Qibin Sun
Infocomm Research

A Natural Language-Based Interface for Querying a Video Database

Onur Küçüktunç, Uğur Güdükbay, Özgür Ulusoy
Bilkent University

We have recently developed a video database system, BilVideo, which provides integrated support for spatiotemporal, semantic, and low-level feature queries.[1] As a further development for this system, we present a natural language-based interface for query specification. This natural language processing (NLP)-based interface lets users formulate queries as sentences in English by using a part-of-speech (POS) tagging algorithm. The system then groups the specified queries as object-appearance, spatial, and similarity-based object trajectory queries by using POS tagging information.

It sends the queries constructed in the form of Prolog facts to the query processing engine, which interacts with both the knowledge base and object-relational database to respond to user queries that contain a combination of spatiotemporal, semantic, color, shape, and texture video queries. The query processor seamlessly integrates the intermediate query results returned from these two system components. The system sends the final results to Web clients.

What motivates our work is the need for a convenient and flexible natural language-based interface to complement the text-based query interface and the visual query interface of the BilVideo system, because specification of spatial queries using text or visual interfaces is not very easy for novice users. (For examples of how others have attempted to handle these issues, see the "Related Work" sidebar, next page.)

Thus, we developed a natural language-based query interface that's convenient and offers greater flexibility when specifying queries. The POS-based pattern-matching approach we use in identifying queries helps users specify queries without conforming to strict rules. This approach also lets us adjust our query interface easily as we add new query types to the BilVideo system.

BilVideo

BilVideo's spatiotemporal queries contain any combination of spatial, temporal, object-appearance, external-predicate, trajectory-projection, and similarity-based object trajectory conditions. A rule-based system built on a knowledge base processes these types of queries. A feature database maintained by an object-relational database also responds to semantic (keyword, event/activity, and category-based) and low-level feature queries (color, shape, and texture). BilVideo has the following parts:[1]

- an object extractor, which extracts salient objects from video keyframes;
- a fact extractor, which extracts spatiotemporal relations between video objects and store them in the knowledge base as facts;
- a video annotator, which extracts semantic data from video clips and stores them in the feature database to query video data for its semantic content;
- a Web-based visual query interface, which specifies queries by using visual sketches and displays the results; and
- a textual query language, which specifies queries using an extended Structured Query Language (SQL).

Currently BilVideo uses a Web-based visual query interface for query specification (available at http://www.cs.bilkent.edu.tr/~bilmdg/bilvideo). The interface uses specification windows for spatial and trajectory queries. The specification and formation of these queries vary significantly; therefore, the interface creates specific windows to handle each type. Users can query the temporal content of videos by combining these two types of primitive queries with temporal predicates (before, during, and so on). Users can also

1070-986X/07/$25.00 © 2007 IEEE

Published by the IEEE Computer Society

Related Work

The following notable works on the use of natural language-based interfaces for querying in multimedia databases have influenced our approach. Keim and Lum[1] propose a visual query specification for multimedia database systems based on natural language processing combined with visual access to domain knowledge. They note that using a system with visual support makes the process of query specification easier for all types of databases.

Our system, BilVideo, uses an SQL-like query language. It also uses a Prolog-based fact representation for spatiotemporal relations, whereas Keim and Lum base their work on a relational data model and SQL query language, especially for tables and attributes.

Gayatri and Raman[2] describe a natural language interface for a video database. The proposed interface maps the natural language queries to the textual annotation structures formed for the database objects. The principles of conceptual analysis are the foundation for the language understanding approach they use. Their conceptual analyzer consists of three passes: the first pass identifies the category of each word; the second pass identifies the words in their verb form acting as nouns; and the third pass disambiguates and fills the slots of actor–object expectations.

Erözel et al.[3] present a natural language interface for a video data model. In their work, they parse and extract queries in natural language with respect to the basic elements of a video data model (which are objects, activities, and events). They provide spatial properties either by specifying the enclosing rectangle, or some 2D relations, such as above, or right. Nevertheless, they don't consider many spatial queries—including topological and 3D relations—in their work.

Wang et al.[4] provide a Chinese natural language query system for relational databases. They have a similar approach to ours for processing the natural language: segmenting the query into words, parsing the segmented statement, and building the SQL statement. However, they built their work on traditional Chinese rather than English, and it doesn't specify spatial or temporal relations for a video database.

A recent work by Zhang and Nunamaker[5] addresses integrating natural language processing and video indexing. They propose a natural language approach to content-based video indexing and retrieval to identify appropriate video clips that can address users' needs.

References

1. D.A. Keim and V. Lum, "Visual Query Specification in a Multimedia Database System," *Proc. 3rd IEEE Conf. Visualization*, IEEE,1992, pp.194-201.
2. T.R. Gayatri and S. Raman, "Natural Language Interface to Video Database," *Natural Language Eng.*, vol. 7, no.1, 2001, pp.1-27.
3. G. Erözel, N.K. Çiçekli, and İ. Çiçekli, "Natural Language Interface on a Video Data Model," *Proc. Int'l Assoc. for Science and Technology for Development* (IASTED) *Int'l Conf. Databases and Applications*, IASTED, pp.198-203.
4. S. Wang, X.F. Meng, and S. Lui, "Nchiql: A Chinese Natural Language Query System to Databases," *Proc. Int'l Symp. Database Applications in Non-Traditional Environments*, IEEE CS Press, 1999, pp. 453-460.
5. D. Zhang and J.F. Nunamaker, "A Natural Language Approach to Content-Based Video Indexing and Retrieval for Interactive E-Learning," *IEEE Trans. Multimedia*, vol. 6, no. 3, 2004, pp. 450-458.

formulate queries by visual sketches. The system automatically computes most of the relations between salient objects specifying query conditions based on these sketches.[2]

Figure 1 shows an example query specification by visual sketches in BilVideo. Suppose that a user wants to retrieve the video segments where James Kelly is to the right of his assistant. The user specifies the relation type (spatial), adds the objects mentioned in the query one by one (James Kelly and his assistant) and draws the minimum bounding rectangles of the objects to express their relative spatial positions. The system combines the query predicates (east, appear, and so on) to form the compound query (see Figure 2).

Part-of-speech tagging process

Part-of-speech tagging is the process of marking up the words in a text with their corresponding parts of speech (see http://en.wikipedia.org/wiki/Part_of_speech_tagging). This process is harder than just having a list of words identified by their parts of speech, because some words can represent more than one part of speech at different times. In many languages, a huge percentage of word forms are ambiguous. There are eight parts of speech in English: nouns, verbs, adjectives, prepositions, pronouns, adverbs, conjunctions, and interjections. However, there are many more categories and subcategories in practice.

Some current major algorithms for POS tagging are the Viterbi algorithm,[3] Brill Tagger,[4] and Baum-Welch algorithm[5] (also known as the forward-backward algorithm). We used the MontyTagger, which is a rule-based POS tagger. It's based on Eric Brill's 1994 transformation-based learning POS tagger and uses a Brill-compatible lexicon and rule files.

Part-of-speech tagging is an indispensable part of natural-language-processing systems. The following excerpt from the MontyTagger Web site (see http://web.media.mit.edu/~hugo/montytagger) describes how the MontyTagger parses the sentences and extracts the correct parts of speech:

> MontyTagger annotates English text with part-of-speech information, e.g., 'dog' as a noun, or 'dog' as a verb. MontyTagger takes a bit of text as input, e.g., "Jack likes apples" and produces an output for the same text where each word is annotated with its part-of-speech, e.g., "Jack/NNP likes/VBZ apples/NNS".

MontyTagger uses the Penn Treebank tag set (see http://www.mozart-oz.org/mogul/doc/lager/

brill-tagger/penn.html). Table 1 (next page) displays a subset of the tags frequently used in our application.

Understanding queries after tagging

BilVideo query language supports three basic types of queries: object queries, spatial queries, and similarity-based object trajectory queries. We can group spatial relations into three subcategories: topological relations, directional relations, and 3D relations. We group these queries to define a general pattern for each group. After tagging all the words in a sentence, the order of POS information gives us an idea of how to define a pattern for this query.

Object appearance and spatial queries

Object queries retrieve segments of a video where the object appears. Table 2 gives examples of this query type. Spatial queries define spatial properties of objects with respect to other objects (see Table 3). There are three subcategories:

- topological relations describe order in 2D space (such as disjoint, touch, inside, contain, overlap, cover, and coveredby);
- directional relations describe the position of objects (directions, such as north, south, east, west, northeast, northwest, southeast, southwest, and neighborhood descriptions, such as left, right, below, and above); and
- 3D relations describe object positions in 3D space (infrontof, strictlyinfrontof, behind, strictlybehind, touchfrombehind, touchedfrombehind, and samelevel).

Similarity-based object trajectory queries

It's impossible to give all the information about a moving object in only one sentence; we have to define the lists of directions, displacements, and intervals to model a trajectory fact. In other words, formulating similarity-based object trajectory queries using visual sketches is more appropriate as compared to using spoken language for this purpose. Table 4 (on page 86) gives an example of this query type.

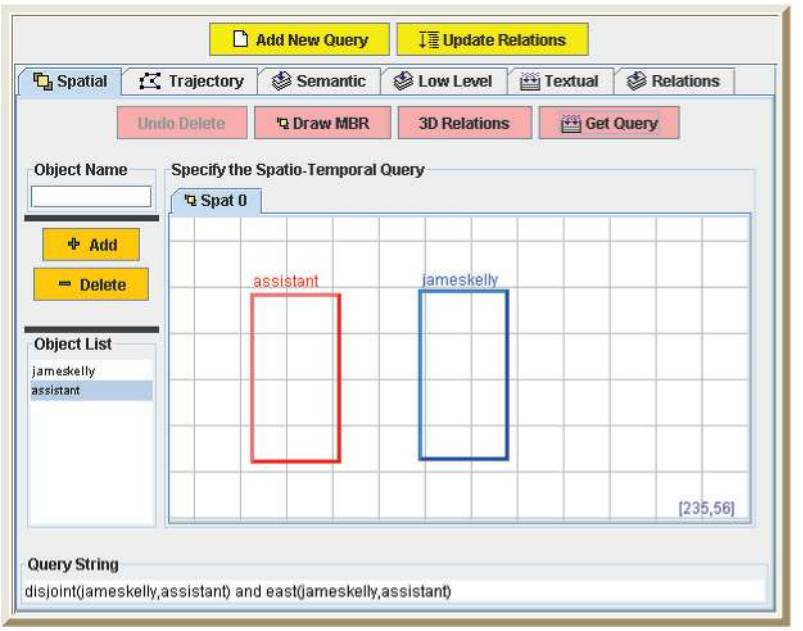

Figure 1. Query specification by visual sketches in BilVideo.

Query-construction algorithm

When the user inputs a query sentence, our algorithm (see Figure 3, page 87) constructs a query the processor can execute. The following example illustrates the steps in processing a sentence and constructing the corresponding query by using the algorithm in Figure 3. Our sentence is "Alfredo Pele was in front of his class."

1. We can't find the word "where" in the sentence.

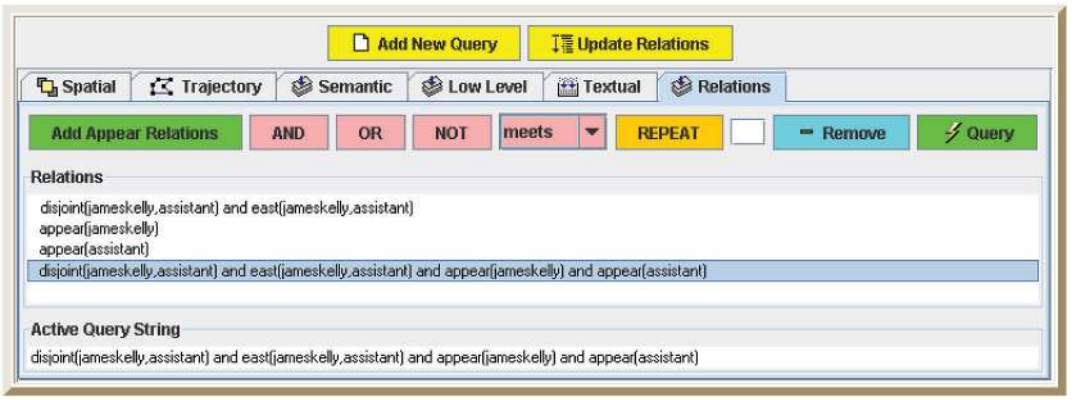

Figure 2. BilVideo query format generated for the visual query specified in Figure 1.

Table 1. A subset of the Penn Treebank tag set.

Part-of-Speech Tag	Description	Example
CC	coordination conjunction	and
DT	determiner	the
IN	preposition, conjunction	in, of, like
JJ	adjective	green
NN	noun, singular or mass	table
NNS	noun plural	tables
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
RB	adverb	however, usually
VB	verb, base form	take
VBD	verb, past tense	took
VBG	verb, gerund/present	taking
VBN	verb, past participle	taken
VBP	verb, singular present, non-third person	take
VBZ	verb, singular present, third person	takes

Table 2. Object queries.

Relation	Query Example	Results of Tagging	Possible Tag Ordering
Appear	James Kelly appears with his assistant	James/NNP Kelly/NNP appears/VBZ with/IN his/PRP$ assistant/NN	NN VB IN NN
Appear	James Kelly and his assistant appear.	James/NNP Kelly/NNP and/CC his/PRP$ assistant/NN appear/VB	NN CC NN VB

2. There are no delimiters.

3. After getting the output from the tagger, the system creates a sentence: "Alfredo/NNP Pele/NNP was/VBD in/IN front/NN of/IN his/PRP$ class/NN."

4. The system merges the proper noun "Alfredo/NNP Pele/NNP," making the sentence "AlfredoPele/NNP was/VBD in/IN front/NN of/IN his/PRP$ class/NN."

5. The system creates a word instance for each word/type pair.

6. Nouns, verbs, and prepositions are important for identifying the query, but the system can ignore the possessive pronoun "his" for our sentence.

7. The system compares the sentence pattern ("NN VB IN NN IN NN") to the query patterns that exist in the BilVideo query language, which matches a spatial query (a 3D relation).

8. No combined objects exist in our sentence, so query processing ends up with the result "infrontof(AlfredoPele, class)."

Possible problems and proposed solutions

Prolog's language format has brought up some challenging issues in processing sentences. Additionally, multiple subjects or objects as well as proper nouns present problems for sentence-processing. In this section, we consider problematic instances and show how we resolved them.

Table 3. Spatial relations: topological, directional, and 3D.

Relation	Query Example	Results of Tagging	Possible Tag Ordering
Disjoint	The player is disjoint from the ball.	The/DT player/NNP is/VBZ disjoint/NN from/IN the/DT ball/NN	NN VB NN IN NN
Overlap	The carpet overlaps the wall.	The/DT carpet/NNP overlaps/VBZ the/DT wall/NN	NN VB NN
Covers	The blanket covers the bed.	The/DT blanket/NNP covers/VBZ the/DT bed/NN	NN VB NN
North	Atlanta-Galleria is north of Atlanta.	Atlanta/NNP -/: Galleria/NNP is/VBZ north/RB of/IN Atlanta/NNP	NN VB RB IN NN
Left	The hall is on the left of the driveway.	The/DT hall/NNP is/VBZ on/IN the/DT left/VBN of/IN the/DT driveway/NN	NN VB IN VB IN NN
Strictlyinfrontof	David is strictly in front of the council.	David/NNP is/VBZ strictly/RB in/IN front/NN of/IN the/DT council/NN	NN VB RB IN NN IN NN
Samelevel	The target is the same level as the player.	The/DT target/NNP is/VBZ same/JJ level/NN as/IN the/DT player/NN	NN VB JJ NN IN NN
Behind	The ball is behind the goalkeeper.	The/DT ball/NN is/VBZ behind/IN the/DT goalkeeper/NN	NN VB IN NN

Proper nouns

For reasons related to database access, the system eliminates spaces in proper nouns to merge them into a single word. Proper nouns processed in this way are either a name and surname pair, or an alias of a well-known person. For example, "James Kelly" in a sentence becomes "JamesKelly" in the query. James Kelly will be tagged as "James/NNP Kelly/NNP," so we can consider this proper noun pair as the name of a person. Similarly, "King/NNP of/IN Italy/NNP" and "Alexander/NNP the/DT Great/NNP" must be taken as a single proper noun. In this case, proper nouns are separated with a preposition "/IN" or a determiner "/DT."

Using "and" for sentences

When the system finds a conjunction at the beginning of the clause, it understands that this clause is connected to the previous clause. It uses this information to understand multiple queries in a sentence, such as "Retrieve all segments in video clips where Ali is behind the wall and Ali is inside the house."

Using "and" for subjects or objects

When the system finds a conjunction while processing the sentence, the query construction algorithm combines two nouns that are separated by this conjunction. This process is required for checking the pattern of the query sentence with multiple subjects or objects.

For example, in the sentence "Apples and bananas are fruits," the algorithm tags the subject of the sentence as "Apples/NN and/CC bananas/NN." First it processes apple, putting it into the word array. Since the next word is a conjunction, we remove apple from the word array and put it into the stack. If the next word is also a noun (as in this example), we put an "&" sign between the nouns, and then put these words together into the word array as a single noun.

Queries using SQL

For a sentence like "Retrieve all news video clip segments where James Kelly moves west," the program needs to identify queries for "Retrieve all news video clip segments" and the rest. Because of this, we can use "where" to separate the sentence.

Combined subjects or objects in a query

If there's a combined subject or object in a query, we have to separate it. For example, "appear(JamesKelly & assistant)" must be transformed into "appear(JamesKelly) and appear(assistant)."

Table 4. Similarity-based object trajectory queries.

Relation	Query Example	Results of Tagging	Possible Tag Ordering
Moves	James Kelly	James/NNP Kelly/NNP	NN NN RB
	moves north	moves/NNS north/RB	

```
1. Replace all 'where' occurrences with a dot sign;
2. Split text into sentences with delimiters;
// The delimiters are period, comma, exclamation mark,
// question mark, semicolon, colon, etc.
foreach Sentence do
    3. TaggedSentence = MontyTagger(Sentence);
    4. Merge proper nouns in TaggedSentence;
    5. Split TaggedSentence into tokens;
    // Each token has the structure 'word/type'
    // Thus, we create instances of word class
    // by using 'word' and 'type' identifiers
    foreach Token in TaggedSentence do
        6. if the type of word is one of the types that we can use then
            // for example, verb, noun, preposition, adverb, conjunction
            Insert Word into the word array for the sentence
        else
            Discard any other parts-of-speech
            // such as articles 'a' and 'the'
        end
    end
    7. Compare the pattern of the processed sentence
            to the pre-described query patterns
    8. if there is a combined object in the query then
        Separate it into simple objects
    end
end
```

Figure 3. Query-construction algorithm.

Examples

We can also express the query visually specified in Figure 1 in natural language as "Retrieve segments where jameskelly is to the right of his assistant" (see Figure 4, next page). This NLP query specification is much easier and more convenient. Figure 5 shows the result of the query in terms of a set of video frame intervals and Figure 6 (page 89) shows still frames from the queried video and indicates the frames that are returned as the query solution.

The examples in Tables 5 and 6 (next page) illustrate different query sentence types and the corresponding queries in BilVideo query format. For the similarity-based trajectory query, it assumes a default similarity threshold (sthreshold) value of 0.75 and a time gap (tgap) value of one.[2]

The querying capabilities of BilVideo limits sentence structures and natural language representations supported by our NLP interface. Although it supports a wide variety of query types—including object queries, spatial (topo-

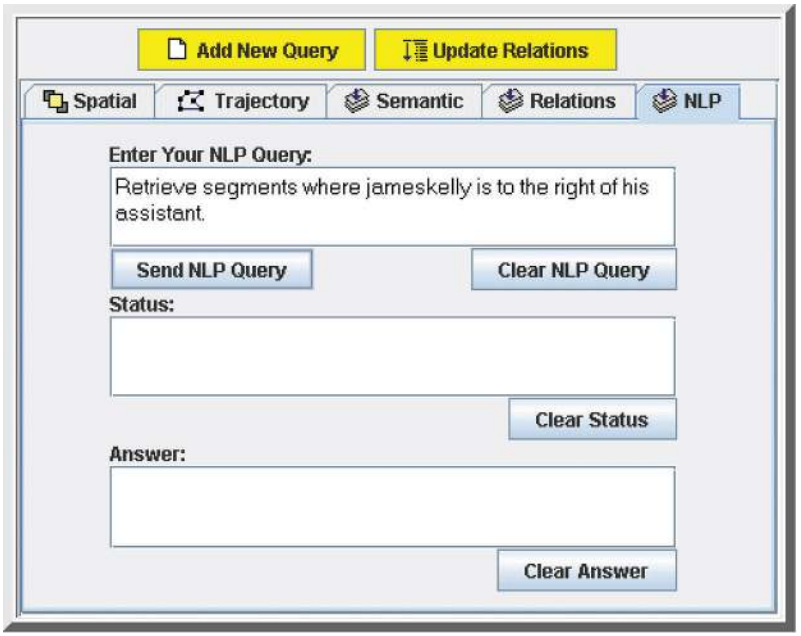

Figure 4. Query specification in natural language.

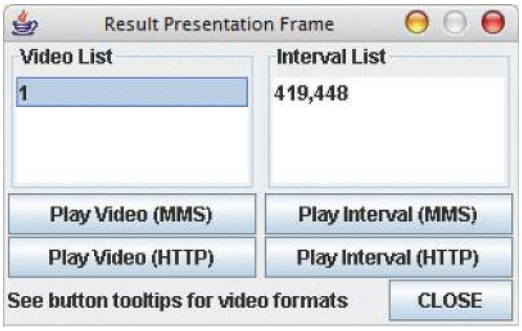

Figure 5. The result of the query given in Figure 4 as a set of video frame intervals.

logical, directional, and 3D) queries, and similarity-based object trajectory queries—the set of relationships that might be formulated between video objects is not exhaustively complete, like in any other video querying system.

In addition to the sentence structures currently supported, the NLP interface can also process more complex sentences by rewriting the corresponding query as a combination of basic query types. As an example, "A is between B and C" is a complex query because our topological, directional, and 3D relations generally have two arguments. The interface can convert it as "(right (A, B) && left(A, C)) || (left(A, B) && right(A, C))," as shown in Table 6. It can also express the same relation between A, B, and C as a combination of directional relations, such as north, west, and so on.

Our current works include the implementation of hand gesture-based interaction for the specification of spatial relations between objects of interest in videos, especially the 3D relations.

Table 5. Query examples.

Type	Natural Language Processing Query	BilVideo Query Format
Object	James Kelly appears with his assistant.	appear(JamesKelly) and appear(assistant)
Topological	The carpet overlaps the wall.	overlaps(carpet, wall)
Directional	The galleria is northeast of Atlanta.	northeast(galleria, Atlanta)
Neighborhood	The hall is on the right of the driveway.	right(hall, driveway)
3D relation	Alfredo is in front of his class.	infrontof(Alfredo, class)
Trajectory	James Kelly moves north.	(tr(JamesKelly, [[north]]) sthreshold 0.75 tgap 1)
SQL-like query	Retrieve all news video clip segments where James Kelly is on the right of his assistant.	select * from video clip segments where right(JamesKelly, assistant)

Table 6. Complex query examples.

Natural Language Processing Query	BilVideo Query Format
Mars and Venus may contain water.	contain(Mars, water) and contain(Venus, water)
The bird is inside the cage and the house.	inside(bird, cage) and inside(bird, house)
Retrieve all news video clip segments where James Kelly is on the right of his assistant, and James Kelly is at the same level as his assistant.	select * from video clip segments where right(JamesKelly, assistant) and samelevel(JamesKelly, assistant)
James Kelly is between the girl and his assistant.	(right(JamesKelly, girl) and left(JamesKelly, assistant)) or (left(JamesKelly, girl) and right(JamesKelly, assistant))
James Kelly appears without his assistant.	appear(JamesKelly) and not(appear(assistant))

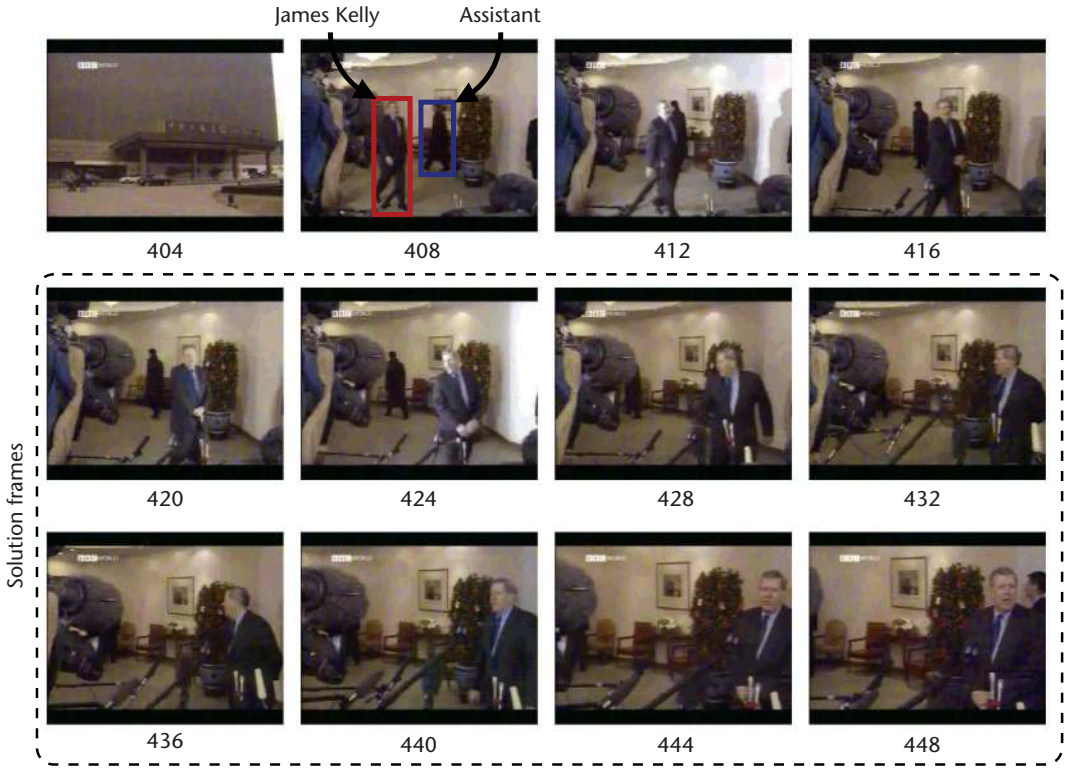

Figure 6. Still frames from the queried video and the frames returned as the query solution.

The implementation is based on taking hand movements as input with the help of a camera. Since different query types require different interaction modalities, a multimodal interface, including visual (mouse-based), natural language-based, and gesture-based components, will be most appropriate for query specification. We are also working on extending BilVideo to provide database support for automated visual surveillance. We are in the process of extending our query types by integrating semantic features (for example, events, subevents, and salient objects) and low-level object features (for example, color, shape, and texture) for surveillance videos. This will be used to support queries related to left-object detection, intruder detection, and so on.

For future work, we're planning to elaborate on the use of adjectives and identify negative meanings in query sentences. We're also planning to incorporate speech recognition into our NLP-based interface so that the queries can be entered using spoken language. **MM**

Acknowledgments

This work is supported by the European Union Sixth Framework Program under grant number FP6-507752 (MUSCLE NoE Project) and the Scientific and Technical Research Council of Turkey (TÜBİTAK) under grant number EEEAG-105E065. We're grateful to Kirsten Ward for her proofreading and suggestions.

References

1. M.E. Dönderler et al., "BilVideo: A Video Database Management System," *IEEE MultiMedia*, vol. 10, no. 1, 2003, pp. 66-70.
2. M.E. Dönderler, Ö. Ulusoy, and U. Güdükbay, "Rule-Based Spatiotemporal Query Processing for Video Databases," *Very Large Databases* (VLDB) *J.*, vol. 13, no. 1, 2004, pp. 86-103.
3. A.J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm," *IEEE Trans. Information Theory*, vol. IT-13, 1967, pp. 260-269.
4. E. Brill, "A Simple Rule-Based Part-of-Speech Tagger," *Proc. Third Conf. Applied Natural Language Processing*, Assoc. for Computational Linguistics, 1992, pp.152-155.
5. L.E. Baum, "An Inequality and Associated Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *Inequalities*, vol. 3, 1972, pp.1-8.

Readers may contact Özgür Ulusoy at oulusoy@cs.bilkent.edu.tr.

Contact Multimedia at Work editor Qibin Sun at qibin@i2r.a-star.edu.sg.