ADVISE: Advanced Digital Video Information Segmentation Engine

Chung Wing Ng and Michael R. Lyu

Department of Computer Science and Engineering

The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR.

{cwng, lyu}@cse.cuhk.edu.hk

ABSTRACT

This paper describes the design of ADVISE, Advanced Digital Video Information Segmentation Engine, which is a web-based video retrieval system. The system aims at providing a visual summarization of video contents, such that users can efficiently determine whether they are interested in the video before they have downloaded it from the Internet. ADVISE consists of three major modules. The first module is responsible for automatic structuring of videos. The second module stores the structure as a Video Table-Of-Contents (VTOC) in XML format, and presents the VTOC on the Web with XSL. The third module provides users options to personalize a video into its summary using SMIL. The system architecture and some implementation screen shots are presented.

Keywords

Video Structuring, Video Summarization, XML, SMIL.

1. INTRODUCTION

Video over Internet is getting more popular now than ever before, due to the rapid growth of Internet bandwidth and the growing use of video in education, entertainment, and information sharing. Among the vast video sources, it is more efficient for users to search for their desired pieces if descriptions of video are provided. A meaningful video description can help users to know the contents at once, so they do not need to waste time on downloading huge clips that they may not be interested. As a result, video descriptions can enhance efficient browsing and retrieval of video contents. Textual description extracted from video caption text is a commonly used solution, however, text may not always well describe the video as the contents are delivered by combining visual, audio, and textual information. Therefore, we propose the system ADVISE, Advanced Digital Video Information Segmentation Engine, which generates visual video descriptions and summarizes videos into a short length, such that it can solve the above problem.

The ADVISE system is built on the Web. There are two ways to help the Internet users to efficiently browse the contents of the shared videos. In the first one, a Video Table-Of-Contents (VTOC) is generated to describe the shared video. The VTOC functions like the table-of-contents in a book by showing the video contents in a structured format. Representative images extracted from the video according to the structure are embedded in the VTOC, such that they can demonstrate what have been shown in the video, and how they synthesize the video contents. After browsing the VTOC, users should then have an abstract idea of the video contents. To further describe a video, the second method provided in ADVISE is the generation of a video summary. Since the image-based VTOC is a static presentation, it may not be sufficient to describe a video. Therefore ADVISE generates the summarized video by scanning through the video segments according to the structure of the VTOC within a short duration. Several options are allowed for the users to customize the video summary, and SMIL is used for the summary presentation. The system architecture is briefly described in the following section.

2. SYSTEM ARCHITECTURE

There are three major modules in the ADVISE system. They are Video Structuring module, VTOC presentation module, and SMIL generation module. Figure 1 shows the system architecture of ADVISE.

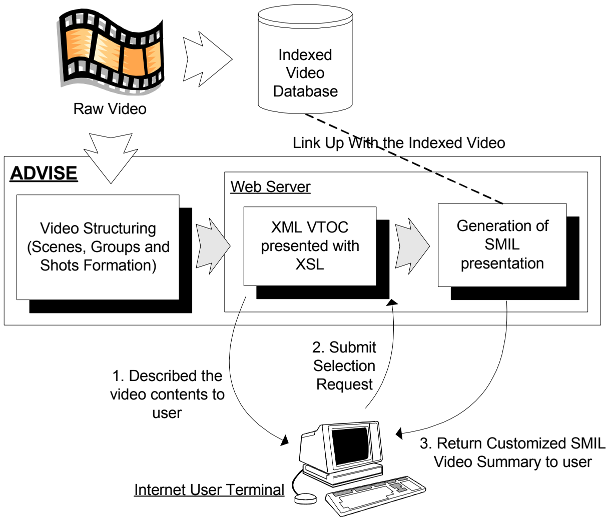

Raw Video
Indexed Video Database
Link Up With the Indexed Video
ADVISE
Web Server
Video Structuring (Scenes, Groups and Shots Formation)
XML VTOC presented with XSL
Generation of SMIL presentation
1. Described the video contents to user
2. Submit Selection Request
3. Return Customized SMIL Video Summary to user
Internet User Terminal

Figure 1: System Architecture of ADVISE

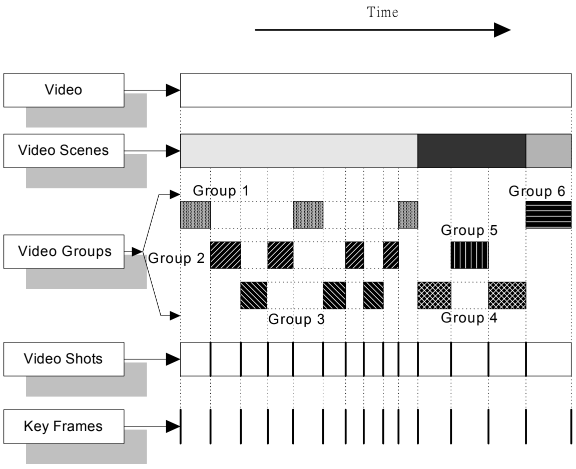


Figure 2: Hierarchy of Video Components

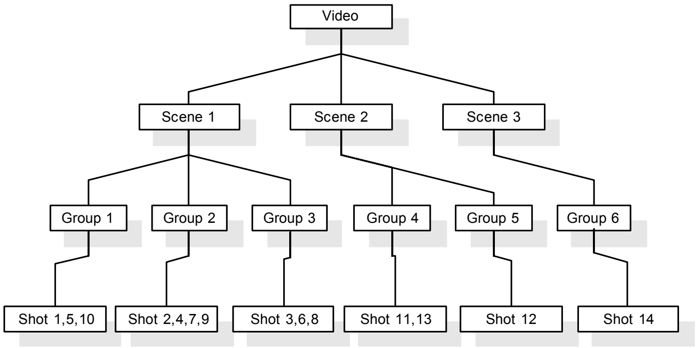


Figure 3: Video Tree Structure

2.1 Video Structuring Module

In this module, a video is structured according to its contents. We divide a video into five levels of components [3]. They are video key frames, video shots, video groups, video scenes, and the whole video. The hierarchy of these components in a video is shown in Figure 2. The video components are further organized into a four-level tree structure as shown in Figure 3. The video tree structure can well demonstrate the organization of the video contents.

In the video structuring process, different video features, for example, audio energy, color histograms, and video caption text, can be used to generate the video components. Our current implementation employs the color histograms feature because it can be calculated from video frames efficiently. We use weighted regional color histograms and an adaptive threshold approach to improve the accuracy of the video structuring process. The tree structure generated by the color histograms approach is already quite adequate, but we can also use multiple video features to further improve the structure.

2.2 VTOC Presentation Module

In this module, VTOC is generated according to the video tree structure from the previous module. The structure is stored in the XML format [4]. We define a set of elements in Figure 4 to describe the tree structure. A sample XML video tree structure is shown in Figure 5.

XML elements	Description	Child nodes	Associated attributes
<advise> … </advise>	encapsulate the whole structure	video	-
<video> … </video>	root level component of video	scene	src
<scene> … </scene>	video scene component	group	id
<group> … </group>	video group component	shot	id
<shot> … </shot>	video shot component	time, keyframe	id
<time />	store the time of a shot	-	value
<keyframe />	point to the key frame image	-	img

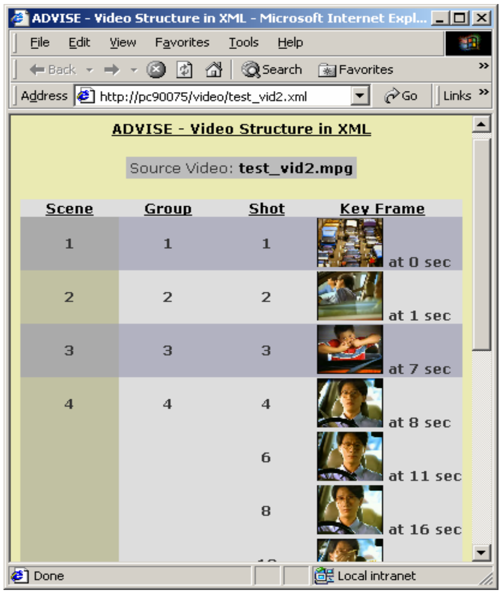
Figure 4: Set of XML elements defined

We use XML for storage because of several advantages. First, we can build an organized and compact data structure for using the nested hierarchy of XML. Also, XML is extensible and searchable as it is in a plain-text format. Besides, XML can be used as a standard information protocol for exchanging data between different system modules.

It is also convenient to transform the XML into a web-based presentation by using XSL [5]. XSL provides sorting and filtering functions such that the output web presentation can be organized according to the values of the XML elements. We regard the resulting web-based presentation of the video tree structure as VTOC. The video components are ordered along scenes, groups and shots. The key frame image of each shot is shown on VTOC together with the corresponding time instance in the video. Therefore, with VTOC, we can easily know what and when a video shot shown in the video. This visual video description can give us an abstract idea of the video contents. A sample VTOC presentation is shown on Figure 6.

```
<?xml version="1.0"?>
<!DOCTYPE advise SYSTEM "./toc.dtd">
<advise>
<video src="rstp:// source video on server">
<scene id="1">
  <group id="1">
    <shot id="1">
      <time value="0"/>
      <keyframe img="./sh_1.jpg"/>
    </shot>
    <shot id="3">
      <time value="11"/>
      <keyframe img="./sh_3.jpg"/>
    </shot>
  </group>
  <group id="2">
    <shot id="2">
      <time value="7"/>
      <keyframe img="./sh_2.jpg"/>
    </shot>
  </group>
</scene>
<scene id="2">
  <group id="3">
    <shot id="4">
      <time value="20"/>
      <keyframe img="./sh_4.jpg"/>
    </shot>
  </group>
</scene>
</video>
</advise>
```

Figure 5. XML Video Tree Structure


ADVISE - Video Structure in XML - Microsoft Internet Expl...
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://pc90075/video/test_vid2.xml
Go
Links
Done
Local intranet

ADVISE - Video Structure in XML

Source Video: **test_vid2.mpg**

Scene	Group	Shot	Key Frame
1	1	1	at 0 sec
2	2	2	at 1 sec
3	3	3	at 7 sec
4	4	4	at 8 sec
		6	at 11 sec
		8	at 16 sec

Figure 6: VTOC Presentation

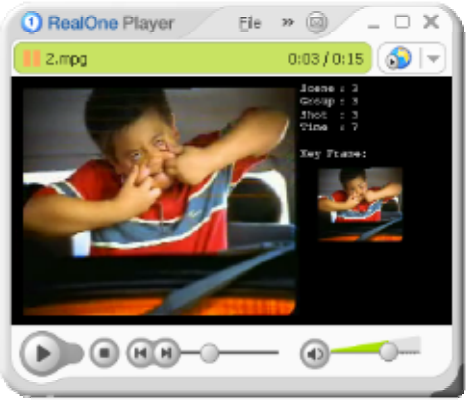
2.3 SMIL Generation Module

In this module, we customize the video into its summary using SMIL [6]. There are two good reasons for using SMIL [1]. First, SMIL benefits from the XML plain-text property. Web server can instantly generate the corresponding SMIL presentation with server-side scripting languages. The video segments are wrapped by SMIL and played according to the users' preferences. The second reason is the network and client adaptability of SMIL. It can dynamically configure the most appropriate media object for streaming, which depends on client display capabilities and connection speed. It would be convenient for the SMIL browser on the client to handle these limitations, instead of including making considerations at the SMIL generation.

The video summary is generated by a PHP script running at the web server upon receiving any user requests. It summarizes a video by combining a few-second-portion of each video shot into a sequence in the SMIL presentation. Then we are able to have a quick scanning of those selected video segments to know the contents. The script reads the video tree structure stored in XML, and then transforms it into corresponding SMIL codes. While video shots are arranged in order in the output SMIL, the script selects certain duration for each video segment. We allow users to request for different lengths of video summary, so that the script determines how much portion of each video shot should be included. A shorter summary may not give enough information about the contents while a longer summary takes more time to browse. We also allow users to request video summary of only selected components on the VTOC. They can firstly choose some interested video components from the VTOC, and browse the generated summary for more details of those components. A sample source of SMIL generated by the script is shown in Figure 7. The customized SMIL video summary is shown in Figure 8.

```
<?xml version="1.0"?>
<smil xmlns="http://www.w3.org/2000/SMIL20/CR/Language">
<head>
<layout type="text/smil-basic-layout">
<root-layout width="550" height="300" background-color="black"/>
<region id="video" left="5" top="5" height="288" width="352" fit="fill"/>
<region id="description" left="360" top="5" height="80" width="150"/>
<region id="keyframe" left="380" top="90" height="90" width="110"/>
</layout>
</head>
<body>
<seq>
<par>
<video src="rtsp:// source video on server" clip-begin="0s"
clip-end="3s" region="video" fill="freeze"/>
<textstream src="desc.rt" clip-begin="0s" clip-end="3s"
region="description" fill="freeze"/>
<img src="./sh_1.jpg" region="keyframe" fill="freeze"/>
</par>
<par>
<video src="rtsp:// source video on server" clip-begin="7s"
clip-end="10s" region="video" fill="freeze"/>
<textstream src="desc.rt" clip-begin="7s" clip-end="10s"
region="description" fill="freeze"/>
<img src="./sh_2.jpg" region="keyframe" fill="freeze"/>
</par>
</seq>
</body>
</smil>
```

Figure 7: A Sample Source for SMIL



Scene : 2
Group : 3
Shot : 3
Time : 7

Figure 8: A Customized SMIL Video Summary

3. CONCLUSION

The ADVISE system presented in this paper improves video browsing and retrieval. The system provides two ways to describe a video. The first one is the web-based presentation called VTOC, and the second one is the generation of SMIL video summary. These two methods can help users to know the contents of a video quickly before they spend much time to download the whole video. We are currently integrating the system with video matching techniques [2], such that users can make queries on searching similar videos. After the integration, we can further improve the efficiency of video retrieval from abundant sources in the Internet.

4. ACKNOWLEDGEMENTS

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. CUHK4222/01E).

5. REFERENCES

[1] R. Hjelsvold, S. Vdaygiri, and Y. Léauté. Web-based Personalization and Management of Interactive Video. In *Proceedings of the Tenth International World Wide Web Conference*, page 129-139, May 2001.

[2] C.W. Ng, I. King, and M.R. Lyu. Video Comparison Using Tree Matching Algorithm. In *Proceedings of The International Conference on Imaging Science, Systems, and Technology*, volume 1, pages 184-190, Las Vegas, Nevada, USA, June 2001.

[3] Y. Rui, T.S. Hunag, and S. Mehrotra. Constructing Table-of-Content for Videos. In *ACM Multimedia Systems Journal, Special Issue Multimedia Systems on Video Libraries*, volume 7, no. 5, pages 359-368, Sept 1999.

[4] W3C Recommendation, Extensible Makeup Language (XML) 1.0 Specification (Second Edition). http://www.w3.org/TR/2000/REC-xml-20001006, 6 October 2000.

[5] W3C Recommendation, Extensible Stylesheet Language (XSL) 1.0 Specification. http://www.w3.org/TR/2001/REC-xsl-20011015, 15 October, 2001.

[6] W3C Recommendation, Synchronized Multimedia Integration Language (SMIL) 2.0 Specification. http://www.w3.org/TR/smil20/, 7 August 2001.