BROWSING AND RETRIEVING VIDEO CONTENT IN A UNIFIED FRAMEWORK

Yong Rui, Thomas S. Huang and Sharad Mehrotra
Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

Abstract - In this paper, we first review the recent research progress in video analysis, representation, browsing, and retrieval. Motivated by the mechanism used to access book's content, we then present novel techniques for constructing video Table-of-Contents and index to facilitate accessing video's content. We further explore the relationship between video browsing and retrieval and propose a unified framework to incorporate both entities in a seamless way. Preliminary research results justify our proposed framework for providing access to videos based on their content.

INTRODUCTION

Research on how to efficiently access the video content has become increasingly active in the past few years [9, 1, 10, 4]. Considerable progress has been made in video analysis, representation, browsing, and retrieval, the four fundamental bases for accessing video content. *Video analysis* deals with the *signal processing* part of the video system, including shot boundary detection, key frame extraction, etc. *Video representation* concerns with the *structure* of the video. An examples of the video representations is the tree structured key frame hierarchy [8, 10]. Build on top of the video representation, *video browsing* deals with how to use the representation structure to help the viewers browse the video content. Finally, *video retrieval* concerns about retrieving interesting video objects to the viewer. The four research areas' relationship is illustrated in Figure 1.

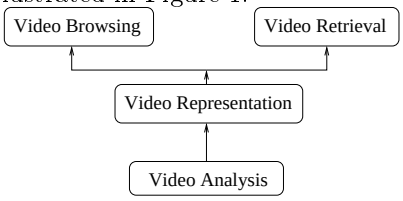

Figure 1: Relations between the four research areas

So far, most of the research effort has gone into video analysis. Though it is the basis for all the other research activities, it is not the ultimate goal. Relatively less research exists on video representation, browsing and retrieval.

From Figure 1, video browsing and retrieval are on the very top of the diagram. They *directly* support users' access to the video content. Both the browsing and retrieval are equally important. An analogy explains this argument. 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-Contents (ToC), finding which chapters or sections suit his need. If he has specific questions (queries) in mind, such as finding a terminology or a key word, he will go to the index page and find the corresponding book sessions containing that question. In short, book's ToC helps a reader *browse* and book's index helps a reader *retrieve*. 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.

What is even more important in video domain is that the ToC and index should be inter-related. For a continuous long medium type like video, such "back and forth" mechanism between browsing and retrieval is crucial. The video library users may have to browse the video first before they know what to retrieve. On the other hand, after retrieving some video objects, it will guide the users to browse the video in the correct direction.

The goal of this paper is to explore novel techniques for constructing both the video ToC and video index and integrate them into a unified framework. The rest of the paper is organized as follows. In section 2, important video terminologies are first introduced. Video analysis is then reviewed and discussed in section 3. In section 4, we describe various video representations. Build on top of section 4, we review video browsing and retrieval techniques in section 5. Our proposed unified framework for video browsing and retrieval is presented in section 6. Conclusions are given in section 7.

TERMINOLOGIES

Video shot: is an unbroken sequence of frames recorded from a single camera. It is the building block of video streams. *Key frame*: is the frame which can represent the salient content of a shot. Depending on the content complexity of the shot, one or more key frames can be extracted. *Video scene*: is defined as a collection of semantically related and temporally adjacent shots, depicting and conveying a high-level concept or story. While shots are marked by physical boundaries, scenes are marked by semantic boundaries [1]. In summary, the video stream can be structured into a hierarchy consisting

[1] Some of the early literatures in video parsing misused the phrase *scene change detection* for *shot boundary detection*. To avoid any later confusion, we will use *shot boundary detection* for the detection of physical shot boundaries while using *scene boundary detection* for the detection of semantic scene boundaries.

five levels: video, scene, shot, and key frame, from top to bottom increasing in granularity [4].

VIDEO ANALYSIS

Shot boundary detection

It is beneficial to first decompose the video clip into shots before any processing is done. In general, automatic shot boundary detection techniques can be classified into five categories [2], i.e. *pixel based*, *statistics based*, *transform based*, *feature based*, and *histogram based*. So far, the histogram based approach is the most popular approach used in shot boundary detection. Several researchers claim that it achieves good trade-off between accuracy and speed [9].

Key frame extraction

After the shot boundaries are detected, corresponding key frames can then be extracted. Simple approaches may just extract the first and last frames of each shot as the key frames. More sophisticated extraction techniques can be based on visual content indicator [11] and shot motion indicator [7].

VIDEO REPRESENTATION

Considering that each video frame is a 2D object and the temporal axis makes up the third dimension, a video stream spans a 3D space. Video representation is the mapping from the 3D space to the 2D view screen.

Sequential Key Frame Representation

After obtaining shots and key frames, an obvious and simple video representation is to sequentially layout the key frames of the video, from top to bottom and from left to right. This simple technique works well only when the number of key frames is not too many.

Scene Based Representation

To provide the user with better access to the video, constructing a video representation at a semantic level is needed [4, 1]. In [1], a scene transition graph (STG) of video representation is proposed and constructed. Video sequence is first segmented into shots. Shots are then clustered by using *time-constrained clustering*. STG is then constructed based on the time flow of clusters.

Video Mosaic Representation

Instead of representing the video structure based on the video-scene-shot-frame hierarchy as discussed above, this approach takes a different perspective [3]. The mixed information within a shot is decomposed into three components: extended spatial information, extended temporal information, Geometric information [3].

VIDEO BROWSING AND RETRIEVAL

These two functionalities are the ultimate goals of a video access system, and they are closely related to the previous section's video representation. The first 3 representations are suitable for video browsing while the last representation could be used in video retrieval.

For "Sequential Key Frame Representation", the browsing is obviously a sequential browsing, scanning from the top-left key frame to the bottom-right key frame. For STG representation, a major characteristic is its indication of time flow embedded within the representation. By following the time flow, the viewer can browsing through the video clip.

Unlike the other video representations, the mosaic representation is especially suitable for video retrieval. The three components, moving objects, backgrounds, and camera motions, are perfect candidates for video index. After constructing such a video index, queries such as "find me a car moving like this", "find me a conference room having that environment", etc. can be effectively supported.

A UNIFIED FRAMEWORK

As we have reviewed in the previous sections, considerable progress has been made in each of the areas of video analysis, representation, browsing, and retrieval. However, so far, the interaction among these components is still limited and we still lack a unified framework to glue them together. This is especially crucial, given the characteristics of the video media type: long and unstructured. In our lab, we have been conducting research to explore the synergy between browsing and retrieval.

Video Browsing

Of the video representations for browsing, "Scene Based Representation" is the most effective one [4, 1]. We have proposed a scene based video ToC representation in [4]. In this representation, a video clip is structured into the scene-shot-frame hierarchy, based on which the ToC is constructed. This ToC frees the viewer from doing tedious "fast forward" and "rewind", and provides the viewer with non-linear access to the video content.

Video Retrieval

Constructing index for videos is far more complex than constructing index for books. For books, the index is normally based on key words or terms that readers will be interested in. For videos, the viewer's interests may cover a wide range. All of the followings are good candidates for video index: Keywords, Frames, Objects and backgrounds.

Frames, objects, and backgrounds are visual entities which the viewers may be interested in. In addition, conventional entities, such as key words, are also important. Increasingly, researchers are realizing that the visual information alone is not enough to support effective retrieval. The conventional key words,

together with the visual entities, will supplement to each other and constitute an effective retrieval system. Our previous work in image retrieval [5] supports frame-based retrieval. We are currently implementing our key word based retrieval by analyzing the video close-caption; and object and background based retrieval based on the Mosaic Representation [3].

A Unified Framework

The above two subsections described our video browsing and retrieval techniques separately. In this section, we will integrate them into a unified framework. For browsing, we have entities like scenes, shots, and key frames; and for retrieval, we have entities such as key words, frames, objects, and backgrounds. They co-exist for their own purposes (to support browsing and retrieval). Exploring further, we will find that they are inherently related to each other. An object has a life cycle within a shot, and a shot's content is captured by its objects and backgrounds. Figure 2 illustrates the unified framework.

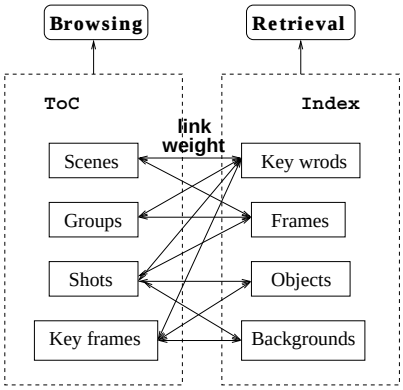

Figure 2: A unified framework

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 more precisely.

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:

- Digitize the video (using *Broadway* for Windows) and transcribe the corresponding close-caption text (using SunBelt Inc.'s *TextGrabber*).
- Synchronize the video and close-caption by time stamps.
- For each shot, extract its corresponding transcribed text.

- Parse the text information by using a key word extractor *AZTagger*.
- The link weight of a shot and a key word is: $lw = tf \times idf$, where tf and idf stand for *term frequency* and *inverse document frequency* for that key word [6].

CONCLUSIONS

This paper introduced video ToC and index and presented techniques for constructing them. It also proposed a unified framework for video browsing and retrieval; thus providing video viewer better mechanism to access the video content.

ACKNOWLEDGMENTS

This work was supported in part by ARL Cooperative Agreement No. DAAL01-96-2-0003 and in part by a CSE Fellowship of University of Illinois.

References

[1] Ruud M. Bolle, Boon-Lock Yeo, and Minerva M. Yeung. Video query: Beyond the keywords. Technical report, IBM Research Report, Oct 17 1996.

[2] John S. Boreczky and Lawrence A. Rowe. Comparison of video shot boundary detection techniques. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*, 1996.

[3] Michal Irani and P. Anandan. Video indexing based on mosaic representations. *Proceedings of The IEEE*, 86(5), 1998.

[4] Yong Rui, Thomas S. Huang, and Sharad Mehrotra. Exploring video structures beyond the shots. In *Proc. of IEEE conf. Multimedia Computing and Systems*, 1998.

[5] Yong Rui, Thomas S. Huang, Michael Ortega, and Sharad Mehrotra. Relevance feedback: A power tool in interactive content-based image retrieval. *IEEE Tran on Circuits and Systems for Video Technology, Special Issue on Interactive Multimedia Systems for the Internet*, Sept 1998.

[6] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1983.

[7] Wayne Wolf. Key frame selection by motion analysis. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, 1996.

[8] H. Zhang, S. W. Smoliar, and J. J. Wu. Content-based video browsing tools. In *Porc. IS&T/SPIE Conf. on Multimedia Computing and Networking*, 1995.

[9] HongJiang Zhang, A. Kankanhalli, and S. W. Smoliar. Automatic partitioning of full-motion video. *ACM Multimedia Systems*, 1(1), 1993.

[10] Di Zhong, HongJiang Zhang, and Shih-Fu Chang. Clustering methods for video browsing and annotation. Technical report, Columbia Univ., 1997.

[11] Yueting Zhuang, Yong Rui, Thomas S. Huang, and Sharad Mehrotra. Adaptive key frame extraction using unsupervised clustering. In *Proc. IEEE Int. Conf. on Image Proc.*, 1998.