# A Natural Language Approach to Content-Based Video Indexing and Retrieval for Interactive E-Learning

Dongsong Zhang and Jay F. Nunamaker

***Abstract*—As a powerful and expressive nontextual media that can capture and present information, instructional videos are extensively used in e-Learning (Web-based distance learning). Since each video may cover many subjects, it is critical for an e-Learning environment to have content-based video searching capabilities to meet diverse individual learning needs. In this paper, we present an interactive multimedia-based e-Learning environment that enables users to interact with it to obtain knowledge in the form of logically segmented video clips. We propose a natural language approach to content-based video indexing and retrieval to identify appropriate video clips that can address users' needs. The method integrates natural language processing, named entity extraction, frame-based indexing, and information retrieval techniques to explore knowledge-on-demand in a video-based interactive e-Learning environment. A preliminary evaluation shows that precision and recall of this approach are better than those of the traditional keyword based approach.**

***Index Terms*—Learning by asking, natural language processing, video indexing and retrieval.**

## I. Introduction

As ONE OF THE key driving forces in the 21st Century, information technology is changing the fundamental ways people learn. In order to increase access to knowledge by the rapidly growing population and to meet the needs of lifelong learning, acquisition of knowledge is not restricted to taking place in traditional classrooms. Learning methods are becoming more and more portable, flexible, and adaptive.

The Internet has been widely adopted as a medium for network-enabled transfer of skills, information, and knowledge [1]. Web-based distance learning, empowered by the Internet and telecommunication technologies, supports significant improvement in the delivery of online courses and training. Traditionally, online learning material was primarily in text format. Today, advances in communication and multimedia technologies make providing multimedia learning content to remote students via the Internet a reality, enabling users to take advantage of diverse human senses and increase their interest in learning. However, to accomplish this, it is critical to be able to index and retrieve multimedia content in an efficient and effective manner. It is far more challenging to deal with multimedia data than to work with pure texts.

In this paper, we propose a novel approach to content-based video indexing and retrieval in an interactive e-Learning environment. We consider e-Learning to be the type of learning situation in which a learner receives electronic education or training material via the Internet. First of all, we will introduce the Learning by Asking (LBA) project, which aims at developing an interactive multimedia-based e-Learning environment. Then, following an overview of different video indexing schemes, we propose a natural language approach to content-based video indexing and retrieval, which has been implemented in LBA. Our approach integrates several information technologies including information retrieval, information extraction, natural language processing, and question-answering. Finally, an evaluation study and future work will be discussed.

## II. Multimedia in E-Learning

The latest multimedia technology carries multimedia content such as audio, video, image, and text over ever-increasing network bandwidth. It is having a dramatic impact on both the process and product of learning. Multimedia instructions have become very attractive and promising in e-Learning because 1) they tap the feelings and emotions of people, and 2) a multisensory learning environment can maximize learners' ability to retain information [2]. Research has shown that multimedia instructions can enhance an individual's problem-solving skills and improve learning effectiveness [3]. For example, video is by far one of the most powerful and expressive nontextual media that can capture and present information [4]. Many researchers have conducted studies on distance education that makes use of instructional videos, and results have shown that the performance level is comparable to that of traditional classroom learning [5], [6]. With advancements in multimedia technology, a number of interactive learning systems based on instructional videos have been developed [7]–[9].

In a video-based e-Learning system, the key challenge is to provide learners with easy, intuitive, and fast access to instructional videos in which they are interested [10]. In comparison with browsing a text, in which a quick glance is sufficient to filter information, browsing a video is much more time consuming. The real difference between a video and a text is that video has constant-rate outputs that cannot be changed without

Manuscript received November 1, 2001; revised September 22, 2002. This work was supported by the Ford Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bernard Merialdo.

D. Zhang is with the Department of Information Systems, University of Maryland-Baltimore County, Baltimore, MD 21250 USA (e-mail: zhangd@umbc.edu).

J. F. Nunamaker is with the Center for the Management of Information, The University of Arizona, Tucson, AZ 85721 USA (e-mail: jnunamaker@cmi.arizona.edu).

Digital Object Identifier 10.1109/TMM.2004.827505

impacting negatively on people's ability to extract information. Consider trying to find video clips of interest in a four-hour long videotape. In the analog domain, this task would be tedious and make people frustrated because they have to view and listen to the whole video in order to identify all relevant parts. Some control features of VCRs, such as fast-forward and rewind, do not help much, since there is no distinguishable audio during fast-forward or rewind operations. Simply digitizing the video will not make the job easier unless digital videos provide structural support in terms of content. Therefore, content-based video indexing and retrieval capability is highly desirable in an interactive e-Learning environment.

In most reported e-Learning research, instructional videos such as videotaped lectures were available to remotely located students through either Web broadcasting or online access. Although videos were digitized, they were not processed into a hierarchical structure in terms of content. The lack of structure of videos can cause several problems in an e-Learning environment. For example, an instructional video may last a couple of hours and cover many subtopics, but students have little control over its content. It is difficult for students to skip a part of a video that they already know or are not interested in. It is very difficult to search for a specific subtopic in an unstructured video. Consider a video lecture lasting two hours. Without content-based video searching capability, people must spend two hours to watch the whole video in order to find all its content that is relevant to a subject, although parts or even most of the video may be irrelevant to their interest. Obviously, to make instructional videos sufficiently well-structured and organized to enable Knowledge-on-Demand (KoD) is a next goal in interactive e-Learning. In recent years, intensive effort to index and analyze video content based on its structural properties has been emerging.

## III. Learning by Asking Project

Learning By Asking (LBA) is a research project sponsored by the Ford Foundation. Its general objective is to develop an interactive multimedia-based e-Learning environment that enables efficient and flexible access to instructional videos in an online knowledge repository. People learn by interacting with the LBA system.

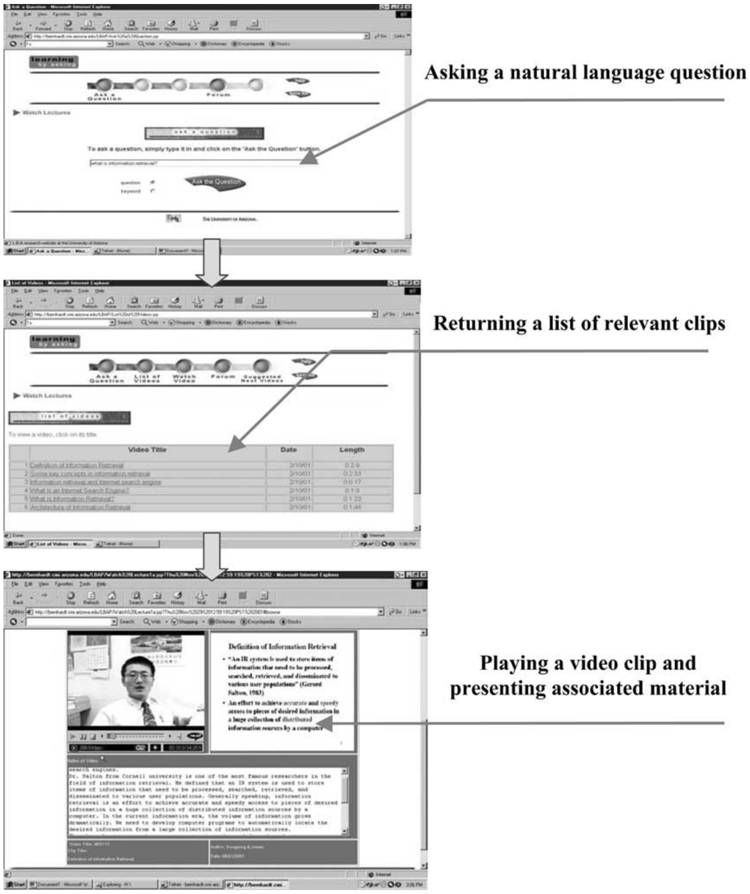
The system mainly consists of three components: a thin-client, a Web server and a video streaming server. The basic idea behind the LBA system can be described as follows. First, some domain experts are videotaped during their lectures or interviews that may take a couple of hours. The content of a video probably consists of many subtopics in the domain, varying from basic concepts to strategies, solutions, or potential applications. We consider such videos as domain expert knowledge. Videos then are digitized and logically segmented into individual clips based on content so that each clip focuses on a specific subtopic. Although a number of segmentation algorithms are available, content-based video segmentation still requires human assistance. In this study, we performed manual logical segmentation by identifying the time boundaries of each clip within a video. A clip is a stream of contiguous frames uniquely identified by its starting and ending time and is relatively neutral in meaning. The length of each video clip is normally in the range of one to three minutes. All videos are stored in a digital video library on a video streaming server. They are retrieved in response to users' requests.

As an easy-to-use system, LBA only requires a thin-client—a user only needs a Web browser, a video player (RealPlayer from RealNetworks), and a sound card installed on his computer to access instructional content. S/He can ask LBA a question in everyday English and watch appropriate video clips and associated material retrieved by the system to answer the question.

In order to provide users with clip searching and analysis capability, a video metadata library containing various descriptive data about video clips such as titles, file size, names of speakers, keywords, starting/ending time, and content templates have been created on a Web server. Learning proceeds via continuous interaction between the user and the LBA system. A question asked by a user will be sent to the Web server, on which the primary information processing and retrieval will take place. After question analysis, the system will search the video metadata library for the best match. In other words, video clips whose content is likely to answer the question will be identified and ranked based on their relevance. Finally, the links of selected relevant clips, as well as other related material such as presentation slides and lecture notes, will be delivered to the user's computer. The user can play any returned clip immediately by clickling the corresponding link(s). Instructional videos, as well as associated material, are presented to the user in a cohesive and interactive manner (Fig. 1). It is implemented via logical connections among the content stored in the database. At any time, users can choose either to replay a clip as many times as desired or to ask new questions and repeat the above questioning process.

During this learning process, an intelligent Learning Assistant Module (LAM) in LBA will automatically generate dynamic and personalized learning guidance based on individual users' learning history, which includes the questions being asked or previously asked. For example, if a user were to ask 'What is skin cancer?', the LBA system might prompt a follow-up suggestion like 'Do you also want to know how to prevent skin cancer?' In addition, an online discussion forum integrated into LBA enables users to exchange ideas or post further questions about a video and to receive comments from their peers or qualified experts.

The current LBA system uses Apache on a Gateway dual Pentium III 800MHZ server as the Web server and RealSystem Server from RealNetworks as video streaming server. On the Web server, we have developed Java Servlet programs to deliver videos and other learning material to users. The video metadata is stored in a MySQL database, which can be accessed by Java Servlet programs via JDBC (Java Database Connectivity). When a relevant clip is identified, the LBA system obtains its location and time boundaries from the metadata library and dynamically generates a .smil file that will be used by the Real-



To ask a question, simply type it in and click on the "Ask the Question" button.

what is information retrieval?

○ question ○ keyword

To view a video, click on its title.

| | Video Title | Date | Length |
|---|---|---|---|
| 1 | Definition of Information Retrieval | 3/10/01 | 0:2:9 |
| 2 | Some key concepts in information retrieval | 3/10/01 | 0:2:33 |
| 3 | Information retrieval and Internet search engine | 3/10/01 | 0:0:17 |
| 4 | What is an Internet Search Engine? | 3/10/01 | 0:1:9 |
| 5 | What is Information Retrieval? | 3/10/01 | 0:1:23 |
| 6 | Architecture of Information Retrieval | 3/10/01 | 0:1:46 |

Definition of Information Retrieval

- "An IR system is used to store items of information that need to be processed, searched, retrieved, and disseminated to various user populations" (Gerard Salton, 1983)
- An effort to achieve accurate and speedy access to pieces of desired information in a huge collection of distributed information sources by a computer

```
search engines.
Dr. Salton from Cornell university is one of the most famous researchers in the
field of information retrieval. He defined that an IR system is used to store
items of information that need to be processed, searched, retrieved, and
disseminated to various user populations. Generally speaking, information
retrieval is an effort to achieve accurate and speedy access to pieces of desired
information in a huge collection of distributed information sources by a
computer. In the current information era, the volume of information grows
dramatically. We need to develop computer programs to automatically locate the
desired information from a large collection of information sources.
```

Fig. 1. Learning by asking questions.

System streaming server to retrieve the video clip. An example of .smil file is as follows.

```
< smil >
< body >
< video src ="rtsp : //videodemo.cmi.arizona.edu:
554/ramgen/online.rm"clip − begin ="1093s"clip −
end ="1182s"/ >
< /body >
< /smil >
```

The LBA system can be adopted for a large variety of applications, especially those where both visual and auditory information are involved, such as distance education, remote software technical support, online workforce training, and healthcare consultation. A typical example is workforce training. Traditionally, employees in a company have to leave their duties, travel to a designated place, and stay there for a few days to obtain training. From a financial point of view, in addition to high traveling costs, this type of centralized training can entail losing business due to employees' absence from their offices. From a knowledge management perspective, companies can manage and reuse their knowledge (training material) in a more efficient way. Under certain circumstances, it can be more cost and time effective to offer training through a system like LBA, which enables individual training at any location whenever it is needed.

## IV. Related Work on Video Indexing and Retrieval

Search is one of the core activities of a digital video library. Its result is a list of candidate videos whose content satisfies a query. Video indexing is a process of tagging videos and organizing them in an effective manner for fast access and retrieval. Automation of indexing can significantly reduce processing cost while eliminating tedious work. In the past decade, a variety of video indexing techniques have been investigated, involving a wide range of topics from computer vision, pattern recognition, speech recognition, natural language processing, image processing, video analysis, and information retrieval. Those techniques can be generally classified into four categories based on different cues that algorithms use in video indexing and retrieval, including visual features, auditory features, texts, and task models.

- *Visual Feature Based Approach*

  This method normally indexes videos based on their visual features such as shape, texture, and color histograms. It involves extensive image processing. Some algorithms use key frames automatically extracted from videos as indices. The basic idea is that every video clip has a representative frame that provide a visual cue to its content. Those representative frames (key frames) are automatically extracted from original videos based on their image features [11]–[13]. Visual-feature based approaches normally use image queries. The video retrieval relies on a set of similarity measures between image features of a query and those of key frames, which can be performed at three abstraction levels: raw data level, feature level, and semantic level [14].

Some indexing algorithms use objects and their attributes, as well as spatial and/or temporal relations among objects in a video to label and index video sequences [15]–[17]. For example, Gunsel *et al.* [17] introduced an approach to temporal video partitioning and content-based video indexing, in which the basic indexing unit was "life-span of a video object, rather than a camera shot or story unit." They indexed motion and shape information of video object planes tracked at each frame and provided an object-based access to video data.

The disadvantages of a visual-feature based approach are that users usually do not have an image handy to formulate a query and that content-based image retrieval has not reached a semantic level that is directly useful to users [18]. When a video has few scene changes, as in a typical videotaped lecture, a visual feature based approach will encounter serious problems in extracting key frames. In addition, defining quantitative measures of key frame similarity still remains a challenging research topic.

- *Auditory Feature Based Indexing*

  Sound is an essential component of a video. The audio track provides a rich source of information to supplement understanding of any video content. Audio information can also be used in video indexing [19]. Image/sound relationships are critical to the perception and understanding of video content. In a number of studies, both the auditory and visual information of videos have been used to extract high-level semantic information as indices [20], [21]. The parsing and indexing of audio-source and video-source often lead to the extraction of a speaker label and of a talking-face mapping of the source over time. Integration of these audio and visual mappings constrained by interaction rules results in higher levels of video abstraction and even partial detection of its context [22]. There are several useful methods for classifying and indexing sounds, such as simile (one sound is similar to another sound or a group of sounds in terms of certain characteristics), acoustical features (e.g., pitch, timbre, and loudness), and subjective features (e.g., describing sounds using personal descriptive language) [23]. Cambridge University has developed retrieval methods for video mail based on keyword spotting in the soundtrack via integrating speech recognition methods and information retrieval technology [24]. In addition, the audio track of a video is also often used to generate a text transcript by means of a speech recognition system for text-based analysis and retrieval [20].

- *Text Based Approach*

  In this approach, videos are indexed by keywords that are automatically extracted from texts related to videos [10], [25]. The retrieval methods rely on keyword search in the free text obtained either from closed captions or from transcriptions of the video soundtrack via speech recognition [26]. Other methods use text identified from video images for video indexing [27], [28]. For example, Lienhart [27] proposed a method to automatically recognize text appearing in videos by using OCR software for video indexing. After a user specified a search string, video sequences were retrieved through either exact substring match or approximate substring match. Text-based video indexing is straightforward and easy to implement. It allows random access to specific points within a video when a particular keyword appears. The major disadvantage of this approach is the loss of the context of search terms. For example, let us consider that one of two videos in a video digital library describes the symptoms of skin cancer and the other explains how to prevent skin cancer. Although both videos contain the same term "skin cancer", they have disparate contexts and address different questions. It will be difficult to distinguish those two videos by using a keyword-spotting approach.

- *Task Based Approach*

  In addition to visual, auditory, and textual cues, the semantic features of tasks can also be used to create video indexes. In general, there have been two ways to apply this approach. One is to create a structured content frame for each video clip as index. A frame is a data object that plays a role similar to that of a record in a relational database. It often consists of a set of fields, usually called slots, that provide various semantic information of a video clip [19], [29]. Burke and Kass [7] developed a video indexing scheme based on "Universal Indexing Frame" to retrieve video clips for presentation in a case-based teaching environment. The index frame contained slots such as "Anomaly", "Theme", "Goal", and "Plan", which explicitly indicated the points of interest or anomalies in a video story.

  The other type of method is to build a task model for a collection of videos within a particular domain. Researchers decompose a task into multiple subtasks or subgoals and generate a hierarchical structure, called a task model, for video indexing and retrieval [8], [30]. For example, in order to train novice transportation planners, Johnson *et al.* [30] developed a Trans-ASK system that contained 21 hours of video detailing the experience of United States Transportation Command personnel in planning for operations such as Desert Shield and Desert Storm. The videos were segmented into a collection of video clips in which experts told "war stories" of their actual experiences. In that system, a six-level hierarchy of objectives and targets was manually created, ranging from national security objectives to individual targets. Video clips were indexed according to a hierarchy of questions in the task model.

  There are some limitations of current task-based approaches. First, a task model is domain dependent and inflexible. Second, task frames or models are mainly created by human experts. It can be very time consuming and inefficient.

The reality is that, for applications such as digital library and interactive e-Learning, learners are ordinarily interested in querying and retrieving specific videos in terms of "what the video is about" rather than "what the video looks like"

[31]. Basically, attempts to automate video indexing based on semantic content face a challenge to generate a semantic representation of video content automatically and to measure the relevancy between a question and video content with that representation. Currently, most content-based video indexing and retrieval schemes rely on image processing and pattern recognition techniques. However, for an instructional video such as a videotaped lecture or seminar, there are few visual cues that can be used to distinguish different content or topics in the video, which mainly contains a talking head with few scene changes.

There has been extensive research effort on automatic video segmentation. A variety of segmentation schemes have been developed. They use rule-based methods based on input from multimedia streams (audio, video, and closed captions), machine learning techniques, or topic detection and tracking (TDT) methods [32]–[34]. For example, Boykin and Merlino [33] developed an automatically induced segmentation system using the Hidden Markov Model (HMM). In this research, we manually segmented instructional videos, such as videotaped lectures or interviews, into individual clips based on their content. The segmentation was done at a logical level instead of a physical level by simply identifying time boundaries of each clip within an entire video. We did it by hand because 1) unlike a TV news video consisting of a sequence of reports with frequent scene changes and interruptions for commercials that are commonly used as important clues for automatic video segmentation, most instructional videos have only one speaker and lack cues for segmentation, and 2) video segmentation is not the focus of this research. Therefore, in our study, video segmentation was a manual process. It normally took about 0.5–1 hours to segment a one-hour video using a software package called Final Cut Pro (Apple Computer, Inc.).

## V. A Natural Language Approach to Video Indexing and Retrieval

The LBA system aims at enabling users to ask questions in everyday English via a Web interface and then to watch a list of retrieved instructional video clips that likely contain the answers. Two primary advantages of using natural language queries over keyword queries are 1) natural language queries are more expressive than keywords so users find it easier to specify their needs, and 2) they provide more context than keyword queries, reducing the vagueness of users' description of their interests. In the past few years, considerable research has been done in the question-answering area [35], [36]. A number of information retrieval systems such as AskJeeves (http://askjeeves.com/), Synthetic Interviews [8], FAQ Finder [37], and START (http://www.ai.mit.edu/projects/infolab/) allow users to ask natural language questions.

Currently, we are primarily interested in investigating the following research questions.

- How to understand a natural language question?
- How to represent the semantic content of an instructional video clip?
- How to measure the relevance between a question and video clips for content-based retrieval?

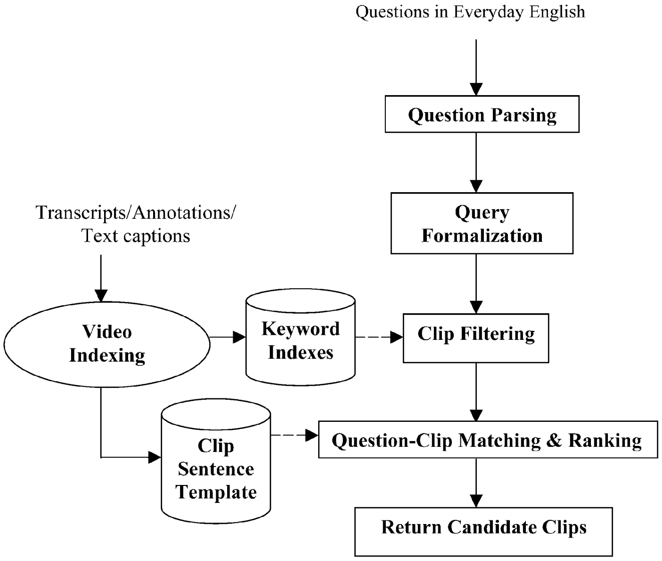


Fig. 2. Diagram of question understanding and clip retrieval in LBA.

Since we mainly deal with interview or lecture videos with few scene changes, commonly used indexing schemes based on visual features are not appropriate in our case. Task-model and auditory-feature based approaches are either too sophisticated or not sufficiently realistic to meet our needs. Transcribed texts or lecture notes are the major sources of information about video content. Carnegie Mellon University's Informedia Digital Video Library project employed a keyword-based search engine that uses keyword spotting, stop words, word stemming, and TF-IDF (term frequency and inverse document frequency) term weighting to search video transcriptions in order to retrieve relevant videos [38]. Their searching scheme requires a user to specify keywords of interests, and analyzes the relevance of video transcriptions based on word occurrence, not at a semantic level. Therefore, we developed a two-phase natural language approach to addressing the above research questions by integrating natural language processing (NLP) technology, named entity (NE) extraction, keyword and frame-based video indexing schemes, and information retrieval (IR) techniques. A diagram of question understanding and clip retrieval in the LBA system is shown in Fig. 2. The whole process proceeds automatically.

The goal of NLP research is to build computational models of language so that people can write computer programs capable of accomplishing various tasks involving natural language. It lays the foundation for question and clip understanding. In our approach, we adopted various NLP techniques, such as morphological analysis and part-of-speech tagging, to analyze natural language questions and video-related text to obtain both syntactic and semantic information.

The NE extraction task aims to analyze unrestricted text in order to extract specific types of information. Typically, this task consists of three subtasks—extraction of entity names, temporal expressions, and number expressions. The expressions to be annotated are "unique identifiers" of entities (e.g., organizations, persons, and locations), times (e.g., dates and time), and quantities (e.g., monetary values and percentages) [39]. The NE extraction task attempts to understand only those sentences/portions in a document that contain the specified information. It has direct practical value in annotating text so that documents can be searched for names, places, and dates, etc. A variety of techniques for NE extraction, such as syntactic analysis, semantic analysis, and machine learning, have been developed and evaluated [39]–[41]. The integration of NE extraction component enables the LBA system to understand video content by capturing semantic information from text. In this research, the LBA system integrates a commercial software package called Conexor iSkim [42] to perform natural language processing and NE tasks when LBA generates templates (frames) for both questions and clip content.

*Natural Language Question Parsing and Formalization:* After a user submits a question in conversational English through the Web interface, such as "What is skin cancer?", the LBA system starts to parse it by using Conexor iSkim, an NLP tool that can analyze an English sentence and produce five types of information for words in the sentence: part-of-speech (POS), lemma, morphology, light syntax, and named entity recognition. Given an example sentence:

"*Learning is an indispensable activity in our lives,*"

Results from iSkim will be as shown at the bottom of the page. The objective of question parsing is to automatically 1) determine the type of an expected answer, 2) determine the focus of the question, and 3) identify different roles of important words or phrases in a question and generate a template representation. We define a question template as having the following structure to represent the content of any given question.

```
< QUESTION TEMPLATE >:=
 Answer Type (type of information
  a question is looking for)
 Question Focus (the core noun)
 Person (named person)
 Organization (named organizations)
 Governor (key verbs)
 Objects (other noun or noun phrases)
 Number (numbers)
 Time (year, date, etc.)
 Location (country, region, city, etc.)
```

To determine whether a clip is relevant to a question or not, we need to clarify what the question is asking for. Ths slot *Answer Type* indicates the semantic characteristic of an expected answer that is determined according to question words. For example, if a question asks "When was Linux first released?", the *Answer Type* is ⟨TIME⟩. We define totally nine answer types based on previous research in question-answering [36], [43]: Person, Organization, Object, Location, Time, Number, Reason, Definition, and Undefined. Among them, names of persons and organizations, locations, time, and numbers contained in a question are obtained by combining NE analysis results from iSkim and some heuristic rules. When iSkim analyzes a sentence, it will automatically assign some predefined tags to the named-entity words it identifies, such as ⟨LOC⟩ for locations, ⟨TEMP⟩ for temporal expression, and ⟨ORG⟩ for names of organizations. Such extracted NE information will fill corresponding slots in the template of the current question. A list of rules is used to help the system determine the answer type. For example, one of the rules is

```
If a question starts with 'What' + person
Answer type =< Person >
```

*Question Focus* refers to the core noun or noun phrase in a question that indicates what the question is all about. Normally, the first noun or noun phrase after the questioning word in a question is its focus. For example, in the question "What is the largest country in the world?", "largest country" is the question focus. The LBA system uses shallow parsing to obtain the important information contained in a given question and fill slots in the question template. A stopword dictionary, which contains about 610 words, is utilized to filter out functional terms in the question, such as "a", "the", and "is", since they contribute little to the meaning of videos. As a result of this phase, a question template with appropriate slot values will be automatically generated for the question currently being asked.

*Video Clip Indexing:* In LBA, the textual description of video content is derived automatically either from a transcription of soundtracks of a video or from lecturers' scripts or lecture notes made during course preparation. A keyword indexing table and a sentence template table are then created automatically for video indexing. The keyword-indexing table contains keywords and phrases extracted from transcriptions or lecture notes, along with their TF-IDF term weights in each video clip [44]. In order to match question templates, the transcribed sentences in each candidate clip will be automatically analyzed and represented in a similar way. Therefore, a

| | Word | Stem | POS and other Info |
|---|---|---|---|
| 1 | Learning | learn | ING |
| 2 | is | be | V PRES SG3 & VA |
| 3 | an | an | DET SG & > N |
| 4 | indispensable | indispensable | A ABS & > N |
| 5 | activity | activity | N NOOM SG &NH |
| 6 | in | in | PREP &N < |
| 7 | our | we | PRON PERS GEN PLI & > N |
| 8 | lives | life | N NOM SG &NH |

sentence template table for candidate clips is also created. In the same manner as for questions, each sentence in a clip is tagged for part-of-speech and named entities by iSkim, and automatically generates a similar template with the same structure as the question template but without *Answer Type* and *Question Focus* slots. In this template table, each sentence template is a single record. These records are used for measuring the degree of relevance between a question and a video clip.

*Phase 1: Clip Filtering:* Considering the large number of video clips and the overhead of detailed linguistic analysis, it is unrealistic and unnecessary to search exhaustively through all video clips during video retrieval. Therefore, the first phase of video retrieval in LBA is to use a search engine based on some information retrieval techniques as a clip filter to select a small set of precandidate clips from the video repository that likely contain answers to the current question. The assumption here is that if a clip does not contain important keywords/phrases in the question, we can reasonably assume that this clip is not relevant. The main objective of this phase is to reduce system's total processing time and improve retrieval efficiency. In LBA, each document is a transcript file of an individual clip.

When a query comes in, the question parsing process automatically identifies keywords, including noun or noun phrases, verbs, quoted expressions, and named entities, from a question. Those nonstopwords, as well as their synonyms obtained from WordNet Synsets (http://www.cogsci.princeton.edu/~wn) and a domain ontology—a hierarchical concept space that defines semantic relationships between different terms in this domain, are passed to the clip filtering process. When a term has a large number of synonyms, the system chooses the first five, which are the most frequently used synonyms for this term. Our search engine issues a Boolean query to the keyword-indexing table and filters out clips whose transcriptions do not contain all the keywords in the question. The remaining clips, called precandidates, are passed to the next phase for further consideration.

A set of heuristics is used in the process of keyword extraction. Some examples are as follows.

- All named entities are considered keywords.
- When a quoted expression occurs in a question or in a sentence, it is recognized as a keyword.
- If two nouns are adjacent, such as "computer network", the whole phrase is treated as a keyword.
- All verbs from a question are selected as keywords, but in a stemmed format. For example, if the verb "computes" occurs in a question, then its stem "compute" is selected as a keyword.

*Phase 2: Question-Clip Matching:* For each precandidate clip identified after the clip filtering phase, LBA measures its relevancy to the question by calculating a similarity score between the question template and clip sentence templates. The similarity score of any sentence within a clip is computed by combining the following factors:

- Matched_Slots_Score (MSS): Compares the slot values of the question template with those of the clip sentence templates. Non-variant term occurrences (exact match) are weighted 2.0, morphological variants (different terms with the same root) are weighted 1.5, and synonyms are weighted 1.0. The total MSS for the sentence "$j$" is computed as follows:

$$MSS_j = \sum_{i=1}^{7} MSS_{ij} \tag{1}$$

  where $MSS_{ij}$ refers to the similarity score of slot "$i$" in sentence "$j$" to the same slot in the current question, except *Answer Type* and *Question Focus* slots.
- Same_WordSequence_Score (SWS): Computes the number of keywords in the question that appear in the same sequence in the current sentence.
- AnswerType_Found_Score (AFS): the original score is 0. If either *Answer Type* or *Question Focus* of the question is found in this sentence, three points are added.

The combined matching score $M_j$ of the sentence $j$ in a clip to a question is defined as follows:

$$M_j = \mathrm{MSS}_j + \mathrm{SWS}_j + \mathrm{AFS}_j. \tag{2}$$

After obtaining combined matching scores for all sentences in the clip transcription using formula (2), we use a slide-window approach to calculate the relevance of this clip to the question. Currently, we set the window size equal to 5. The basic idea is that starting from the first sentence of the clip transcription or lecture notes, we sum the individual combined matching scores of five sentences within the current window, which starts from the first sentence and ends after the fifth sentence, to obtain a window matching score. Next, we move the window toward the end of transcript one sentence. At this point, the new window starts from the second sentence and ends after the sixth. We then obtain the matching score of the current window in the same way. This process (moving the window one sentence at a time) continues until the last sentence of the transcript is reached. Finally, we select the highest window matching score as the relevance score of this clip to the current question. The assumption of this approach is that if a clip can answer a question, there should be a portion somewhere in that clip that shares most content/context with the question. The LBA system can also dynamically generate a brief, query-based text summarization for each candidate clip, including a few high-scoring sentences. This type of summary can provide the context in which query terms occur and highlight the clip content. Finally, the links to those clips are returned to users' local computers in decreasing order of their relevance.

## VI. Evaluation

We conducted a preliminary evaluation of video retrieval using the LBA system. The test-bed was a collection of 468 video clips in the community development domain, which covered many important issues such as fundamental concepts and practical strategies in this field. Those clips were prepared by a domain expert for use in training new employees of community development agencies. No clip was longer than 3 min. The headlines of clips defined by the expert were used to generate question prompts in the evaluation. For example, one of the questions was "How to analyze a financial environment?" The average length of a query was 5.81 (words).

TABLE I
EVALUATION RESULTS

| | Precision | Recall |
|---|---|---|
| **Traditional Search** | 28.4% | 48.5% |
| **Two-phase Approach** | 37.2% | 62.6% |

We submitted 30 natural language questions generated by an expert and examined clips whose relevance scores were larger than a threshold for each question. For each of those questions, there were at least one and at most five relevant clips in the video repository. The retrieval results were analyzed using well-accepted measures for information retrieval effectiveness: precision and recall. Precision refers to the reliability of the clips retrieved, calculated as the number of appropriate clips returned divided by the total number of returned clips. Recall refers to the ratio of the number of relevant clips that were retrieved to the number of relevant clips that should have been retrieved.

We did a comparison test using the same set of 30 questions. First, we used traditional keyword based searching, which relied on a full-text indexing strategy and TF-IDF weighting scheme. Our two-phase retrieval system in LBA was then used. Since our ultimate goal in video searching is to make the list of returned relevant clips as short as possible without missing any of interest, the system was programmed to return only the top ten clips to users if more than ten candidate clips were identified. If there were fewer than ten clips, the system would return all of them. The results are shown in Table I.

The results showed that precision and recall of the two-phase approach were significantly higher than those of the traditional approach $(P < 0.05)$. We also recognize that in order to save users' time, a video retrieval system should be able to report correctly when there is no relevant clip available to a particular question. We therefore created 12 new questions that did not have any corresponding clips in our community development collection. When the system could not find any clips, it would play a "Sorry" video clip prepared in advance saying "Sorry. We cannot find any relevant information for you at this time." We found that our two-phase approach correctly reported no relevant clips available for 10 out of 12 testing questions.

Our future evaluation will be performed on a much larger digital video library. We also had some other discoveries during the evaluation. First of all, although the NLP research has had remarkable achievement in the past decades, it has not reached the anticipated level because of the tremendous complexity and ambiguity of natural language. Second, we believe further semantic analysis of questions and clip transcriptions can improve retrieval performance. We plan to develop a set of grammar rules to identify more detailed semantic relationships between terms within a sentence. Those relationships will be defined based on Fillmore's Case Grammar [45].

## VII. CONCLUSION

Videos have been more and more used in interactive e-Learning. How to make them searchable to satisfy individual needs is an important and challenging task. This paper introduces LBA, an interactive e-Learning environment that allows users to learn by asking questions and retrieving/viewing relevant video clips returned as responses. The system provides people with a sense of being in communication with a mentor in real-time. We have developed a novel two-phase approach to conducting content-based video indexing and retrieval to identify video clips appropriate to addressing users' interests. The approach integrates natural language processing, named entity extraction, text and frame based video indexing and information retrieval techniques. The relevance of video clips to questions is measured based on the similarity between generated templates of questions and clip content. This research explores a new way to access instructional videos in interactive e-Learning. Some preliminary results have shown that this approach achieves higher precision and recall than the traditional keyword-based approach.

## REFERENCES

[1] L. Carswell, "Teaching via the internet: The impact of the internet as a communication medium on distance learning introductory computing students," in *Conf. Integrating Technology into Computer Science Education* , Uppsala, Sweden, 1997, pp. 1–5.

[2] M. R. Syed, "Diminishing the distance in distance education," *IEEE Multimedia*, vol. 8, pp. 18–21, 2001.

[3] S. Carville and D. Mitchell, "It's a bit like star trek': The effectiveness of video conferencing," *Innov. Educ. Training Int.*, vol. 37, pp. 42–49, 2000.

[4] A. Hampapur and R. Jain, "Video data management systems: Metadata and architecture," in *Multimedia Data Management*, W. Klas and A. Sheth, Eds. New York: McGraw-Hill, 1998, ch. 9.

[5] H. Simpson, H. L. Pugh, and S. W. Parchman, *The Use of Videoteletraining to Deliver Hands-On Training: Concept Test and Evaluation*. San Diego, CA: Navy Personal Research and Development Center, 1992.

[6] C. Morales, C. Cory, and D. Bozell, "A comparative efficiency study between a live lecture and a web-based live-switched multi-camera streaming video distance learning instructional unit," in *2001 Information Resources Management Assoc. Int. Conf.*, Toronto, ON, Canada, 2001.

[7] R. Burke and A. Kass, "Supporting learning through active retrieval of video stories," *Expert Syst. Applicat.*, vol. 9, pp. 361–378, 1995.

[8] D. Marinelli and S. Stevens, "Synthetic interviews: The art of creating a "Dyad" between humans and machine-based characters," in *Interactive Voice Technology for Telecommunications Applications'98*, 1998, pp. 43–48.

[9] M. d. G. Pimentel, Y. Ishiguro, G. D. Abowd, B. Kerimbaev, and M. Guzdial, "Supporting educational activities through dynamic web interfaces," *Interact. Comput.*, vol. 13, pp. 353–374, 2001.

[10] A. Amir, G. Ashour, and S. Srinivasan, "Toward automatic real time preparation of online video proceedings for conference talks and presentations," in *34th Hawaii Int. Conf. System Sciences*, Maui, HI, 2001, pp. 1662–1669.

[11] H. S. Chang, S. Sull, and S. U. Lee, "Efficient video indexing scheme for content-based retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 1269–1279, Dec. 1999.

[12] N. Dimitrova, T. McGee, and H. Elenbass, "Video keyframe extraction and filtering: A keyframe is not a keyframe to everyone," in *Proc. ACM Conf. Information and Knowledge Management*, Las Vegas, NV, Nov. 10–14, 1999, pp. 113–120.

[13] E. Ardizzone and M. Cascia, "Automatic video database indexing and retrieval," *Multimedia Tools Applicat.*, vol. 4, pp. 29–56, 1997.

[14] D. Papadias, M. Mantzourogiannis, P. Kalnis, N. Mamoulis, and I. Ahmad, "Content-based retrieval using heuristic search," in *22nd Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval* , Berkeley, CA, 1999, pp. 168–175.

[15] D. Zhong and S. F. Chang, "Video object model and segmentation for content-based video indexing," in *1997 IEEE Int. Symp. Circuits and Systems* , Hong Kong, 1997, pp. 1492–1495.

[16] S.-Y. Kim and Y. M. Ro, "Fast content-based MPEG video indexing using object motion histogram," in *IEEE Region 10 Conf. (TENCON 99)* , Cheju Island, Korea, 1999, pp. 1506–1509.

[17] B. Gunsel, A. M. Tekalp, and P. J. L. Van Beek, "Object-based video indexing for virtual-studio production," in *IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition* , San Juan, Puerto Rico, 1997, pp. 769–774.

[18] S. Dagtas, W. Al-khatib, A. Ghafoor, and R. L. Kashyap, "Models for motion-based video indexing and retrieval," *IEEE Trans. Image Processing*, vol. 9, pp. 88–101, Jan. 2000.

[19] S. W. Smoliar and H. Zhang, "Content based video indexing and retrieval," *IEEE Multimedia*, vol. 1, pp. 62–72, 1994.

[20] H. D. Wactlar, M. G. Christel, Y. Gong, and A. G. Hauptmann, "Lessons learned from building a terabyte digital video library," *IEEE Computer*, vol. 32, pp. 66–73, 1999.

[21] Y.-L. Chang, W. Zeng, I. Kamel, and R. Alonso, "Integrated image and speech analysis for content-based video indexing," in *Third IEEE Int. Multimedia Computing and Systems* , Hiroshima, Japan, 1996, pp. 306–313.

[22] S. Tsekeridou and I. Pitas, "Content-based video parsing and indexing based on audio-visual interaction," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 11, pp. 522–535, Apr. 2001.

[23] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, vol. 3, pp. 27–36, 1996.

[24] M. Brown, J. Foote, G. Jones, K. Sparck-Jones, and S. Young, "Open-Vocabulary speech indexing for voice and video mail retrieval," in *Fourth ACM Int. Multimedia Conf.* , Boston, MA, 1996, pp. 307–316.

[25] W. Li, S. Gauch, J. Gauch, and K. M. Pua, "Vision: A digital video library," in *1st ACM Int. Conf. Digital Libraries* , Bethesda, MD, 1996, pp. 19–27.

[26] H. D. Wactlar, A. G. Hauptmann, M. G. Christel, R. A. Houghton, and A. M. Olligschlaeger, "Complementary video and audio analysis for broadcast news archives," *Commun. ACM*, vol. 43, pp. 42–47, 2000.

[27] R. Lienhart, "Automatic text recognition for video indexing," in *Fourth ACM Int. Conf. Multimedia* , Boston, MA, 1996, pp. 11–20.

[28] Y. Kuwano, H. A. Taniguchi, S. Mori, and H. K. Kurakake, "Telop-on-demand: Video structuring and retrieval based on text recognition," in *2000 IEEE Int. Conf. Multimedia and Expo (ICME 2000) United States*, New York, 2000, pp. 759–762.

[29] R. Burke, "Conceptual indexing and active retrieval of video for interactive learning environments," *Knowl.-Based Syst.*, vol. 9, pp. 491–499, 1996.

[30] C. Johnson, L. Birnbaum, R. Bareiss, and T. Hinrichs, "War stories: Harnessing organizational memories to support task performance," *Intelligence*, pp. 17–31, 2000.

[31] H. Jiang and A. K. Elmagarmid, "WVTDB—A semantic content-based video database system on the world wide web," *IEEE Trans. Knowl. Data Eng.*, vol. 10, pp. 947–966, Nov./Dec. 1998.

[32] O. Javed, S. Khan, Z. Rasheed, and M. Shah, "A framework for segmentation of interview videos," in *IASTED Int. Conf. Internet and Multimedia Systems and Applications United States*, Las Vegas, NV, 2000.

[33] S. Boykin and A. Merlino, "Machine learning of event segmentation for news on demand," *Commun. ACM*, vol. 43, pp. 35–41, 2000.

[34] P. Chiu, A. Girgensohn, W. Polak, E. Rieffel, and L. Wilcox, "A genetic algorithm for video segmentation and summarization," in *2000 IEEE Int. Conf. Multimedia and Expo* , vol. 3, New York, 2000, pp. 1329–1332.

[35] E. M. Voorhees and D. M. Tice, "The TREC-8 question answering track evaluation," in *Eighth Text REtrieval Conference (TREC-8)* , Gaithersburg, MD, 1999.

[36] K. C. Litkowski, "Syntactic clues and lexical resources in question-answering," in *Ninth Text REtrieval Conference (TREC-9)* , Gaithersburg, MD, 2000, pp. 83–106.

[37] R. Burke, K. Hammond, V. Kulyukin, S. Lytinen, N. Tomuro, and S. Schoenberg, "Question answering from frequently-asked question files: Experiences with the FAQ finder system," *AI Mag.*, vol. 18, pp. 57–66, 1997.

[38] H. D. Wactlar. Informedia—Search and summarization in the video medium. presented at Imagina 2000 Conf.. [Online] Available http://www.informedia.cs.cmu.edu/documents/imagina2000.pdf

[39] N. Chinchor. MUC-7 named entity task definition. presented at Seventh Message Understanding Conf. (MUC-7). [Online] Available http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html

[40] R. Grishman, "The NYU system for MUC-6 or where's the syntax?," in *Sixth Message Understanding Conf. (MUC-6)* , Columbia, MD, 1995, pp. 167–175.

[41] D. M. Bikel, R. Schwartz, and R. M. Weischedel, "An algorithm that learns what's in a name," *Mach. Learn.*, vol. 34, pp. 211–231, 1999.

[42] A. Voutilainen, "Helsinki taggers and parsers for english," in *Corpora Galore: Analysis and Techniques in Describing English* , J. M. Kirk, Ed. Amsterdam, The Netherlands/Atlanda, GA: Rodopi, 2000.

[43] R. Srihari and W. Li, "Information extraction supported question answering," in *Eighth Text REtrieval Conf. (TREC-8)* , Gaithersburg, MD, 1998, pp. 185–196.

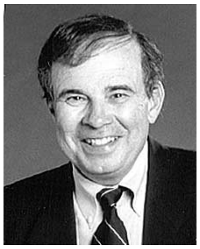
[44] G. Salton, *The SMART Retrieval System*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[45] W. A. Cook, *Case Grammar Theory*. Washington, D.C.: Georgetown Univ. Press, 1989.

**Dongsong Zhang** received the Ph.D. degree in management from the College of Business and Public Administration, University of Arizona, Tucson, in 2002.

He is an Assistant Professor with the Department of Information Systems, University of Maryland-Baltimore County, Baltimore. His primary research interests include interactive e-Learning, intelligent information systems, knowledge discovery, and computer-supported collaboration and communication. His work has been published or will appear in the *Communications of the ACM* (CACM), IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, *Information Systems Frontier*, and *Communications of AIS*, among others.

Dr. Zhang is a member of ACM.



**Jay F. Nunamaker** is Regents and Soldwedel Professor of MIS, computer science and communication, and Director of the Center for the Management of Information, University of Arizona, Tucson. Under his leadership, the MIS Department achieved national recognition as a top five ranked MIS department. He was featured in the July 1997 *Forbes Magazine* issue on technology as one of eight innovators in information tecnology. He is is known for his research in collaboration systems and knowledge management. He specializes in group decision-making and deliberation, automation of systems development, databases, expert systems, and systems analysis and design. His information and laboratories can be found on Navy ships, in third-world countries, in corporate businesses throughout the world, and in the White House.

Dr. Nunamaker was elected as a Fellow of the Association of Information Systems in 1999 and in 1998 was recognized as one of the top four most productive MIS researchers over five years.