In Proceedings of 2003 AAAI Spring Symposium on New Directions
in Question Answering, March 2003, Palo Alto, California

Answering Questions about Moving Objects in Surveillance Videos

Boris Katz, Jimmy Lin, Chris Stauffer, and Eric Grimson

MIT Artificial Intelligence Laboratory
200 Technology Square
Cambridge, MA 02139
{boris,jimmylin,stauffer,welg}@ai.mit.edu

Abstract

Current question answering systems succeed in many respects regarding questions about textual documents. However, information exists in other media, which provides both opportunities and challenges for question answering. We present results in extending question answering capabilities to video footage captured in a surveillance setting. Our prototype system, called Spot, can answer questions about moving objects that appear within the video. We situate this novel application of vision and language technology within a larger framework designed to integrate language and vision systems under a common representation. We believe that our framework will support the next generation of multimodal natural language information access systems.

Introduction

Although many advances have been made in question answering over the last few years, most existing systems are exclusively text-based (Voorhees, 2001; Voorhees, 2002). While such systems are undoubtedly useful, information exists in many other types of media as well; a truly effective information access system should not only be able to answer questions about text, but also about pictures, movies, sounds, etc. Furthermore, text is often not the most appropriate answer to user queries. An intelligent information access system should be able to choose the answer format that will best satisfy users' information needs.

We are extending question answering capabilities into new domains. In particular, our focus has been on video footage captured in a surveillance setting. We have developed Spot, a system that answers questions about moving objects found within video footage. In response to user questions, our system returns dynamically generated video clips that directly satisfy the user query. For example, when the user asks "Show me all southbound cars," Spot displays a video that consisted solely of cars heading south; all other traffic, both pedestrian and vehicular, is discarded. This technology is made possible by integrating a motion-tracking system and a natural language system, both developed at the MIT AI Laboratory.

In addition to our prototype video-surveillance question answering system, we are currently developing a generalized framework for integrating vision and language systems in order to exploit the synergies that arise from fusing multimodal information streams. To this end, we are developing a Common Linguistic-Visual Representation (CLiViR) that captures the salient aspects of both language and vision. This generalized framework supports four major capabilities: event recognition, event querying using natural language, natural language event summarization, and event monitoring.

Although there exist information retrieval systems that operate on video clips and still images (Aslandogan and Yu, 1999; Smeaton *et al.*, 2001), the vast majority of them treat multimedia segments as opaque objects. For the most part, current multimedia information retrieval systems utilize textual data, such as captions and transcribed speech, as descriptors of content for indexing purposes. For many types of media, such textual metadata is hard to obtain. Furthermore, the content of multimedia segments cannot be adequately captured by representations purely derived from text; such representations will necessarily be impoverished. Although there has been research on automatically extracting features from video and images, it has been limited to such information as color, shape, and texture; such low-level features alone are insufficient to capture the semantic content of non-textual segments. In addition, automatically translating user queries into sets of such low-level features is a challenge yet to be overcome.

We believe that in certain domains it is possible to directly analyze video input and generate representations that capture the semantics of the events by bringing to bear computer vision technology. We attempt to break the inter-media barrier by developing shared representations that are capable of bridging different modalities.

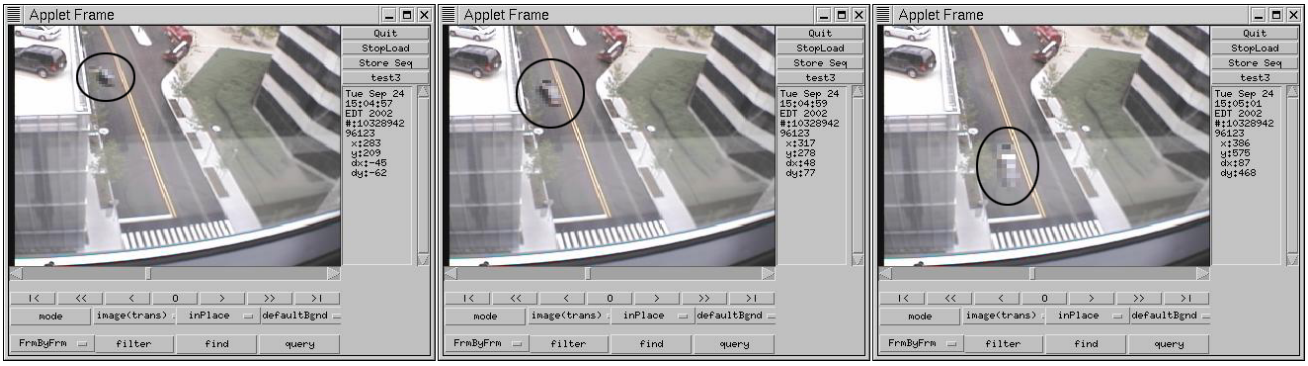

Figure 1: Still frames taken from Spot's answer to "Show me all cars leaving the garage."

Spot: A Prototype

We have built a prototype information access system, called Spot, that answers interesting questions about video surveillance footage taken around the Technology Square area in Cambridge, Massachusetts. The scene consists of a large parking garage to the west, an office building to the east, and a north-south roadway that runs between the two structures. The area experiences moderate to heavy amounts of both pedestrian and vehicular traffic, depending on the time of day. A typical segment of the video footage shows cars leaving and entering the parking garage, vehicles (e.g., cars and delivery trucks) driving both northbound and southbound, and pedestrians walking around.

In response to a natural language question, our Spot system is able to filter raw footage and dynamically assemble an abridged video clip satisfying the user request. Currently, we focus on various types of motion within the scene. For example, when a user asks "Show me all cars leaving the garage," the system responds with a video clip showing only cars exiting the garage; all other vehicular and pedestrian traffic is discarded. Figure 1 shows several still frames from the answer.

Currently, the system can answer a variety of interesting questions, e.g.,

> Did any cars leave the garage towards the north?
> Display cars exiting the garage towards the south.
> Show me cars entering Technology Square.
> Give me all southbound cars.

Our Spot system is a proof of concept demonstrating the viability of question answering for video surveillance. Much in the same way that traditional question answering systems can respond to queries about textual documents, Spot allows users to ask interesting questions about objects moving in a particular scene.

Underlying technology

Spot is the product of computer vision and natural language understanding technology. Our prototype is supported by two systems developed at the MIT AI Laboratory: a real-time motion tracking system and the START Natural Language System.

Robust Object Tracking

By combining the latest in both computer vision and machine learning techniques, we have developed systems that can robustly track multiple moving objects in both indoor and outdoor settings. Under a bottom-up, data-driven framework called Perceptual Data Mining (PDM) (Stauffer, 2002), we have created autonomous perceptual systems that can be introduced into almost any environment and, through experience, learn to model the active objects of that environment (Stauffer and Grimson, 2000). Over the last five years, we have processed billions of images. Using novel attention mechanisms and adaptive background estimation techniques, our system can isolate moving objects from stationary background scenery.

Our motion-tracking algorithm is based on an adaptive background subtraction method that models each pixel as a mixture of Gaussians and uses an on-line approximation to update the model. The Gaussians are then evaluated using a simple heuristic to decide whether or not a pixel is part of the background process. Foreground pixels are segmented into regions by a two-pass, connected components algorithm. Objects are tracked across frames by using a linearly predictive multiple hypotheses tracking algorithm, which incorporates both position and size. Our approach is able to robustly ignore environmental effects, e.g., flags fluttering or trees swaying in the wind, etc., and handle different weather conditions, e.g., rain or snow. Furthermore, our system is capable of maintaining tracks through cluttered areas, dealing with objects overlapping in the visual field, and adjusting to gradual lighting changes.

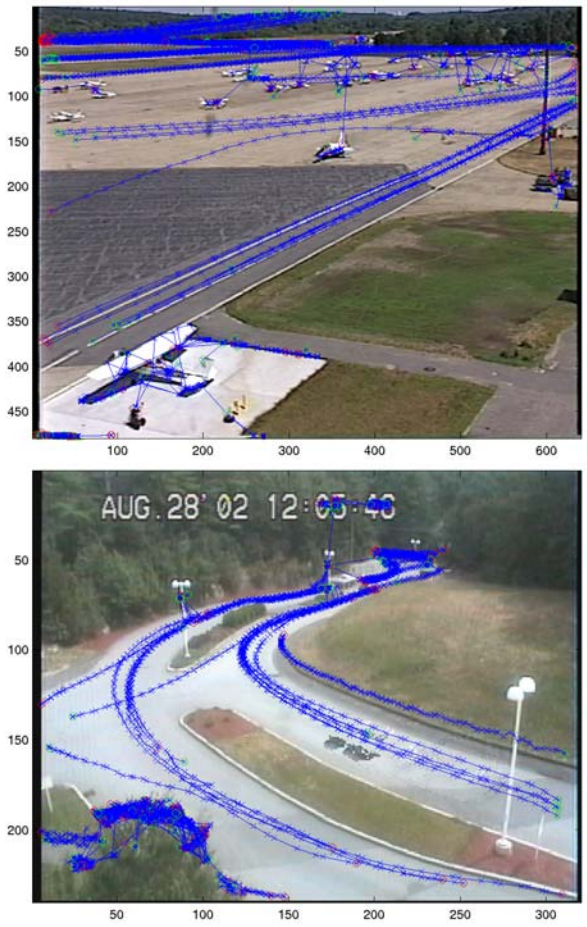

Figure 2: A composite of motion tracks detected by our vision system in two different settings: an airport tarmac (above), and an entrance gate to an office park (below).

With our technology, it is possible to observe and characterize motions in a particular scene over long periods of time. By applying unsupervised classification techniques to the observed trajectories of moving objects, we can categorize patterns of usage in a site; these include common paths of movement through the site based on type of object, as well as common patterns of usage as a function of time of day. As an example, Figure 2 shows two scenes, a tarmac setting at an airport and a gate of an office complex, with motion tracks superimposed. From the airport environment, we are able to observe tracks of airplanes taking off and landing in the distance, typical taxi paths, and motion of cars along the roads. From the office gate setting, we are able to observe cars entering and leaving, as well as pedestrians walking along the road. This classification provides us with a basis for flagging unusual behaviors, for retrieving similar instances of behaviors, and for gathering statistics on site usage.

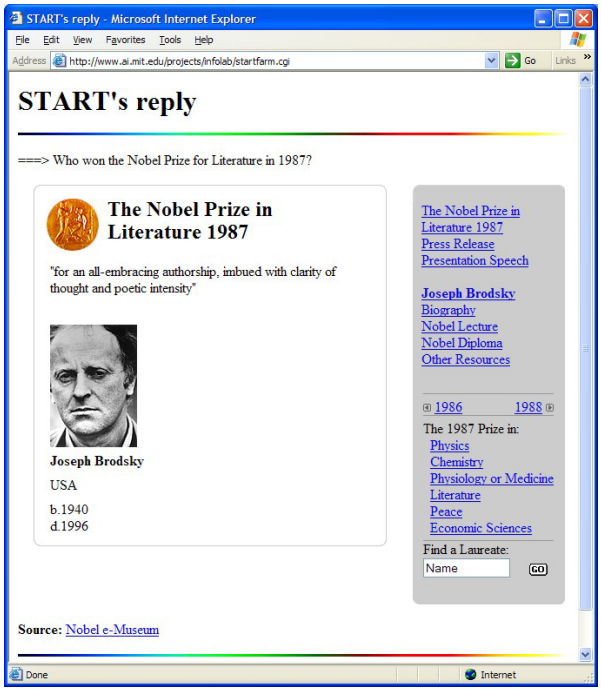

Figure 3: START answering the question "Who won the Nobel Prize for Literature in 1987?"

Natural Language Understanding

The other component that supports the Spot system is natural language understanding technology, in the form of the START information access system (Katz, 1997; Katz *et al.,* 2002). START is grounded in a technique called natural language annotation, in which English phrases and sentences are used to describe information segments and the types of questions that they are capable of answering. The system then parses these annotations and stores the parsed structures (called ternary expressions) with pointers back to the original information segments they describe. To answer a question, the user query is compared to the annotations stored in the knowledge base. Because this match occurs at the level of syntactic structures, linguistically sophisticated machinery such as synonymy/hyponymy, ontologies, and structural transformational rules are all brought to bear on the matching process. Linguistic techniques allow the system to achieve capabilities beyond simple keyword matching, for example, handling complex syntactic alternations involving verb arguments (Katz and Levin, 1988). If a match is found between ternary expressions derived from annotations and those derived from the query, the segment corresponding to the annotations is returned to the user as the answer. Figure 3 shows a screenshot of START answering a user question with both text and images.

An important feature of the annotation concept is that any information segment can be annotated: not only text, but also images, multimedia, and even procedures. For example, pictures of famous people or flags of countries in the world could be annotated with appropriate phrases and retrieved in response to user queries. Multimedia items such as recordings of "hello" in various languages could be treated in the same manner. A procedure for calculating distances between two locations or a procedure for calculating the current time in any world city could also be similarly annotated, as well as database queries, which give START access to large quantities of both structured and semistructured information (Katz *et al.*, 2002). Annotation of structured queries allows START to treat the Web as if it were a uniform "virtual" database; organized in this fashion, Web resources serve as valuable knowledge sources for question answering.

START, the first question answering system available on the World Wide Web, came on-line in December, 1993. Since then, it has engaged in exchanges with hundreds of thousands of users all over the world, supplying them with useful knowledge. Currently, our system answers millions of natural language questions about places (e.g., cities, countries, lakes, coordinates, weather, maps, demographics, political and economic systems), movies (e.g., titles, actors, directors), people (e.g., birth dates, biographies), dictionary definitions, and much, much more.

Integrating Vision and Language

In response to a user query in English, Spot dynamically applies *filters* over raw video footage to generate new video clips containing only information that satisfies the users' information need. START assists in the process by understanding the user query and translating the information need into the correct filter. This section explains how the translation process works in greater detail.

The basic unit of output from the motion tracking system is a *track*, which traces the motion of a single object over time. A track is comprised of a sequence of *track instances*. A track instance depicts a tracked object at a particular moment in time. Each track instance is tagged with a unique identifier and timestamp, and also contains information about the object's screen coordinates, size, velocity vector, and other meta information. In addition, each track instance contains the actual image of the object and its silhouette, making it possible to reconstruct a movie of the object and its motion.

We have developed a symbolic representation that abstracts away from the raw tracking stream. Our representation captures only the salient aspects of the motion observed, and allows for a more compact description of the visual data. A compact, yet expressive representation, is crucial because the tracking information provided by the vision system is continuous and extremely large—a single camera is capable of generating over one hundred thousand images an hour, most of which may not be relevant to the user.

The basis of our representation is adapted from Jackendoff's representation of motion and paths (Jackendoff, 1983; Jackendoff, 1990). This representation is a component of Lexical Conceptual Structures (LCS), a syntactically-grounded semantic representation that captures many cross-linguistic generalizations. It has been successfully used in applications such as interlingual machine translation (Dorr, 1992) and intelligent tutoring (Dorr, 1995).

As an example, a car leaving the garage would be represented in the following (simplified) expression:

$$\text{MOVE}(\text{Object}_{213}, [_{\text{PATH}}\ \text{Source}(\text{Garage}_{57})])$$

In this representation, objects moving along paths are specified by a series of *path primitives*. Path primitives capture the relationship between the object in motion and other (mostly stationary) objects; they often correspond to prepositions in natural language. Path primitives are ideal for serving as the intermediary between vision and language: They correspond to features that can be easily extracted from raw video (using only screen coordinates and other tracking information). In addition, natural language queries about motion are most naturally specified in terms of prepositions, which correspond directly to path primitives. Thus, our framework for capturing motion and paths serves not only as an expressive representation language, but also functions as a powerful query language.

Paths in our representation can be minimally specified, i.e., with a single path primitive, or they can describe complicated paths in detail. The well-known verse of a popular winter song:

> Over the river and through the woods, to grandmother's house we go...

could be expressed as

$$\begin{aligned}&\text{MOVE}(\text{We},\\ &\quad [_{\text{PATH}}\ \ \text{Over}(\text{River}_{35})\\ &\qquad \text{Through}(\text{Woods}_{23})\\ &\qquad \text{Destination}(\text{GrandmothersHouse}_{1})])\end{aligned}$$

This representation would allow us to answer a variety of questions, e.g.,

> Where did we go?
> What did we pass to get to grandmother's house?
> Did we go through anything on our way?

Spot utilizes this representation as a query language to filter surveillance footage. Consider the query "Show me all cars leaving the garage." Using natural language annotations, START translates the English question into the directive:

Filter([$_{PATH}$ Source($Garage_{57}$)])

We assume that prior to answering the question, the system has already been taught the location of the garage, in terms of screen coordinates. This query can then be fulfilled by a simple script that checks if the beginning of each track lies within the indicated region (to within an error tolerance). The result is a dynamically generated video that contains only motion tracks satisfying the filtering profile. Because START understands the user question, it can easily handle variations in language, e.g.,

Did any cars leave the garage?
Give me all cars that exited the garage.
Display cars leaving the garage.

As another example, our natural language annotation technology allows Spot to translate the question "Show me cars entering Technology Square" into the query

Filter([$_{PATH}$ Source($Road_{North}$) Direction(South)])

Accordingly, Spot displays a video showing only cars entering Technology Square.

Related Work

Object and event recognition in the general domain is far beyond the capabilities of current technology. Instead, current video and image retrieval systems rely on low-level features such as color, texture, and shapes that can be automatically extracted (see (Aslandogan and Yu, 1999; Yoshitaka and Ichikawa, 1999) for a survey). However, such systems are fundamentally incapable of capturing high-level semantics.

A method of overcoming the limitations of low-level feature-indexing is to utilize textual annotations that may accompany multimedia content. Image retrieval systems have been built around the use of image captions (Smeaton and Qigley, 1996; Wactlar *et al.*, 2000); similarly, video retrieval systems have incorporated textual transcripts (either taken from closed-captions or generated by speech recognition systems) and other manually entered annotations (Smeaton *et al.*, 2001). There are several drawbacks to this approach: In many cases, descriptive annotations cannot be obtained automatically and require human labor to gather. In addition, unstructured text may not be the best representation for multimedia content, and systems performing multimedia retrieval on textual annotations must contend with well known problems in natural language processing, e.g., ambiguity, alternations, etc.

Other attempts to automatically extract higher-level semantics from multimedia segments include pseudo-semantic classification, where items are broken down into broad, generic categories like nature *vs.* man-made or indoor *vs.* outdoor (Smith *et al.*, 2001; Chen *et al.*, 1999). Object recognition technology has also been applied to image and video retrieval, although most efforts have been focused on specific objects, e.g., faces, numbers, etc. Video Semantic Directed Graph (Day *et al.*, 1995), an object oriented framework for representing video sequences, focuses on modeling physical objects and their appearance or disappearance rather than events and activities. In contrast to these approaches, we are attempting to develop automatic methods of extracting and modeling high-level semantic events.

Next Steps

Although our current path representation is limited in scope and in the types of questions that it can answer, we believe that Jackendoff's LCS representation serves as a solid basic foundation upon which to build more expressive structures. Our immediate goal is to augment our existing representation with attributes such as time, speed, color, size, etc. These additions would allow a larger variety of questions to be answered:

Show me the last delivery truck that stopped in front of the office.
Show me all pedestrians walking north.
Display all blue cars entering the garage.

Furthermore, we are in the process of developing a language that can manipulate these primitive building blocks to craft more complex events. A significant portion of this effort is to explore the possible ways in which primitive events can be related to each other; through this process we hope to develop a meaningful set of "connectors" to explicitly express such relations. A primitive event language grounded in human intuitions would allow non-experts to group a particular sequence of events into a larger unit. It would allow users to ask very interesting questions like:

In the last hour, did any car pull up to the curb and let out passengers?
Have any trucks circled the building more than twice within the last day?
Show me any instance of a man getting into a car and then getting out of the car within five minutes.

With traditional video-retrieval systems, it is often very difficult to construct meaningful queries in terms of low-level features such as textures and colors, e.g., a car dropping a passenger off at the curb or a van making a u-turn. Yet, many of these complex events can be easily specified in natural language. We believe that by building a suitable abstraction between language and vision capable of capturing high-level semantic events, we can build systems that afford users effective access to video information

Furthermore, traditional information-retrieval-based question answering technology can be integrated with a system like Spot to form a generalized multimodal information access system. Such an integration would provide users with powerful tools for analyzing situations from different perspectives. For example, consider a busy freeway intersection: integration of video and textual data would allow a user to understand anomalous traffic patterns (detected by video surveillance) by consulting traffic and weather reports (textual data).

Input using multiple modalities is another direction that we would like to explore. In our domain of visual surveillance, gestures could play an important role as a possible mode of querying. For example, a user could ask "Show me all cars that went like this [gesturing an indirect path the leads from the garage to intersection." Or "Did anyone leave this building [pointing to a specific office building] in the last hour?"

A Common Framework

As part of our attempts at integrating vision and language, we are developing a common representational framework called CLiViR (Common Linguistic-Visual Representation). The goal is to develop practically-grounded shared structures that bridge the visual and linguistic domains. Our desire is to capture the salient aspects of both visual and linguistic data, while discarding irrelevant details. Our framework supports four major capabilities that cross the boundaries of language and vision (see Figure 4):

- **Event Recognition.** CLiViR serves as an "event language" for describing visual scenes. Our representation abstracts away from the raw video feeds into a symbolic structure that can be further analyzed and manipulated.

- **Event Querying.** English queries can be translated into queries in CLiViR, and then matched against the recognized events. With such capabilities, users can ask questions in English and get back appropriate answers, either video clips or textual descriptions. Because the matching is done on symbolic representations, the system can provide concise responses that capture large variations in the video data.

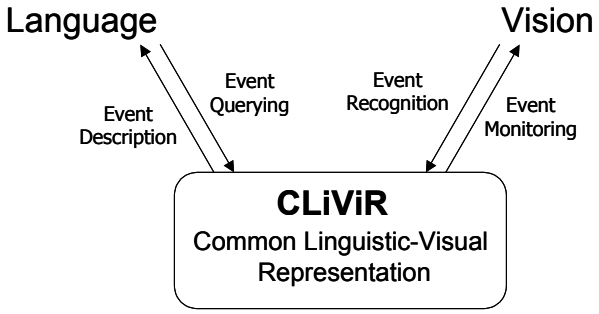

Figure 4: The basic structure of an integrated language and vision system.

- **Event Description.** Our common representation also supports natural language generation, so that users can request "digested summaries" of a particular scene, e.g., "In the last five minutes, five cars passed by the front. A blue van stopped across the street for approximately a minute and then drove off."

- **Event Monitoring.** CLiViR also allows users to issue standing queries that filter the incoming video, e.g., "Notify me whenever a black sedan pulls up into the driveway."

In our development of CLiViR, we are drawing from a large body of research in artificial intelligence, knowledge representation, and other cognitively-inspired computational theories. By leveraging previous work in related areas, we can synthesize elaborate and expressive theories of meaning. On the other hand, our commitment to work with unaltered, real-world video will ground our system in realistic scenarios, ensuring robustness and scalability. In essence, we are developing a platform capable of validating theories of representation against real-world data.

An important aspect of video sequences that cannot be easily captured by Jackendoff's LCS representation is the notion of time and temporal intervals. For modeling temporal aspects of activity, we believe that Allen's work on qualitative relations between temporal intervals (Allen, 1983) is highly relevant. Correspondingly, there has been some work on relations in the spatial domain that we could capitalize on (Chang *et al.*, 1987; Egenhofer and Franzosa, 1991). In addition, we believe that representations focusing on change and transitions between states (Borchardt, 1992), rather than states themselves, will also be helpful in the development of CLiViR.

An especially important aspect of the CLiViR representation is that it can capture activity at multiple levels of abstraction and utilize multiple parallel representations: an impossible problem at one level of abstraction becomes trivial at another; a difficult problem can be transformed

into a simple problem by switching representations. However, the challenge will be relating these different structures and knowing when to apply them.

Conclusion

The integration of vision and language systems presents both difficult challenges and exciting opportunities for information access systems. We believe that we have taken an important step in the right direction. Not only have we demonstrated question answering on video surveillance footage, but also sketched a general framework for integrating the visual and linguistic domains under a shared representation.

Psychologists have long believed that language and vision are two very important aspects of cognition that contribute to what we ascribe as "intelligence." Humans are able to effortlessly integrate visual and linguistic data to reason and learn, an ability far beyond that of present day computers. By structuring a framework that bridges vision and language, not only can we build more effective information access systems, but perhaps we can also shed some light on the wonders of human cognition.

Acknowledgements

We would like to thank Gary Borchardt, Greg Marton, and Stefanie Tellex for their comments on earlier drafts of this paper.

References

Allen, J. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*. 26(11): 832-843.

Aslandogan, Y. and Yu, C. 1999. Techniques and systems for image and video retrieval. *IEEE Transactions on Knowledge and Data Engineering*. 11(1):56-63.

Borchardt, G.C. 1992. Understanding causal descriptions of physical systems," *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI-92).*

Chang, S.K.; Shi, Q.; and Yan, C. 1987. Iconic indexing by 2-D string. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 9(3):413-428.

Chen, J.Y.; Taskiran, C.; Albiol, A.; Delp, E.J.; and Bouman, C.A. 1999. ViBE: a video indexing and browsing environment. *Proceedings of the SPIE Conference on Multimedia Storage and Archiving Systems IV*.

Day, Y.F.; Dagtas, S.; Iino, M.; Khokhar, A.; and Ghafoor, A. 1995. Object-oriented conceptual modeling of video data. *Proceedings of the Eleventh International Conference on Data Engineering (ICDE-95).*

Dorr, B. 1992. The Use of lexical semantics in interlingual machine translation. *Machine Translation*, 7(3):135-193.

Dorr, B.; Hendler, J.; Blanksteen S.; and Migdalof, B. 1995. Use of LCS and discourse for intelligent tutoring: On beyond syntax. In M. Holland, J. Kaplan, and M. Sams (eds.), *Intelligent Language Tutors: Balancing Theory and Technology*. Hillsdale, NJ: Lawrence Erlbaum Associates. pp. 288-309.

Egenhofer M. and Franzosa R. 1991. Point-set topological spatial relations. *International Journal of Geographic Information Systems.* 5(2):161-174.

Jackendoff, R. 1983. *Semantics and Cognition.* Cambridge, Massachusetts: MIT Press.

Jackendoff, R. 1990. *Semantic Sturctures*. Cambridge, Massachusetts: MIT Press.

Katz, B. and Levin, B. 1988. Exploiting lexical regularities in designing natural language systems. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING'88).*

Katz, B. 1997. Annotating the World Wide Web using natural language. In *Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97)*.

Katz, B.; Felshin, S.; Yuret, D.; Ibrahim, A.; Lin, J.; Marton, G.; McFarland, A.J.; and Temelkuran, B. 2002. Omnibase: Uniform access to heterogeneous data for question answering. *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB 2002)*.

Smeaton, A. and Qigley, I. 1996. Experiments on using semantic distances between words in image caption retrieval. *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-96).*

Smeaton, A.; Over, P.; and Taban, R. The TREC-2001 video track report. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*.

Smith J.; Srinivasan S.; Amir A.; Basu S.; Iyengar, G.; Lin, C.Y.; Naphade, M.; Ponceleon, D.; and Tseng, B. 2001. Integrating features, models, and semantics for TREC Video Retrieval. *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*.

Stauffer, C., and Grimson, W.E.L. 2000. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8):747–757.

Stauffer, C. 2002. *Perceptual Data Mining*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Voorhees, E.M. 2001. Overview of the TREC 2001 question answering track. *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*.

Voorhees, E.M. 2002. Overview of the TREC 2002 question answering track. *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*.

Wactlar, H.; Hauptmann, A.; Christel, M.; Houghton R.; Olligschlaeger. 2000. Complementary video and audio analysis for broadcast news archives. *Communications of the ACM*. 32(2):42-47.

Yoshitaka, A. and Ichikawa, T. 1999. A survey on content-based retrieval for multimedia databases. *IEEE Transactions on Knowledge and Data Engineering.* 11(1):81-93.