APPLYING SEMANTIC ASSOCIATION TO SUPPORT CONTENT-BASED VIDEO RETRIEVAL

Yueting Zhuang†∗, Yong Rui, Thomas S. Huang, Sharad Mehrotra

†Department of Computer Science
Zhejiang University, Hangzhou, 310027, P.R.China
Beckman Institute and Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
E-mail:{yzhuang,yrui,huang}@ifp.uiuc.edu and sharad@cs.uiuc.edu

ABSTRACT

The traditional approach to video retrieval is to first annotate the video by textual information (titles and key words) and then the queries will be searched based on this keyword set. Since automatic annotation has not yet been available, this work needs great amount of labor and has been proved to be unrealistic in applications. Another approach, which seems to be at the other extreme, is to utilize the low-level video content, such as color, texture, shape, motion features and so on, in an attempt to get rid of the need of key words annotation.

We hold the view in this paper that a user preferable query form should include both the keywords and video contents. In this paper, we will explore the semantic aspect based on video TOC structuring [1]. Close-captioning is used to extract a basic keywords set. *WordNet*, an electronic lexical system, is used to provide semantic association. The approach has been applied in Web-MARS VIR and the running result has shown that the retrieval performance is greatly improved.

1. INTRODUCTION

More and more videos are emerging on the Web. Internet has become a huge reservoir of video. How to retrieve the video efficiently is a key issue in the database society as well as in information retrieval area.

The traditional approach is to first annotate video by textual information and consequently queries are made based on the set of keywords and even go further, annotate object along with their various relations [2, 3]. For example, in [3], the system manually annotated detailed information from video, thus the system can deal with very complicated queries, such as "*find all people who appear in frames in which Gene Kelly and Giner Rogers are getting married*". Apparently it is impractical to do the annotation manually since there are huge amount of video data existing on the web. On the other hand, even to those who are experienced in permuting keywords to locate the documents in a database, the frustration is still obvious when confronted with a query like "*find me a clip of two seconds in which a red car racing along a hillside road on a bright day disappears as the road bends around the hill*" [4]. In cases such as aerial surveillance situation, this *fully human assisted database schemes* approach is not realistic.

An alternative approach is to ignore the semantic meaning, and answer queries totally based on the image feature content, such as color, texture, layout, aiming towards getting rid of the time-consuming annotation process which can not avoid subjectivity [5]. There have been many research projects and systems in this area.

In this paper, we hold the viewpoint that a user's preferable query will contain both the visual content and subjective keywords that is to be matched with the database. One way to annotate the basic keyword set from a video is to use the close-captions if available. Our emphasis in this paper is the application of semantic association to extend the content-based retrieval performance of the video database. A tradeoff is made between fully human assisted database schemes and fully image content based methods.

This work was supported in part by the Army Research Laboratory under Cooperative Agreement No. DAAL01-96-2-0003 and in part by NSF/DARPA/NASA Digital Library Initiative Program under Cooperative Agreement No. 94-11318. ∗ Also in part by 863 High-Tech No. 863-306-04-03-3 and NSF of China.

2. OVERALL SYSTEM STRUCTURE

Because of the large amount of video information, neither manual indexing nor manual annotation, although accurate in content and semantic meaning, is possible. So the philosophy behind our work is to maximize the procedure that can be automatically conducted.

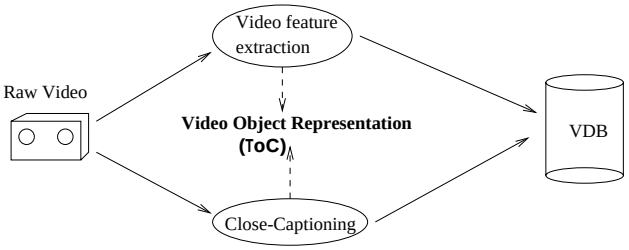

Figure 1: Integrated video extraction model in Web-MARS VIR

Figure 1 shows the video extraction structure in Web-MARS VIR.[1] Video content, such as shot boundary, key frames, are extracted and populated into the video database (VDB) automatically. For those close-captioned videos, semantic content(text, keywords) is also acquired and stored into the VDB automatically.

The video object representation is the key issue. In the next section, we will provide a hierarchical video representation structure called *ToC*, where slots are provided to accommodate the semantics.

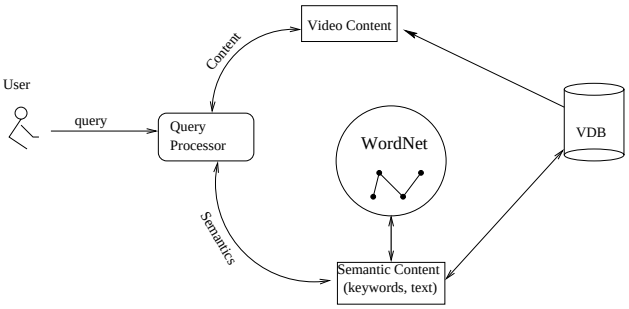

Figure 2: Query Processing Model in Web-MARS VIR

Figure 2 shows the query processing model in which both the video content and semantics are used.

Consider the following scenario. Suppose a user presents a query:

show me a shot that has as much action as this one?(Given an example image at the same time)

The system will match the word "action" in the database. If it is included in VDB, the related video shots are returned. Otherwise, trigger *WordNet* to find out the *synset* and then query into the VDB again. In this case, "action" can mean either "military action" or "human activity". Since WordNet has the property that every noun in it is included in a single tree, we can always trace "action" back to the root node. The semantics of the nodes along the way is compared with that in the database. The result is convergent.

[1]MARS is *Multimedia Analysis and Retrieval System* developed at UIUC[6] and is being extended to the web. Web-MARS VIR is the Video Information Retrieval part of the system.

3. VIDEO OBJECT REPRESENTATION— TABLE-OF-CONTENT(TOC)

In order to index and browse through video, the first important thing to do is to find an efficient way to represent the video. One of the popular existing approaches to representing video contents is the structural modeling approach[7].

How does a reader efficiently access a 1000-page book's content? Without reading the whole book, he will probably first go to the book's Table-of-Content (ToC), finding which chapters or sections meet his need. If he has specific questions (queries) in mind, such as finding a term or a key word, he will go to the index page and find the corresponding book sections containing that question. In short, a book's ToC helps a reader *browse* and a book's index helps a reader *retrieve(search)*. The former is useful when the reader does not have any specific question in mind and will make his information need more specific and concrete via browsing the ToC. The latter is useful when the reader has a specific information requirement. Both aspects are equally important in helping users access the book's content. For current videos, unfortunately, we lack both the ToC and the index. Techniques are needed for constructing ToC and index to facilitate the video access.

The video stream can be structured into a hierarchy consisting five levels: video, scene, group, shot, and key frame, from top to bottom increasing in granularity [1].

In order to support not only content-based browsing but also content-based retrieval, we need to incorporate semantics in the ToC structure. Basically, the way how the ToC is related to semantics is shown in figure 3.

Some entities, e.g. key words, are associated with all of its counterparts, while others, e.g. objects, have a defined life cycle. The link weights are real numbers within [0,1], indicating how strong the two entities' link is. For example, if shot 1 of video A has a 0.9 link weight to key word "dog", it indicates that "dog" is an important content in that shot. The link weights enable the viewer to go "back and forth" between the ToC and index. Each round of such a "back and forth" helps the viewer to locate the information of interest

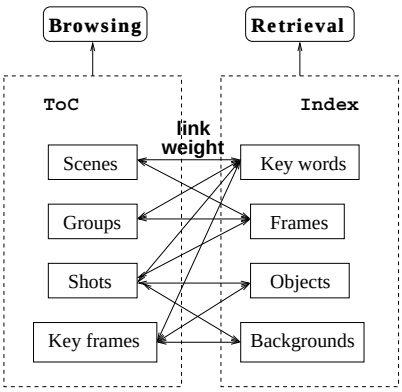

Figure 3: Exploring semantics for each part in video structure

more precisely.

4. VIDEO SEMANTICS AND KNOWLEDGE

A question arises as to how to acquire the semantics of video efficiently and automatically?

There are various ways of finding the link weights between the entities. For example, to associate key words to shots, the following procedure is performed:

4.1. Close-captioning

The semantics content used to assist the content-based retrieval is extracted from the closed-caption signal that was originally intended for the consumption of the hearing impaired viewers. Closed captions are incorporated into the video program by an encoded composite data signal during line 21 of field one of the standard NTSC video signal. The text contains program related material, and in many cases, is an exact representation of the speech contained in the audio portion of the video program[8]. When it is decoded, each frame contains control characters, or up to two alphanumeric characters. These characters over several frames will generate words or sentences. The control characters is used to determine the attributes of the text such as color, font, indent, and location on the screen and will be used by the closed-caption decoder.

In our system, we first digitize the video using *Broadway* for Windows and transcribe the corresponding close-caption text using SunBelt Inc.'s *TextGrabber*. Then we synchronize the video and close-caption by time-stamps.

The following is a sample close caption text extracted from *Independance Day*.

```
. . . . . .
It's so fuzzy.
    [Horn honks]
 [horn honks]
 oh, no.
    Good morning, lucas.
 You see these? I got  a whole god damn
    crop full of these.
 If your father's  not in the air in 20
     minutes...
 ......
 [Whispers]
 all right, go ahead.  Put it on.
 General, you might  want to watch this.
    Tv: Ladies and gentlemen...
 ......
```

Table 1: Content of close-captioning result

Events	Start frame	End frame	Keyword
horn honks	12	200	Lucas ...
whispers	2201	2430	go ahead ...

Table 1 shows the mapping from close-captioned text to the physical position(start frame, end frame) of the video clips.

4.2. Parsing of transcribed text

The link weight of a shot and a key word is calculated as: $lw = tf \times idf$, where tf and idf stand for *term frequency* and *inverse document frequency* for that key word [9].

For each shot, we extract its corresponding transcribed text and parse the text information by using a key word extractor *AZTagger*.

4.3. WordNet

WordNet is an electronic lexical system developed by George Miller and his colleagues at Princeton University [10]. The noun portion of WordNet is designed around the concept of *synset* which is a set of closely related synonyms representing a word sense(meaning). Every word that is in the WordNet has one or more senses and for each sense it has a distinct set of synonyms, and a distinct set of words related through other relationships such as hypernyms / hyponyms (IS_A relation), holonyms (MEMBER_OF relation) and meronyms (PART_OF relation).

Several noval features of WordNet is as follows:

- Distance between Concept

WordNet provides IS_A relation between concepts, an important feature in measuring the distance of concept. For example, "man" and "gentleman" is linked by IS_A. If the transcribed text in the database is "man", while the term in user's query is "gentleman", WordNet provides such kind of semantics association.

- Causal Relation Between Actions

  WordNet also provides "Cause-to" relation between actions(verb) which can be used in the explanation of actions. For example: "kill" causes "die".

- Synonyms between Adjectives

  For example, similar words to "beautiful" are "attractive", "charming", "fine-looking", "pretty" and so on. WordNet provides synonyms between adjectives.

5. EXPERIMENT SETUP

We have incorporated the above principles in the Web-MARS VIR. Informix Universal Server version 9.12 for UNIX is used as the DBMS for video storage and indexing. The data source of the video database is designed to be from web sites whose URLs are reported by a web-crawler. Currently, for the testing of VIR, we have loaded 20 video clips and later videos will be added from sources given by web-crawler. Figure 4 shows the result of query "gentleman and lady". Note that in the database only "man" and "woman", which come from close-captioning, are stored. Using of WordNet causes the flexible result.

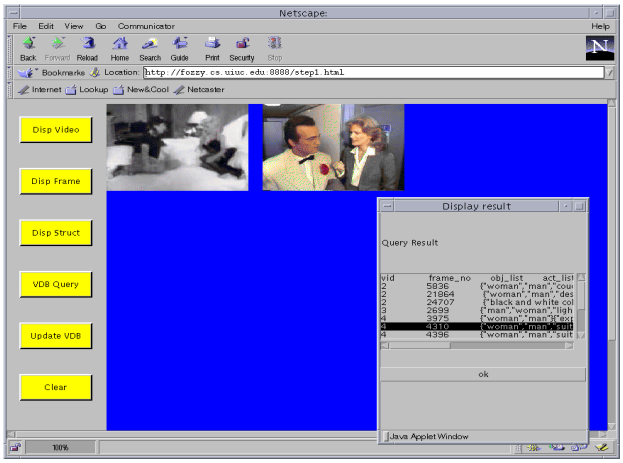

Figure 4: An example of query result

6. CONCLUSION

In this paper, we have explored the semantics association of video based on ToC structuring. Close-captioning is used to extract the basic set of keywords. Then WordNet is applied to improve the retrieval performance. In the end, experiment result has shown its effectiveness.

7. REFERENCES

[1] Y. Rui, T. S. Huang, and S. Mehrotra, "Exploring video structures beyond the shots," in *Proc. of IEEE conf. Multimedia Computing and Systems*, 1998.

[2] S. Chang and E. Jungert, "Pictorial data management based upon the theory of symbolic projections," *Journal of Visual Languages and Computations*, no. 2, pp. 195–215, 1991.

[3] S. Adali, K. Candan, S.-S. Chen, K. Erol, and V. Subrahmanian, "Advanced video information system: Data structures and query processing," *To appear in ACM-Springer Multimedia Systems Journal*.

[4] A. Gupta and R. Jain, "Visual information retrieval," *Communications of the ACM*, vol. 40, no. 5, 1997.

[5] W. Niblack, R. Barber, and et al., "The QBIC project: Querying images by content using color, texture and shape," in *Proc. SPIE Storage and Retrieval for Image and Video Databases*, Feb 1994.

[6] T. S. Huang, S. Mehrotra, and K. Ramchandran, "Multimedia analysis and retrieval system (MARS) project," in *Proc of 33rd Annual Clinic on Library Application of Data Processing - Digital Image Access and Retrieval*, 1996.

[7] B. Rubin and G. Davenport, "Structured content modeling for cinematic information," *SIGCHI Bull*, vol. 21, no. 2, pp. 78 –79, 1989.

[8] B. Shahraray and D. C. Gibbon, "Automated authoring of hypermedia documents of video programs," in *ACM Multimedia 95*, 1995.

[9] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1983.

[10] G.A.Miller, R.Beckwith, C.Fellbaum, D. Gross, and K. Miller, "Introduction to WordNet: An on-line lexical database," *International Journal of Lexicography*, vol. 3(4), pp. 235–244, 1990.