promising path forward to a natural language video retrieval. The task is a challenging one, and we do not expect perfect performance. An interface using our models could successfully find matching clips, enabling video search by natural language query.

6. CONCLUSION

The work described here focuses on single prepositional phrases. We plan to use our models as components of a more complicated natural language understanding system that can handle chains of prepositional phrases (e.g., "from the dining room to the sink along the right side of the island") and a richer variety of words, such as "wandering," "loitering," and "pacing." Our methodology - collecting natural language descriptions of video clips and using the resulting corpus to train and test semantic models - enables us to build up robust models of meaning that are useful for solving real-world problems.

We see natural language as one component of a multimodal retrieval interface, complemented by other input modalities. In addition to natural language queries, our retrieval interface already supports several graphical query methods. Users can draw an example trajectory with the mouse, and the system finds similar clips. They can also query by bounding box, using the mouse to draw boxes indicating regions of interest. These query interfaces complement a natural language interface, making a tool more powerful than either on its own. We are also applying our models to enable a robot to understand natural language directions such as "go out the door to the computers." Our long-term goal is to build models of spatial semantics that work in many different realistic domains.

We have presented a multi-modal retrieval system that finds video clips that match natural language queries such as "to the stove," "along the right side of the island," or "across the kitchen." To train and evaluate our system, we collected a corpus of video clips paired with natural language descriptions. This corpus provides a snapshot of the ways people describe movement in video. Using this corpus, we trained and tested binary classifiers for spatial prepositions

in English, and measured their performance. This methodology enables us to identify important concepts underlying the semantics of spatial prepositions.

6.1 Acknowledgments

We would like to thank our annotators, as well as Nathan Booth, Gregory Marton, Kevin Gold, Jennifer Akana and Piotr Mitros. Stefanie Tellex was supported by the Office of Naval Research under MURI N00014-07-1-0749.

References

[1] R. Cucchiara, C. Grana, A. Prati, and R. Vezzani. Computer vision techniques for PDA accessibility of in-house video surveillance. In *First ACM SIGMM International Workshop on Video surveillance*, pages 87–97, Berkeley, California, 2003. ACM.

[2] J. Demsar and B. Zupan. Orange: From experimental machine learning to interactive data mining. Technical report, Faculty of Computer and Information Science, University of Ljubljana, 2004. URL `http://www.ailab.si/orange`.

[3] 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.

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

[5] Y. Ivanov, A. Sorokin, C. Wren, and I. Kaur. Tracking people in mixed modality systems. Technical Report TR2007-011, Mitsubishi Electric Research Laboratories, 2007.

[6] 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.

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

[8] P. Jodoin, J. Konrad, and V. Saligrama. Modeling background activity for behavior subtraction. In *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on*, pages 1–10, 2008.

[9] 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.

[10] J. D. Kelleher and F. J. Costello. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306, June 2009.

[11] B. Landau and R. Jackendoff. "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16:217–265, 1993.

[12] 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.

[13] T. Lochmatter, P. Roduit, C. Cianci, N. Correll, J. Jacot, and A. Martinoli. Swistrack - a flexible open source tracking software for multi-agent systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008.

[14] G. Marton and L. B. Westrick. Sepia: a framework for natural language semantics. Technical report, Massachusetts Institute of Technology, 2009.

[15] 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.

[16] T. Regier and M. Zheng. Attention to endpoints: A Cross-Linguistic constraint on spatial meaning. *Cognitive Science*, 31(4):705, 2007.

[17] T. P. Regier. *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization*. PhD thesis, University of California at Berkeley, 1992.

[18] 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. ISSN 0031-3203.

[19] 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.

[20] L. Talmy. The fundamental system of spatial schemas in language. In B. Hamp, editor, *From Perception to Meaning: Image Schemas in Cognitive Linguistics*. Mouton de Gruyter, 2005.

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

[22] J. Vlahos. Welcome to the panopticon. *Popular Mechanics*, 185(1):64, 2008. ISSN 00324558.

[23] T. Yamasaki, Y. Nishioka, and K. Aizawa. Interactive retrieval for multi-camera surveillance systems featuring spatio-temporal summarization. In *Proceeding of the 16th ACM international conference on Multimedia*, pages 797–800, Vancouver, British Columbia, Canada, 2008. ACM.

[24] 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.

[25] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008.