ERMSCT[#]: MPEG-7 Based Didactical AV Courseware Building Tool for Heterogeneous Terminals

YUANZHI ZOU[1, 2]

[1]*Institute of Computing Technology, Chinese Academy of Sciences, P.O. BOX 2704, Beijing, 100080, P.R.China*

[2]*Graduate School of the Chinese Academy of Sciences*

TIE-JUN HUANG[3], WENGAO[3]

[3]*Institute of Computing Technology, Chinese Academy of Sciences, P.O. BOX 2704, Beijing, 100080, P.R.China*

Now video courseware is not adaptive to heterogeneous terminals, and its description metadata is too much simple to efficiently and effectively search, store, and evaluate the teaching quality. For much redundancy and domain-knowledge in the didactical AV and the maturation of MPEG-7[1][2], we have implemented ERMSCT[a] that can analyze the didactical video content, produce two variations of it for heterogeneous terminals, as well as annotate with MPEG-7. The purpose of this paper is primarily twofold. The first purpose is to how to use MPEG-7 to design the description scheme about the didactical AV. The second purpose is to show how efficient and effective the video courseware created by content-based multimedia analysis is.

1. Introduction

The didactical video courseware can provide rich content, at the same time, using digital video as an asynchronous method of instruction [3] is found useful, satisfying, easy to use, and easy to learn. But now video courseware is not corresponding to growing popularity and capability of various mobile devices and network bandwidth.

At present, there exist two basic problems in didactical video courseware. First one is that using video courseware is time-consuming and tedious. Statistical results of the survey [3] indicate the lowest rating in the category "Ease of Use". AV source may be analyzed to result in structural information [4] that can guarantee no-linear access to video data and different variations of didactical video for effective use. Second one is that only management metadata in the courseware is too simple to generate efficient and effective organization,

[#] This work has been supported by grant no 2001BA101A07 from educational resource management system (ERMS) project of Chinese Education Ministry since 2003.

[a] This is a didactical AV courseware-building tool of ERMS project.

search, and estimate the teaching quality. In learning technology area, ADLNet (http://www.adlnet.org/) proposes a metadata for learning contents called SCORM (Sharable Content Object Reference Model) that aim the interoperability and reusability of WWW based learning materials in different organizations / WWW servers, but SCORM is unsuitable for the description of video content. As yet, there are many standards describing multimedia sources, i.e. MPEG-7 [1] [2], SMEF [7], SMPTE [8], Dublin Core [9], TV-Anytime [10], and so forth, among them is MPEG-7 the most competitive and complete.

The remainder of the paper is organized as follows. Related work is introduced in Section 2. Section 3 describes an MPEG-7 based description scheme. In Section 4, analysis of the didactical AV and creation of variations of the didactical AV is shown. In section 5, we can find the efficiency and effectiveness of accessing and storing video courseware by experiment. The paper by outlining the future work is concluded in the section 6.

2. Related Work

Building courseware is a very complicated process. There has been some literature with respect to the methods about the development of courseware in recent years. The important characteristic of development of courseware is of interdisciplinarity that relates to developers, i.e. computer scientist, psychologist, graphic designer, media specialist, domain expert and so forth, as well as the consideration of psychological and ergonomical aspects [11]. Good methodology should combine learning theory, instructional design processes, multimedia design, and human-computer interface issues [12]. Situated learning is known to be an effective didactical approach; yet, multimedia systems with built-in support for it are uncommon [13]. A learning management system should enable adaptivity, the retrieval of history and state, comparison of results, tracking for pedagogical research, shared reference databases, and problem scenario databases [14]. In view of researches above, we may realize good courseware had better resolve the two problems that are facilitating interoperability and globalization of data resources and flexibility of metadata management. Although a few researches have related to MPEG-7 based content analyzing [24] [25], there are few literatures relating to MPEG-7 based analyzing video content to build the video courseware, and almost present video courseware only provides simple management descriptions that don’t benefit evaluating teaching quality based on some part of the content in AV.

3. Variations of the Didactical AV Courseware Scheme Based on MPEG-7 MDS

Designing a courseware scheme is based on the requirement as followed; the high

level structure of the courseware is shown in the Figure 1, which is compliant to the Universal Modeling Language (UML) notation [15]. According to MPEG-7 MDS [20], its definition can be implemented.

1. Organize the didactical video on the shot-level.
2. Summarize the shots.
3. Segmentation of the audio corresponds to the video shots.
4. Audio script corresponds to the video shots.
5. Describe variations of AV courseware for heterogeneous terminals.

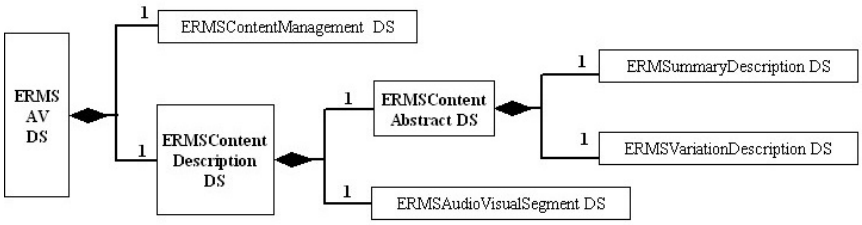

Figure 1. The high level structure of the courseware

The Didactical AV courseware may be described by an instance of ERMS AV DS. The ERMS AV DS consists of ERMS ContentManagement DS and ERMS ContentDescription DS. ERMS ContentManagement DS mainly refers to management metadata, i.e. media information, creation information, usage information, and user information and so forth. ERMS ContentDescription DS includes ERMS ContentAbastract DS and ERMS AudioVisualSegment DS. The summary of one AV courseware may be described in ERMS ContentAbastract DS, and adaptive descriptions for heterogeneous terminals can be described in the ERMS VariationDescription DS. At last the video or audio segment corresponding to each shot can be defined in the ERMS AudioVisualSegment DS.

4. Analysis of Didactical AV and Creation of Variations of Didactical AV

There exist two objectives for analysis of the didactical AV program. First one is to create the summary of the didactical AV and segment the audio content on the shot-level, second one is on the shot-level to semi-automatically recognize voice into audio script and transcode key frames in the shots into Gif format with constraint to 160*160 size and 256 color that is adaptive to PDA devices with embedded IE browser.

4.1. *Detection of Shot Boundary and Extraction of key Frame*

It can be found that there is much redundancy in didactical video because of small variations of the scene and shot changes limited to abrupt and gradual transitions. The shot semantics is corresponding to the audio script. So detection of shot boundary is sufficient to analyze the didactical video. There exist many techniques for detection of the shot boundary [16] [17] [18] [19]. In didactical video, there exists domain-knowledge as follows.

1. The shot change occurs among the teacher, the audience and the tablet or the screen.
2. When slide changes, the shot change will occur.
3. There exist salient changes of color among the teacher, the audience and the tablet or the screen.

So, two consecutive frames from different shots are not possibly to have similar colors, color is one of the most common visual primitives for detection of the shot boundary, then histograms as the descriptors for color are adopted in ERMSCT. If the distance between the color histograms of two consecutive frames was higher than a threshold, ERMSCT can detect the shot change. After detecting all the shots boundary, ERMSCT can rank all the shots by time code, if the time length of the ith shot s_i is smaller than 2 second, incorporate s_i into s_{i-1}, except for i = 1(assume that the time length of a shot is too short to express the complete semantic in the didactical video).

In addition, content-based analysis can't well solve segmentation of video on semantic level [22] [23]. When color histograms are used to detect the shot boundary, the semantic between adjacent shots may likely discontinue. So, ERMSCT can permit the user to manually refine the shots. Browsing the shots through ERMSCT, users can drag the slide bar to mark the end time of a segment of the video or to mark the start time of it, and then ERMSCT can merge adjacent shots with constraint to the rules of which the definition is as follows to automatically merge the shots. The mTimeStart, mTimeEnd, sTimeStart[i], and sTimeEnd[i] express start time of the user's manually marking a video program, end time of the user's manually marking a video program, start time of the shot s_i, and end time of shot s_i, respectively.

1. Set mediaTimeUnit [20] = "PT1N25F"(PAL video format).
2. Assure mTimeStart < mTimeEnd.
3. Localize the positions of mTimeStart and mTimeEnd in the shots, get the mTimeStart position in si, the mTimeEnd position in s_k, assure the condition satisfies k > i and i>1; otherwise, if k<i, then don't process the new shot. The discriminant of merging new shot is shown in Table 1.

Although many methods [21] can be used to extract key frames from a shot, in terms of the didactical video, the slide content of a shot is the most important

and hardly changes, so ERMSCT uses the simple way that has the middle frame be selected as a key frame.

Table 1. The discriminant of merging new shot

Conditions		Outcome
mTimeStart <= sTimeStart[i]+2*25	mTimeEnd >= sTimeStart[j]+2*25	$s_{i-1}+s_j$ (i>1)
		s_1+s_j(i=1)
	mTimeEnd < sTimeStart[j]+2*25	$s_{i-1}+s_{j-1}$ (i>1)
		$s_{i-1}+s_{j-1}$ (i=1)
mTimeStart > sTimeStart[i]+2*25	mTimeEnd >= sTimeStart[j]+2*25	s_i+s_j
	mTimeEnd < sTimeStart[j]+2*25	s_i+s_{j-1}

4.2. *Creation of AV Summary*

Creation of AV summary can provide us a quick way of skipping courseware, and an AV summary may be placed on the application servers to make users know of it. Given the state of the art of current computer vision and image understanding techniques, overall video content understanding is in their infant stages. As regards the didactical video, there is domain-knowledge as follows.

1. When the teacher changes topic, the slide generally varies.
2. The longer a slide keeps invariable, the more important the slide becomes.
3. The audio track is somehow non-smooth and jumpy if each constituent audio segment lasts for only 2 seconds.
4. The first shot relates to the lecture topic, and then the final shot refers to the lecture conclusion. They must be included in the video summary.

So, ERMSCT creates the AV summary using the following major steps:

1. Sort S that denotes a set of all the shots in the didactical AV program and is ranked by time code, S= $\{s_1, s_2 \ldots s_n\}$, select top ten longer shots, and group the ten shot into the summary cluster.
2. If there is not the first or final shot in the summary cluster, then add the first or final shot into the summary cluster.
3. In the summary cluster, take a segmentation of each shot for up to 2 seconds, and concatenate these selected segments together to form the video summary by time code.

4.3. *Creation of Variations of the Didactical AV*

As regards the MPEG-1 format of didactical Av about an hour, its data is about 600MB with 352*288 size, based on the simple description management

metadata, it is time-consuming and tedious to browse the courseware. In addition, real-time decomposition of video becomes impractical on complexity-constrained mobile devices [5] [6], but real-time decomposition of mp3 audio is practical. So there are two types of variations, one consists of the video summary, the key frame pictures (JPEG format), an audio mp3 file and description file for Personal computer and the other is composed of an audio mp3 file, the key frame pictures (Gif format) and a description file for PAD devices. Following session 4.1, segmenting the audio corresponding to the shots, ERMSCT may use the commercial voice recognition software to translate the segments of voice into the audio script and support manually correcting it. The function that can automatically transcode JPEG pictures of 352*288 size into Gif pictures of 160*160 size with 256 colors is also implemented in ERMSCT.

5. Experiment

Here, we select four didactical AV programs relating to four subjects, i.e. soft engineering, logic for mathematicians, computer principle and pattern classification and so forth to build courseware. Parameters relating to four didactical AV programs are as follows in Table 2. The Fo, T, Ns, Fa, SFpc, and SF_{PAD} denote original didactical AV program file size, time duration of original didactical AV program, the number of the shots, data of audio file extracted from original program and transcoded into mp3 format, the ratio of didactical AV file data to variation of its, data of all the key frame files for Personal computer, and data of all the key frame files for PAD device, respectively. In the experiment, lenovo XP218 is selected as a PAD device that can support embedded IE browser, and its performance references are in Table 3.

Table 2. Parameters relating to four didactical AV programs

Subject Name	Fo	T	Ns	Fa	SFpc	SF_{PAD}
Soft engineering	599MB	1H09S	67	55MB	1.38MB	0.9564 MB
logic for mathematicians	604MB	1H04S	75	55.5MB	1.37MB	1.00M B
Computer principle	531MB	53M24 S	57	48.8MB	1.10MB	0.739M B
Pattern Classification	361MB	36M15 S	32	33.1MB	0.768MB	0.475M B

Storage convenience, teaching quality estimation and adaptation to lenovo XP218 may be verified by four types of didactical AV courseware.

Table 3. lenovo XP218 PAD performance references

Operating system	Processor	Memory	Video	Wireless
Microsoft Windows Mobile for Pocket PC2003	Intel PXA255 400MH	FLASH：128MB RAM：64MB Mobile-RAM	240*320, Advanced TFT-LCD，65536 color	Support GPRS, Transmission rate: 43.2kbps

Table 2. illustrates that if a didactical AV program about an hour or so hasn't been processed by ERMSCT, downloading courseware from courseware providers is time-consuming and costly through the Internet, moreover, it is too large to download into PAD flash memory. Data of about an hour or so courseware built by ERMSCT is smaller than 60MB. Storage convenience can be found in Table 4. The rc is defined in formula (1), the rc of four programs is illustrated in Table 4. The rc and SF represent the ratio of didactical AV file data to variation of its and data of all the key frame files, respectively.

$$rc = Fo/(Fa+SF) \quad (1)$$

After four programs are analyzed, data compression is up to above 10, data of courseware mainly consists of audio data and then video information is summarized into sequences of key frame pictures, so rc fluctuates about 10.6.

Table 3. and Table 4. show that an hour or so courseware built by ERMSCT and stored in the flash memory can be browsed with embedded IE in the PAD device, and 128MB flash memory can store two hours courseware; on the other hand, lenovo XP218 PAD supports wireless transmission rate up to 43.2kbps, so it is also fit for downloading courseware for testers to watch the them.(through the internet, the on-line media service test isn't considered here, and in fact it is also fit for on-line receiving audio streams of 8kps rate that can guarantee the essential QoS of voice.)

Table 3. shows that the courseware is based on the granularity of the shot-level, testers can estimate teaching quality about some knowledge in one shot ,and but normal video courseware can't support it.

6. Conclusion

There are two main contributions in ERMSCT. First, ERMSCT provides a way to use MPEG-7 to design the description scheme about the didactical AV.

Second, ERMSCT can make two variations of the didactical AV based on content-based video analysis for heterogeneous terminals. Although ERMSCT can improve the efficiency and effectiveness of storage, search and estimation of teaching quality, there are still some problems in ERMSCT. There exist three main works in the future. First one, content analysis methods need be enhanced, i.e. if more than two teachers exist in a didactical AV program, ERMSCT isn't capable of recognizing these teachers to support the search of courseware based on a given teacher; second one, ERMSCT is limited to content-based AV analysis to make courseware, so it can't integrating other multimedia, e.g. flash animation into courseware; at last, the function that metadata descriptions based on MPEG-7 can be converted into other metadata standard should be considered in the future work.

Table 4. Compression rate illustration in four types of didactical AV courseware

Subject Name	r_c for pc	r_c for PAD
Soft engineering	10.6243	10.7048
logic for mathematicians	10.6207	10.6903
computer principle	10.6413	10.7188
Pattern Classification	10.6590	10.7520

References

1. Nack, F., Lindsay, A.T., *Everything you wanted to know about MPEG-7. 1, Multimedia, IEEE , Volume:* 6 , *Issue:* 3(1999)
2. Nack, F., Lindsay, A.T., *Everything you wanted to know about MPEG-7. 2, Multimedia, IEEE, Volume:* 6, *Issue:* 4 (1999)
3. Filippidis, S.,et al., *Using Digital Video as an Asynchronous Method of Instruction, Advanced Learning Technologies*, 2003. *Proceedings. The 3rd IEEE International Conference on*, 9-11(2003)
4. Jin-Hau Kuo, et al., A*n MPEG-4/7 based architecture for analyzing and retrieving news video programs, ICCE'*2003, *IEEE International Conference on Consumer Electronics*, 17-19(2003)
5. Www.intel.com, *Intel StrongARM SA-1100 Microprocessor for Portable Applications*
6. Richard Han et al., *Uinversal Tuner*：*A Video Streaming System for CPU/Power-Constrained Mobile Devices, vol. 9, ACM, Multimedia* (2001)
7. http://www.bbc.co.uk/guidelines/smef/
8. http://www.smpte.org
9. http://dublincore.org/
10. http://www.tv-anytime.org/

11. Khaldoun Ateyeh, et al., *Modular development of multimedia courseware, Vol.2, IEEE, First International Conference on Web Information Systems* (2000)
12. Uden, L., *Multimedia design framework for courseware, Advanced Learning Technologies*, 2000. *IWALT* 2000. *Proceedings. International Workshop* (2000)
13. Tretiakov, A. et al., *Designing Multimedia Support for Situated Learning, Advanced Learning Technologies*, 2003. *Proceedings. The 3rd IEEE International Conference* (2003)
14. Sessink, O., et al., *Author-Defined Storage in the Next Generation Learning Management Systems, Advanced Learning Technologies*, 2003. *Proceedings. The 3rd IEEE International Conference on* (2003)
15. http://www.uml.org
16. R. Zabih,et al., *Feature-based algorithms for detecting and classifying scene breaks, Proc. ACM Multimedia,San Francisco,CA,*189-200(1993)
17. Dugad, R., et al., *Robust video shot change detection, Multimedia Signal Processing*, 1998 *IEEE Second Workshop* (1998)
18. P. Bouthemy, M., et al., *A unified approach to shot change detection and camera motion characterization, Circuits and Systems for Video Technology, IEEE Transactions on , Volume:* 9 , *Issue:* 7(1999)
19. Rasheed, Z., Shah, M., *Scene detection in Hollywood movies and TV shows, Computer Vision and Pattern Recognition, 2003. Proceedings. IEEE Computer Society Conference, Volume:* 2, 18-20 (2003)
20. ISO/IEC JTC1/SC 29/WG11/N4242
21. Wolf, W., *Key frame selection by motion analysis, Acoustics, Speech, and Signal Processing Conference Proceedings. IEEE International Conference, Volume:* 2(1996)
22. Yong Rui,et al., *Digital image/video library and MPEG-7: standardization and research issues, Acoustics, Speech, and Signal ,Volume* 6, *IEEE*(1998)
23. Smeulders,et al., *Content-Based Image Retrieval at the End of the Early Years Content-based image retrieval at the end of the early years, Pattern Analysis and Machine Intelligence, IEEE Transactions on , Volume:* 22 , *Issue:* 12(2000)
24. Jae-Ho Lee, Gwang-Gook Lee, Whoi-Yul Kim, *Automatic video summarizing tool using MPEG-7 descriptors for personal video recorder Consumer Electronics, IEEE Transactions on, Volume:* 49 , *Issue:* 3 (2003)
25. B.L. Tseng, Ching-Yung Lin, Smith, J.R., *Using MPEG-7 and MPEG-21 for personalizing video, Multimedia, IEEE, Volume:* 11 , *Issue:* 1 (2004)