Towards Surveillance Video Search by Natural Language Query

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 is also a natural test bed for evaluating semantic structures for natural language descriptions of motion on naturalistic data. This paper describes first steps towards a system that grounds the meaning of spatial prepositions in geometric features. This system can be used to search a corpus of surveillance video for clips that match spatial language queries such as "along the hallway" and "across the kitchen." We present experiments characterizing the performance of models for the prepositions "across" and "along," and present a methodology for modeling other spatial prepositions.

Categories and Subject Descriptors

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

Keywords

video retrieval, spatial language

1. INTRODUCTION

In the United States, there are an estimated 30 million surveillance cameras installed, which record four billion hours of video per week. [21] 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.

We are building an interface that finds video clips in surveillance video that contain people moving in ways that match natural language queries such as "across the living room" and "along the right side of the island." A core problem in building a natural language query system for surveillance video is encoding robust visually-grounded models of the meaning of words such as "along" and "across". We present a framework and methodology for computationally grounding the meaning of spatial prepositions. We developed a system that uses these meanings to retrieve clips from a corpus of surveillance video that match natural language queries.

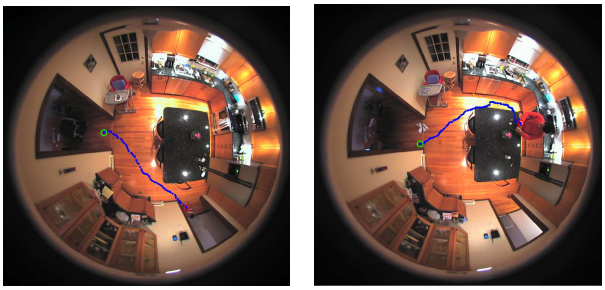
In our approach, the meanings of spatial prepositions are modeled by visual classifiers that take spatial paths as input. These classifiers are trained using labeled path examples. Continuous geometric paths of people in video are converted into a set of features motivated by theories of human spatial language [8, 11, 20]. We evaluate the models by measuring their performance in retrieving video clips that match a natural language query. As part of the evaluation we analyze which features contribute most to the system's performance, and thus best capture the semantics of the spatial preposition in a form usable by a decision tree. Features that work well in this classification task have been empirically shown to capture important aspects of the meaning of spatial prepositions. This methodology is a way to computationally specify and then evaluate theories of spatial semantics.

This work is motivated by the data analysis needs of the Human Speechome Project (HSP) [19], an effort to analyze one child's language development based on a densely sampled long-term audio-visual record of life at home. Approximately 90,000 hours of 960x960 resolution video have been recorded using fisheye lens cameras mounted in the ceiling of rooms in the child's home over a three-year period. Sample frames from this corpus, retrieved by the system for the query "across the kitchen," are shown in Figure 1.

Although this corpus is unique in its purpose and scope, the nature of the video content – people interacting indoors over extended periods of time – is representative of a much larger class of domains. Airports, 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 generating natural language directions.

Previous work in video surveillance has focused on tracking objects in video (e.g., [6, 24]), automatically recognizing unusual events in video such as unattended luggage in public areas or unusual behavior in a home (e.g., [3, 9]), and integrated retrieval interfaces (e.g., [7, 22]). 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. A more detailed literature review appears in Section 4.

CIVR '09 Fira, Santorini



(a) (b)

Figure 1: Frames from two clips returned for the query "across the kitchen." The start point of the trajectory is green, and the end point is red.

2. SYSTEM ARCHITECTURE

Our system finds video clips that match natural language queries such as "along the hallway." When a user enters a natural language query, the system first parses it, then uses a classifier to find matches in the corpus. Prepositions such as "across" and "along" are treated as two-argument functions which take an ordered list of points (the figure) and a polygon (the ground). For the query "along the hallway," the figure is the trajectory of a person in the video clip, and the ground is a polygon representing the hallway. The function returns a boolean representing whether the situation matches the spatial preposition. Other prepositions take different types of arguments: "in" takes a point and a polygon, and "between" takes a point and two polygons. Noun phrases such as "the kitchen" are resolved to polygons by human annotations. A parser extracts the function/argument structure from the query and resolves referring expressions in the query to an annotation. The parser can also extract parts of annotations, understanding expressions such as "the right side of the island." For classification, the path and polygon representation is converted to a set of features. A decision tree for each preposition is learned from labeled examples, making it easy to inspect the resulting classifier and determine which features are most important. The system uses the classifier to find video clips that match the query.

The system searches over a database of *person tracks*. Each person track is an ordered list of points and timestamps corresponding to a person's motion over several seconds of video. Person tracks are automatically extracted from the video 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 with motion templates [2], 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 motion in the scene.

For the evaluation, tracks were filtered in several ways. First, only tracks longer than four seconds in time and three feet in distance were included in the evaluation, in order to eliminate tracks that appear and disappear very quickly. Second, all tracks that any annotator labeled as "bad tracking" were excluded from the evaluation in order to focus the evaluation on the models for spatial prepositions.

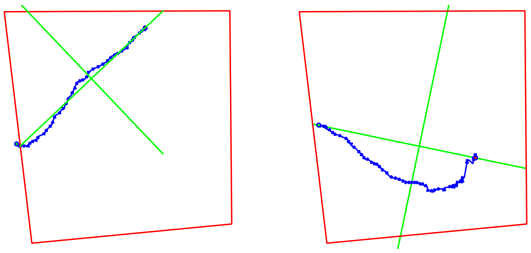


Figure 2: Schematic view of the clips shown in Figure 1. The axes that the figure imposes on the ground are overlayed in green.

3. MODELS FOR SPATIAL PREPOSITIONS

The system retrieves tracks matching a query by creating a computational model of the meaning of spatial prepositions in the query. The computational model is instantiated as a decision tree classifier that decides whether a track is a valid or invalid example of a spatial preposition based on a feature vector. The features are designed to computationally capture the meaning of spatial prepositions, and are described in the following section.

3.1 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 two argument spatial relation that 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: 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 intersects the ground. The axes for two scenes are illustrated in Figure 2. 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 normalized[1] average distance between the figure and the axes it imposes on the ground.

axesToFigureSum The normalized distance between the start of the axes and the start of the figure, plus the distance between the end of the axes and the end of the figure.

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

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

[1]Normalized distances are computed by dividing by the size of the diagonal of the bounding box of the figure and the ground together or in some cases the figure alone, in order to make the model scale invariant.

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

figureCenterOfMassToGroundCentroid The normalized distance between the center of mass of the figure and the centroid of the ground.

figureLengthByCrow The ratio of the length of the figure and the distance between its start and end points.

peakDistance The normalized maximum distance between the figure (over all points on the figure) and the axes (over all points on 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.

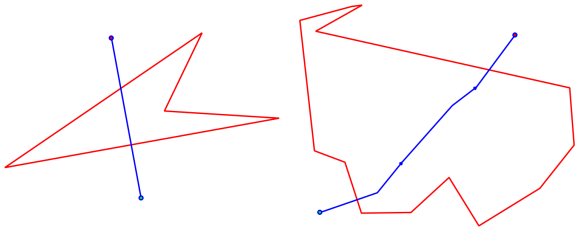
These features capture the degree to which the figure follows the axes, and the appropriateness of the axes to the ground. Our methodology was to invent features, and then let the system learn how to use those features for classification. By seeing what features were most important at this real-world task, we can gain insight into semantic structures underlying the meaning of spatial prepositions. Our evaluation shows that *ratioFigureToAxes* was most important for retrieval on its own. This feature is high when the figure goes from one point on the boundary of the ground to another, but does not capture the appropriateness of the two points: the two points might be very close together. Other features, such as *centroidToAxesOrigin* and *distAlongGroundBtwnAxes*, model this requirement.

To measure the generality of our model for "across," and qualitatively assess its performance, we asked an annotator to draw examples and counter-examples of "across," and measured the system's accuracy on this data set. For this task, the annotator used a mouse to draw a polygon representing the ground, and a series of line segments representing the figure, and labeled the example as "across" or "not across." We asked the annotator to only create clear examples about which other people would agree. Some examples that the system correctly classified from this data set are shown in Figure 3. The overall accuracy of the system (trained only on tracks annotated for "across the kitchen") was 0.68 and the F-score was 0.78. This performance shows that the model for across can handle diverse geometries, beyond what appears in our video data set.

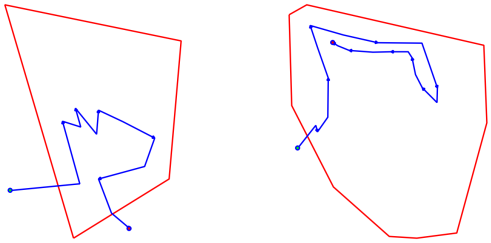
3.2 Along

"Along" is a two argument 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, visualized in Figure 4. In this way, the model reduces noise from the beginning or end of the path if the person started out far away from the ground object but quickly approaches it. The features used to train a decision tree learner to recognize examples of "along" are listed below.

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

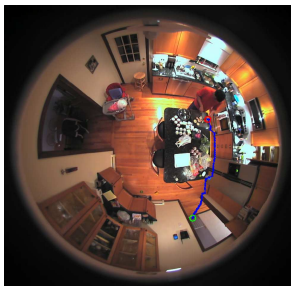


(a) Positive examples.

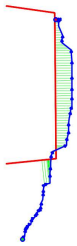


(b) Negative examples.

Figure 3: Examples of "across" from a data set created by one of our annotators. The annotator was free to draw any polygon for the ground and any line segment for the figure. The system correctly classified each of these examples.



(a) A frame of video.



(b) Schematic view of the clip, showing the distances used to compute the *averageDistance* feature in green.

```
peakDistance < 0.266 (value: 0.112)
  | distStartGround < 0.548 (value: 0.340)
  |   Class: True
  | distStartGround >= 0.548 (value: 0.340)
      ...
peakDistance >= 0.266 (value: 0.112)
  ...
```

(c) Part of the decision tree that classified this example.

Figure 4: A clip retrieved by the system for "along the right side of the island," showing the distances used to compute the *averageDistance* feature.

stdErrOfRegression The standard error of a regression line fit to the figure.

figureStartToEnd The normalized distance between the first and last points in the figure.

averageDistance The normalized 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.

standardDeviation The standard deviation of the distance between the figure and the ground.

peakDistance The normalized maximum distance between the figure and the ground.

These features capture the degree to which the figure follows the boundary of the ground, and the degree to which the figure and the ground are linear.

4. RELATED WORK

Our system transforms spatial language queries into a function/argument structure based on the theories of Jackendoff [8], Landau and Jackendoff [11] and Talmy [20]. Their definitions of "across" and "along" focus on the relationship of the figure to the axes of the ground. This work proposes a specific algorithm for computing the axes the figure imposes on the ground, and specifies features to precisely ground the meanings of these prepositions.

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. Our system uses some of the same features, such as center-of-mass distance between the figure and the ground, but uses decision trees, in order to give more insight into the operation of the model. By testing our model on annotations of real video, we are also using a more realistic test set. Regier and Carlson [15] describe the attention vector sum (AVS) algorithm, a precise computational model for the geometric meaning of the word "above" that captures many nuances of human judgements for this term. We expect the AVS algorithm to be an effective feature for a query such as "to the left of the island."

Also using video from the Human Speechome Project, Fleischman et al. [4] 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. [14] describe the Large-Scale Concept Ontology for Multimedia (LSCON), 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, LSCON 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, focusing on 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. [10] 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 [8] 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. [5] 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 [7] 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. [23] 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 a text string are easily repeatable; in contrast, it is difficult to draw (or tell someone else to draw) the exact same path 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 express graphically. The combination of a pen-based interface and a natural language interface is more powerful than either interface on its own.

5. EVALUATION

We evaluate our models for the meanings of "across" and "along" by measuring the performance of a system that uses the models to retrieve tracks that match a natural language query. Data for training and testing were labeled by annotators, who saw a video clip paired with a natural language phrase. The location of a person was marked at each frame of the clip, and annotators were instructed to mark whether the motion of the person in the clip matched the phrase. There was no capability to move the boundaries of video clips or join successive clips; such tracks were simply marked "invalid." In order to measure inter-annotator agreement, two annotators marked each clip.

Table 1 shows agreement scores for the data used in this evaluation. The results are well above chance, indicating that the task is achieving some level of consistency across annotators. However, it was surprisingly difficult to achieve good levels of inter-annotator agreement on this task. We had to iterate several times with annotators to develop a set of instructions. The instructions emphasized that only clear examples should be marked valid, and that there might

Query	Agreement
across the kitchen	0.65
across the living room	0.27
across the dining room	0.39
along the right side of the island	0.66
along the hallway	0.83
through the kitchen	0.68

Table 1: Chance corrected agreement scores for two annotators for various queries using Multi-π as described in Artstein and Poesio [1]. Zero is agreement at chance assuming all coders were drawing from the same underlying distribution, and one is perfect agreement.

not be very many "valid" examples. They also specified that the whole path should match the query: if only part of it matched, it should be marked "invalid" (e.g., a path where someone goes back and forth across the kitchen should be marked "invalid" for the query "across the kitchen"). The instructions included examples of tracks in each category. We chose this annotation methodology because it is fast to annotate, and it directly asks annotators to code the piece of information that we are interested in: whether the track should appear in a result set.

6. RESULTS

For the preposition "across," we report the system's performance for three queries: "across the kitchen," (Figure 5) as well as "across the living room" and "across the dining room" (Figure 6). In all cases the system was trained on one day of data annotated by one of the authors, using tracks only from the kitchen. Annotations for "across the living room" and "across the dining room" were not used during development of the system. We compare the system to a baseline heuristic that returns all long tracks in the kitchen. The length threshold of the baseline was chosen to maximize the F-score over the same training data seen by the system.

The two annotators did not have high agreement for "across the living room" and "across the dining room." Table 2 shows that most of the differences are examples that Coder C marked "valid," and Coder D marked "invalid." This suggests that Coder D was setting the threshold for "across" higher than Coder C.

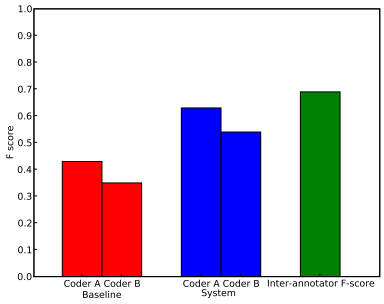


Figure 5: Results for "across the kitchen."

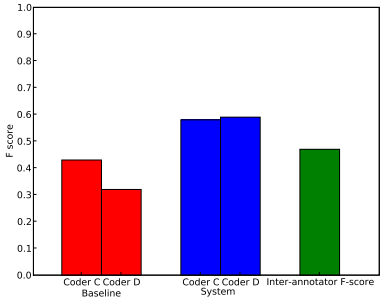


Figure 6: Results for "across the living room" and "across the dining room"

	TP	FP	FN	TN	F-score
Coder C/Baseline	256	242	424	1477	0.43
Coder D/Baseline	104	329	105	1578	0.32
Coder C/System	315	82	365	1637	0.58
Coder D/System	172	200	37	1707	0.59
Coder C/Coder D	197	12	440	1467	0.47

Table 2: The system's performance for "across the living room" and "across the dining room" as confusion matrices showing the number of true positives, false positives, true negatives and false negatives. F-scores for this data are graphed in Figure 6

Our framework is designed to model a variety of spatial prepositions. Here we report preliminary results for the prepositions "along" and "through." Figures 7 and 8 show the system's performance for two queries using the preposition "along." The baseline selects all tracks that intersect the region between the island and the counter. This baseline gives perfect recall, but lower precision and lower overall F-score than the system. Figure 9 shows the system's performance on the query "through the kitchen." The "through" classifier uses the same features as the "across" classifier, but is trained on different data. The baseline for "through" is the "across" classifier.

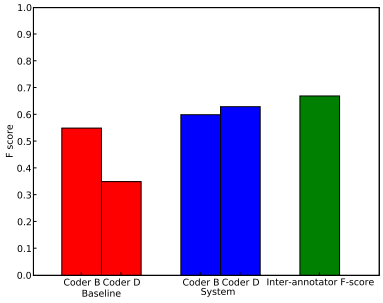


Figure 7: Results for "along the right side of the island."

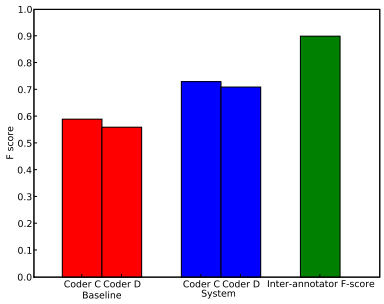


Figure 8: Results for "along the hallway."

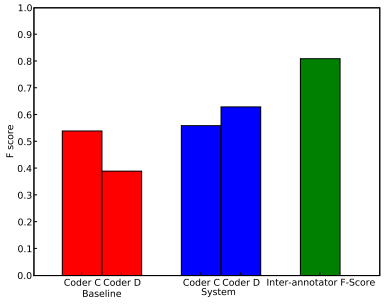


Figure 9: Results for "through the kitchen."

6.1 Discussion

Although our system performs better than a baseline and approaches human performance, it is troubling that there is not a larger spread between baseline performance and human judgements. Our relatively low agreement scores imply a problem with the formulation of the task. For the data graphed for "across the living room" and "across the dining room" in Figure 6, the average track length in pixels is 488. But the length of tracks where the two annotators agree is 449, compared to 635 over tracks where they disagree. These longer tracks may be harder to annotate because they are more likely to contain non-matching activity at the beginning and end of the track. Better event segmentation could solve this problem. Agreement scores may also be low because categorizing hundreds of video clips based on a spatial preposition is an unnatural task. Annotators may also incorporate judgements about a person's intentions and goals into their classification decision.

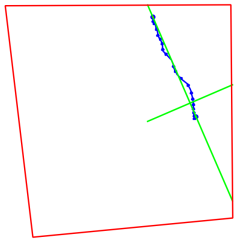
A key result of the evaluation is insight into what features are important to capturing the meaning of spatial prepositions. We present the performance of the classifier when trained on various subsets of features in Figures 12 and 13. In these figures, each horizontal bar represents the F-score of a classifier trained using only certain features. The colors in the bar encode what features were used to train a classifier for that run.

Figure 12(a) shows the performance of each feature on its own for the query "across the kitchen." This graph shows that the single best-performing feature for "across" is *ratioFigureToAxes*, a measure of the length of the figure compared to the length the axes it imposes on the ground. Figure 12(b) shows the performance of every pair of features. None of the features that measure the distance between the figure and the axes perform particularly well, except when paired with *ratioFigureToAxes*. Many of the other high scoring features measure how well the figure cuts across the ground, dividing it into two roughly equal parts. The importance of *ratioFigureToAxes* is also evident when looking directly at the decision tree used to classify examples of "across," shown in Figure 1(a). Here *ratioFigureToAxes* is the first feature used to separate the data. The other features used straightness of the path, and the second two measure the degree to which the path divides the ground into two equal pieces.

```
ratioFigureToAxes<0.786 (value: 0.874)
  ...
ratioFigureToAxes>=0.786 (value: 0.874)
  |  peakDistance<0.440 (value: 0.058)
  |  | axesToFigureSum<0.000 (value: 0.076)
  |  |  ...
  |  | axesToFigureSum>=0.000 (value: 0.076)
  |  |  | distAlongGroundBtwnAxes<0.252 (value: 0.337)
  |  |  |   ...
  |  |  | distAlongGroundBtwnAxes>=0.252 (value: 0.337)
  |  |  |  Class: True
  |  peakDistance>=0.440 (value: 0.058)
  |    ...
```

Figure 10: The decision tree used to classify the example shown in Figure 1 for "across the kitchen." The branches that the system took are shown in red.

(a) The last frame of video in the clip.



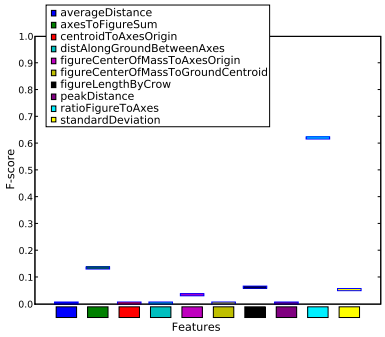
(b) Schematic view of the clip.

```
ratioFigureToAxes<0.786 (value: 0.508)
  | ratioFigureToAxes<0.461(value: 0.508)
  |   ...
  | ratioFigureToAxes>=0.461 (value: 0.508)
  | | figureLengthByCrow<0.796 (value: 0.935)
  | |   ...
  | | figureLengthByCrow>=0.796 (value: 0.935)
  | |   Class: True
ratioFigureToAxes >= 0.786 (value: 0.508)
  ...
```

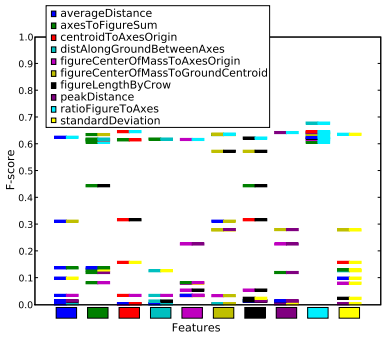
(c) Part of the decision tree that classified this example.

Figure 11: A clip that the system misclassified as "across the kitchen."

For "along," the best performing features measure the distance between the figure and the ground. *distEndGround* and *distStartGround* perform poorly on their own, but are present in all the highest performing pairs. These features involve the start and end points in the video clip, consis-

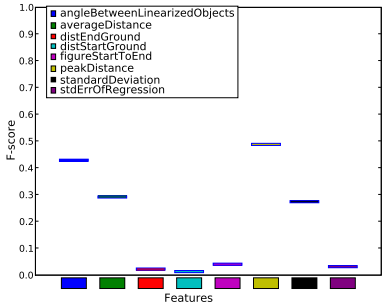


(a) F-scores for each feature alone.

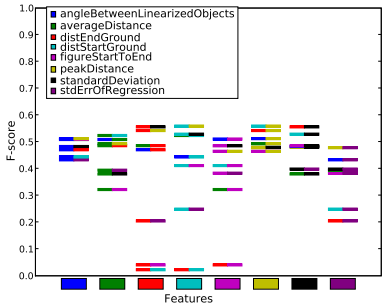


(b) F-scores for all pairs of features

Figure 12: Performance of subsets of features on "across the kitchen."



(a) F-scores for each feature alone.



(b) FR-scores for all pairs of features

Figure 13: Performance of subsets of features on "along the right side of the island."

tent with evidence that people pay more attention to the endpoints of a spatial motion event [16]. Finding a more principled algorithm for identifying boundaries in our video clips could further exploit this heuristic and lead to better performance. Some of the tracks returned for "along the right side of the island" were actually along the left side of the island. When one considers only the geometry of the right side of the island and the figure, as the system does, this makes sense, since the system only sees "the right side of the island" and not anything else in environment. To solve this problem we plan to introduce features involving the visibility of the ground from the figure, with respect to obstacles in the environment.

7. FUTURE WORK

Our next goal is to expand the vocabulary of our system so that it handles open-ended spatial queries. Using our library of geometric features we plan to add spatial prepositions whose meanings rely primarily on geometric information: "through", "around", "toward", and "away from." We have developed an annotation task in which annotators write a natural language description of the motion of a person in the surveillance video. Annotators will fill in the sentence "The person is going ..." with whatever ending makes sense to them given the person's motion. This methodology will enable us to collect a set of potential queries together with video clips that match them. We can then use these annotations to train and evaluate a retrieval system on a much larger vocabulary.

Once the system's lexicon is large enough, it may be possible to define higher-level events by composing lower-level primitives into compound concepts. For example, "chase" could be defined in terms of a sequence of "towards" events. In addition, compound events could be described from simpler events: "putting away the dishes" is when someone goes "from the drying rack to the cupboards over and over." This idea has the potential to enable rapid expansion of the system's vocabulary, but in a context where the quality of definitions of new terms can be robustly tested and evaluated.

8. CONTRIBUTIONS

We presented the first steps towards a system that can find video clips that match a natural language description using models for the meanings of spatial prepositions. Our system models the meaning of the words "across" and "along" with a set of computationally grounded features, and we have identified specific features that enable a classifier to find examples of these prepositions. The library of features can be used to ground the meanings of other spatial prepositions as well. This framework is enabling us to build a retrieval system that starts to bridge the semantic gap, by finding clips that open-ended spatial-language queries.

9. ACKNOWLEDGEMENTS

Thanks to Piotr Mitros, Rony Kubat, Jeff Orkin, Brandon Roy, and Gregory Marton for their comments on earlier drafts of this paper. We would also like to thank Angela Brewster and the other annotators.

References

[1] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, Dec. 2008.

[2] G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, Inc., 1st edition, Oct. 2008.

[3] 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.

[4] 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.

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

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

[7] 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.

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

[9] 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.

[10] 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.

[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] 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.

[15] T. Regier and L. A. Carlson. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology. General*, 130(2):273–98, June 2001. PMID: 11409104.

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

[22] 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.

[23] 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.

[24] 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.