Chapter 12

Answering Questions about Moving Objects in Videos

Boris Katz, Jimmy Lin, Chris Stauffer, and Eric Grimson
MIT Computer Science and Artificial Intelligence Laboratory
200 Technology Square
Cambridge, MA 02139
{boris,jimmylin,stauffer,welg}@csail.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 describe our efforts in extending question answering capabilities to video data: our implemented prototype, Spot, can answer questions about moving objects in a surveillance setting. This novel application of vision and language technology is situated within a larger framework designed to integrate knowledge from multiple domains under a common representation. We believe that our framework will support the next generation of multimodal natural language information access systems.

1. 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, since text is often not the most appropriate response to user requests, an intelligent information access system should be able to choose the answer format that will best satisfy users' information needs.

Our research centers on extending question answering capabilities to new domains: specifically, video footage captured in a surveillance setting. We have developed Spot, a system that answers questions about moving objects found in

The final version of this book chapter appears as:
Boris Katz, Jimmy Lin, Chris Stauffer and Eric Grimson. Answering Questions About Moving Objects in Videos. In Mark T. Maybury, editor, New Directions in Question Answering. Cambridge, Massachusetts: MIT Press, 2004, pages 113-128.

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 consists solely of cars heading south; all other traffic, both pedestrian and vehicular, is omitted. This research effort is made possible by integrating motion-tracking and natural language technologies developed at the MIT Computer Science and Artificial Intelligence Laboratory.

Our prototype video-surveillance question answering system is situated in a general framework for integrating vision and language systems, called CLiViR (Common Linguistic-Visual Representation). In order to exploit the synergies that arise from fusing multimodal information streams, we are developing common representations capable of capturing salient aspects of both language and vision. This general framework supports four major capabilities: event recognition, question answering, event description, and event filtering.

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, the required textual metadata is hard to obtain. Furthermore, the content of multimedia segments cannot be adequately captured by representations purely derived from text; these 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; these low-level features alone are insufficient to capture the semantic content of non-textual segments. In addition, automatically translating user queries into sets of appropriate low-level features is a challenge yet to be overcome.

In certain domains it is possible to integrate computer vision and natural language technology in order to extract high-level events from video input and generate symbolic representations that capture the semantics of those events. We attempt to break the inter-media barrier by developing shared representations that are capable of bridging different modalities.

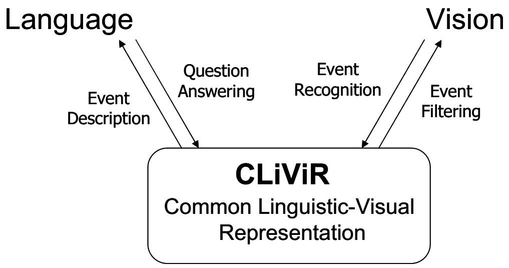


Figure 1: The capabilities provided by CLiViR, our integrated vision and language framework

1.1. A Common Framework

Common Linguistic-Visual Representation (CLiViR) lies at the heart of our effort to integrate vision and language. The central goal is to develop a framework that supports practically-grounded shared representations capable of bridging the visual and linguistic domains. Our desire is to capture the salient aspects of both visual and linguistic data relevant for answering user questions, while discarding irrelevant details. Our framework supports four major capabilities that cross the boundaries of language and vision (see Figure 1):

- **Event Recognition.** CLiViR serves as an "event language" for describing visual scenes. Our representation abstracts away raw video into symbolic structures that can be easily analyzed and manipulated.

- **Question Answering.** Natural language questions can be translated into queries in CLiViR, and then matched against 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.

- **Event Description.** Our common representation 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 in front of the building. A blue van stopped across the street for approximately a minute and then drove off."

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

Our development of CLiViR is driven by both theoretical and practical considerations. On one hand, we believe that by leveraging existing work in knowledge representation and cognitive psychology, it is possible to synthesize theories of meaning expressive enough to capture linguistic and visual knowledge. On the other hand, our commitment to working 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.

1.1 Spot: A Prototype

We have built a prototype information access system, called Spot, that can answer 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, a white 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 video segment shows cars leaving and entering the parking garage, vehicles (e.g., cars and delivery trucks) driving both northbound and southbound, and pedestrians walking around.

Our Spot system analyzes raw footage and attempts to recognize salient events in the video; these events, captured in a symbolic representation, are stored in a knowledge base and indexed for convenient retrieval. When asked a natural language question, Spot first transforms it into a symbolic query, which is then matched against the knowledge base of indexed events. Based on the match, Spot dynamically assembles 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 cars pulling up to the white office building.", the system responds with a video clip showing cars entering the driveway of that

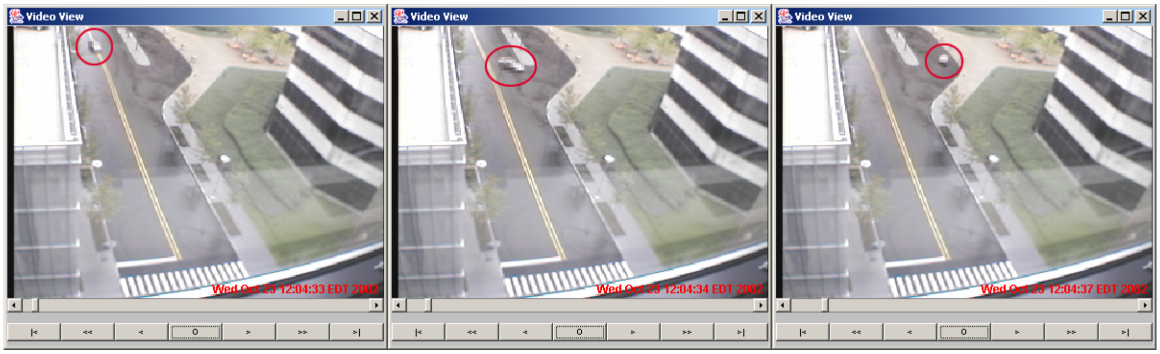

Video View
Wed Oct 23 12:04:33 EDT 2002
|< << < 0 > >> >|
Video View
Wed Oct 23 12:04:34 EDT 2002
|< << < 0 > >> >|
Video View
Wed Oct 23 12:04:37 EDT 2002
|< << < 0 > >> >|

Figure 2: Still frames taken from Spot’s answer to “Show me cars pulling up to the white office building.”

particular building; all other vehicular and pedestrian traffic is omitted. Figure 2 shows several still frames from the answer.

Our prototype can presently answer a variety of interesting questions, e.g.,

> Show me cars dropping off people in front of the white building.
> Did any cars leave the garage towards the north?
> How many cars pulled up in front of the office building?
> Show me cars entering Technology Square.
> Give me all northbound traffic.

Spot is a proof of concept demonstrating the viability of question answering for video surveillance applications. 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.

2. Underlying technology

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

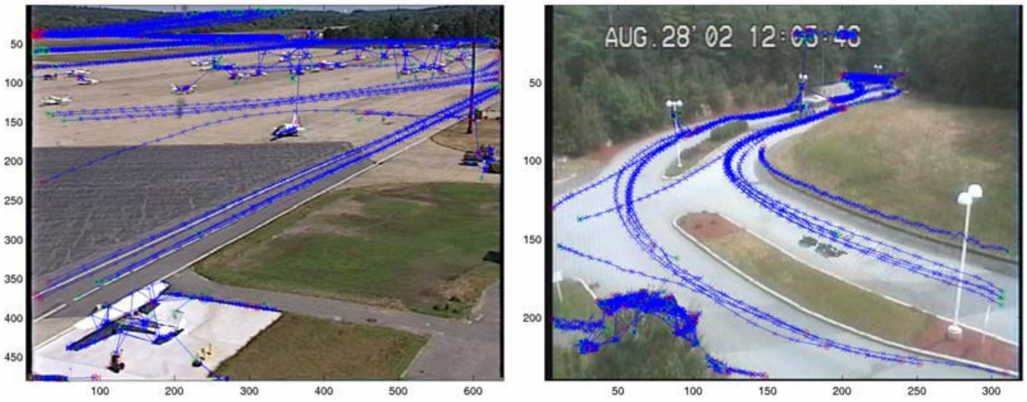


Figure 3: A composite of motion tracks detected by our system in two different settings: an airport tarmac (left), and an entrance gate to an office park (right).

2.1 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.

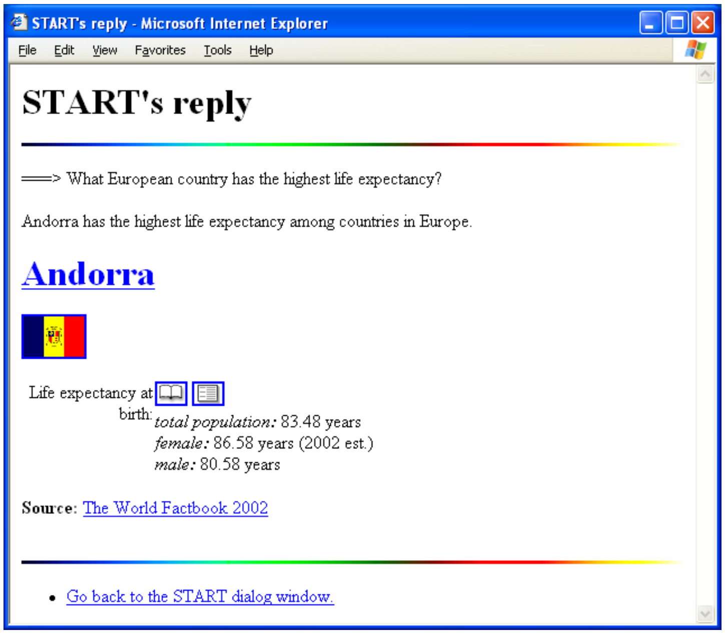
Our motion-tracking algorithm is based on an adaptive background subtraction method that models each pixel as a mixture of Gaussians and uses an on-line approximation to update the model. The Gaussians are then evaluated using a simple heuristic to decide whether or not a pixel is part of the background 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 overlapping objects, and adjusting to gradual lighting changes.

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 3 shows two scenes, a tarmac at an airport and a gate of an office complex, with motion tracks superimposed. In the airport environment, our system is able to detect airplanes taking off and landing in the distance, typical taxi paths, and motion of cars along the roads. In the office gate setting, our system is 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.

2.2 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). In December 1993, START became the first question answering system available for question answering on the World Wide Web. Since then, it has engaged in exchanges with hundreds of thousands of users all over the world, supplying them with useful knowledge. See Chapter 17 in this volume for a more detailed description of our system.

The START system 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 parses these annotations, converts the parsed structures into a set of ternary expressions (Katz, 1988), and stores them in a knowledge base with pointers back to the original information segments they describe. To answer a question, the user query is also converted into a set of ternary



START's reply

==> What European country has the highest life expectancy?

Andorra has the highest life expectancy among countries in Europe.

Andorra

Life expectancy at birth: *total population:* 83.48 years
female: 86.58 years (2002 est.)
male: 80.58 years

Source: The World Factbook 2002

- Go back to the START dialog window.

Figure 4: START answering the question "What European country has the highest life expectancy?"

expressions, which are then 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. 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. Figure 4 shows a screenshot of START answering a user question with both text and images.

3. Integrating Vision and Language

In response to a user's English question, Spot queries its knowledge base of recognized events to dynamically generate a video clip that satisfies the user's information need. START assists in the process by understanding the question and translating it into an appropriate request to the event knowledge base. This section explains how Spot integrates language and vision to answer natural language questions.

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 in 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 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 by the following (simplified) expression:

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

In this representation, objects moving along paths are specified by a series of path primitives. These primitives capture the relationship between the object in motion and other (mostly stationary) objects. 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 prepositional phrases, which correspond to path primitives. Thus, our framework for capturing motion and paths serves not only as an expressive representation, but also as a powerful query language. Our structures can contain any number of path primitives; underspecification can be used to indicate missing or unknown knowledge. As another example, the well-known verse of a popular winter song:

> Over the river and through the woods, to grand-mother's house we go...

could be expressed as

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

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?

An essential component of our representational framework is the mapping from symbolic locations to on-screen coordinates. Currently, this is accomplished manually: a human must identify the interesting landmarks in a particular scene and map the screen location of those landmarks to their symbolic representations. In the future, such mappings could be constructed automatically or semi-automatically by clustering visible behavior into interesting regions and positing the existence of relevant landmarks in the scene.

To efficiently answer natural language questions, we have constructed a knowledge base for storing representations derived from raw video. Spot queries this store using partially instantiated CLiViR structures with unbound variables.

3.1 Answering Questions

Consider the query "Show me all cars leaving the garage." START translates the English question into the following query:

Query(MOVE(car?, [PATH Source(Garage_{57})]))

Since we have not yet integrated object recognition technology into Spot, size is used as a surrogate for object detection, i.e., blobs above a certain size are classified as "cars".

The translation of natural language questions into CLiViR queries is accomplished by augmenting the lexical entries of verbs such as *leave* and *enter* with fragments of CLiViR queries that are dynamically assembled and instantiated during parse-time. Such lexical knowledge is stored alongside the subcategorization frames associated with each verb. In this particular example, the direct object of *leave*, "the garage", is extracted from the parse tree, translated into its corresponding CLiViR location, Garage_{57}, and instantiated as the argument of the "Source" primitive of the path. The mapping from natural language locations to CLiViR locations is specified in terms of nouns and various syntactic constraints on those nouns (e.g., adjectives). Because our lexicon contains information about synonymy, different terms such as *parking structure* and *parking garage* all map to the same CLiViR symbol. Cardinal directions like "north" and "south" are treated as special instances of locations.

The resulting query is then executed against the database of stored event structures. In response, Spot dynamically generates video that contains only motion tracks satisfying the user request. START's syntactic machinery automatically allows Spot to 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, START's knowledge about natural language allows Spot to translate the question "Show me all southbound traffic." into the query

$$\text{Query(MOVE(car?, [PATH Source(Road}_{17}\text{) Direction(South)]))}$$

Accordingly, Spot displays a video showing only vehicles driving from north to south.

4. 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 shape that can be automatically extracted; see Aslandogan and Yu (1999) and Yoshitaka and Ichikawa (1999) for a survey. However, such systems are fundamentally incapable of capturing high-level semantics.

A method for overcoming the limitations of low-level feature indexing is to utilize textual annotations that may accompany multimedia content. For example, 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. Often unstructured text simply is not be the best representation for multimedia content. Furthermore, descriptive annotations often cannot be obtained automatically and require human labor to gather. In contrast, one major advantage offered by our approach stems from the compositional nature of our semantic representation. Instead of using fixed annotations to described pre-defined multimedia segments, our CLiViR framework allows complex events to be built up from simple primitives. Although we must still specify the semantics for individual primitives, a compositional semantics allows unparalleled flexibility and the ability to describe novel unseen events.

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 for extracting and modeling high-level semantic events.

5. 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 existing representations with attributes such as time, speed, color, size, etc. to allow richer questions:

> Show me the last delivery truck that stopped in front of the office.
> How many people walked out of the building in the last 15 minutes?
> Display all blue cars entering the garage from the south this morning.

We are presently in the process of developing a language that can manipulate primitive building blocks such as paths and path primitives to craft more complex events. A significant portion of this effort focuses on exploring possible ways in which primitives can be combined to form more complex events such as sequences or causal chains. From this, we hope to develop a meaningful set of "connectors" to explicitly relate primitive events. This event language would allow users to ask very interesting questions:

> 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 like textures and colors, e.g., for events such as 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. By building a suitable abstraction between language and vision for capturing high-level semantic events, we can build systems that provide users with 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) 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?"

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, Allen's work (1983) on qualitative relations between temporal intervals 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, representations focusing on change and transitions between states (Borchardt, 1992), rather than states themselves, could also be helpful in the development of CLiViR.

6. Challenges

Our endeavor seeks to bring together maturing technologies from different fields to tackle fundamental problems in artificial intelligence. Although increasing trends towards specialization have resulted in many spectacular breakthroughs in narrow domains, the fragmentation of artificial intelligence research into a multitude of sub-disciplines has lead to stagnation of progress in "grand challenges”. We believe that technologies from various disparate areas are mature enough to support integration into a common framework, and that the potential synergies resulting from such integration will allow us to develop applications with far-reaching consequences.

Nevertheless, in pursuing our objectives, we may encounter both theoretical and practical challenges. The inherent assumption of our shared representation is the decompositional nature of knowledge across different modalities. Although preliminary successes are encouraging, it remains an open research question whether complex events can indeed be broken down into simple primitives. Furthermore, can primitives then be assembled to productively describe novel and unanticipated events? If the expansion of a system like Spot into new domains leads to an endless proliferation of new primitives, then perhaps our common representation approach to integrating vision and language is untenable. Perhaps, even if this were the case, the failure of our approach would teach us something about the nature of knowledge itself.

Aside from theoretical considerations, there are practical concerns with the scaling of video knowledge bases. Currently, Spot can successfully handle days worth of video (on the order of hundreds of megabytes). However, real-world applications might require storage of months, and maybe even years, of footage. The vast amounts of data involved, coupled with the rich symbolic structures derived from them, create a potential problem in the storage, indexing, and retrieval of video knowledge. Nevertheless, we are confident that such issues can be addressed through a combination of increased computing power and better algorithms.

7. Conclusion

The integration of vision and language technologies presents both difficult challenges and exciting opportunities for information access systems. We have taken several steps towards crossing the chasm between different modalities—not only have we implemented a prototype question answering system for 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 human intelligence. People are able to effortlessly integrate visual and linguistic data to reason and learn, an ability far beyond that of present day computers. By creating a framework for bridging 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.

8. Acknowledgements

We would like to thank Gary Borchardt, Sue Felshin, Greg Marton, and Stefanie Tellex for their comments on earlier drafts of this paper. This research was funded in part by MIT's Lincoln Laboratory and the Oxygen Project.

9. 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 Structures.* Cambridge, Massachusetts: MIT Press.

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. 1988. Using English for Indexing and Retrieving. In *Proceedings of the 1st RIAO Conference on User-Oriented Content-Based Text and Image Handling (RIAO '88).*

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. In *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB 2002).*

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).*

Smeaton, A.; and Qigley, I. 1996. Experiments on Using Semantic Distances between Words in Image Caption Retrieval. In *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. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*.

Stauffer, C.; and Grimson, W.E.L. 2000. Learning Patterns of Activity Using Real-Time TSracking. *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. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*.

Voorhees, E.M. 2002. Overview of the TREC 2002 Question Answering Track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*.

Wactlar, H.; Hauptmann, A.; Christel, M.; Houghton R.; and Olligschlaeger, A. 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.