Automatic Video Segmentation and Story-Based Authoring in E-Learning

Yi-Chun Liao
Department of Computer Science and Information Engineering,
China University of Technology, Taiwan, R.O.C.
Email: ech_liao@cute.edu.tw

Chun-Hong Huang [1]
Department of Computer information and Network Engineering,
Lunghwa University of Science and Technology, Taiwan, R.O.C.
Email: ch.huang@mail.lhu.edu.tw

Abstract— **In e-learning, video is widely used in many course domains. We address the problem of the lecture recording and the organization of visual information through user's interaction at different steps. Our work focuses on the following three important areas: (1) synchronization method of the material contents and (2) interactions between material and users (3) construction of adaptive presentation methods in solving different knowledge levels.**

This paper proposes a story-based editing and browsing system with the automatic video segmentation. We also point out that a video classification technology can be further integrated to enhance the tool by using visual and audio information. In addition to the semantic segmentation, an instructional video can be edited with an instructor's story. The story-based editing is similar to hypervideo. Hypervideo is used as a hyperlink in a web. An instructor can constructs an instructional material by hypervideo links. For delivering pre-recorded lectures, we start our discussion on a multimedia presentation recording system which we had developed.

Index Terms— **video segmentation, story-based, e-learning, video presentation, video authoring**

I. INTRODUCTION

Video processing has been an important and challenging issue. It has simplified the editing process to automatically create high semantic video data. Usually, the video is viewed as a video document and cut into several units. The two different video segmentation approaches are used:

- Shot-based segmentation: it identifies a transition in content between two frames, and uses a key-frame to represent a video shot [1, 2, 3].
- Object-based segmentation: a frame is divided into objects and background according to the temporal relations and spatial relations. [4, 5, 6].

In many existing researches, it is desirable to identify syntactic and semantic components by using the differences of video contents, for example, sports and news [5, 6, 7, 8]. Ideally, the video will be automatically annotated as a result of machine interpretation of the semantic content of the video. Although the visual content is a major source in a video program, an effective strategy in video-content analysis is to use extractable attributes from multimedia materials.

Many video applications are also proposed in e-learning. They aim at providing a user-experience and user-friendly by adapting the content, re-using videos. A virtual video editing system indicates how principles and techniques of user-controlled video editing have been integrated into four multimedia environments [9]. Many works address the challenge of extracting structure in educational and training media based on the type of material [10, 11, 12]. Multimedia presentation systems were widely developed for many purposes, including the delivery of distance learning lectures. Some video-based applications are used in language learning environment and focus on the interactivity between teacher and student. Yoshiaki Hada [13] used XML to extend traditional videoconference system. The system can record learning scene as video file for teachers. The system also can add revises and comments what teacher wants to into conversational video between teachers and learners.

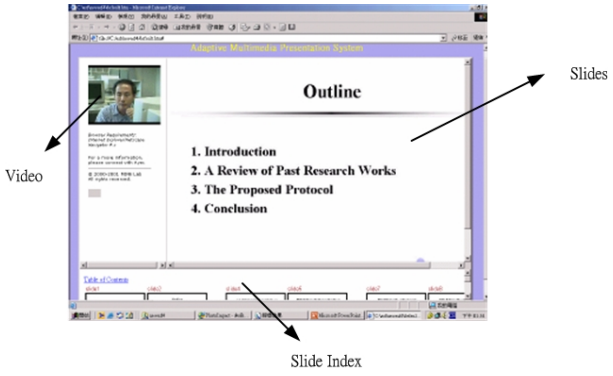
This paper proposes and evaluates a complete and novel lecture video processing framework that includes the segmentation function and recording function in real-time. In our method, it is possible to extract information rapidly. The complex function is not needed in our algorithm to reach good performance. Finally, an instructional video can be editing by story board, and playback.

[1] corresponding author: Chun-Hong Huang

II. VIDEO PROCESSING FRAMEWORK

Lecture presentation systems are used by instructors to enable classrooms created modern environments. The increasing amount of devices that are used within such environments encouraged the joint presentation of different materials prepared in advance or developed during the lecture.

In e-learning, a video application is developed as the lecture presentation system, and the commercial products are found [15, 16, 17]. The system included e-learning standard- SCORM in [17]. Figure 2(a) shows a common template with the three media streams. The synchronization issue is needed to be solved. Advances in this system other annotation messages will be designed as in figure 2(b) [18]. These systems are web-based and apply a video stream, Microsoft PowerPoint slides, and video annotations. The Synchronized Multimedia Integration Language (SMIL) provides the synchronization and interaction mechanism.



(a) Presentation template

(b) Lecture annotation system

Figure 1. Multimedia Presentation System.

The methods and tools discussed here address the needs of two users involved in the presentations: instructors and editors. A teacher can be the course instructor and the editor. In figure 2, a usual video processing in e-learning includes three steps to produce a lecture material. Ideally, the video just recording once and reuse many times by the computer processing and the instructor's author. Furthermore, the instructor's author is considered as a script or a template that is composed and exchanged between the different instructors.

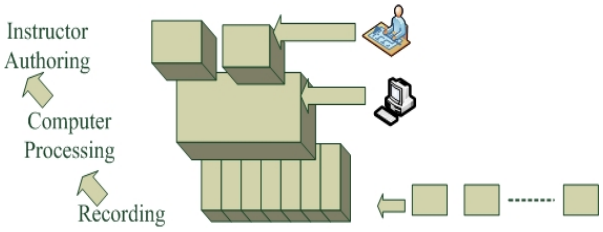


Figure 2. Video Processing Steps in e-learning.

In e-learning, the existing researches provide many good solutions in technique issues. We found that the video material be handled to be suited to use in three steps if we want to take the advantages of the video in e-learning as shown in figure 3.

- **Recording:** A lecture material is recorded in general course environment by capturing device. Or using other recording device finds the voice, text document and so on.
- **Authoring:** An editor post-product the lecture video by the proposed story board.
- **Playback:** For an end user, a simple tool is a good tool like a TV. Indirectly playback the content without terms is perfect.

III. INSTRUCTOR-BASED VIDEO RECORDING

Here we will focus on video processing issue how to provide more powerful and meaningful learning material.

There are two major components in our system: Video Recording and Video Segmentation. In the video recording phase, we focus on the issues of the synchronization and the real-time encoding. In the segmentation phase, by using visual and audio information, we propose a segmentation method.

A. Instructor's Interactions Recording

It is very convenient to use a device to record a learning content to deliver by network. But, deeply thinking about that the interaction between the instructor and the learner is not enough. Where is the interaction?

The recording system has three different multimedia streams: Screen stream, CCD capturing video stream and audio stream. It is a usual course environment with a projector and a whiteboard.

As shown in figure 4, there are two input messages that can be inserted. The interaction mark is an import message that represents which should be paid attention in a video. Additionally, if the presentation slides are used, the title and content of a slide can be used to index a video.

Our solution is to combine video signal with screen output in real-time, and to record the combined signal with a compression method. The advantage of our approach enables computer output (user interactions) to be embedded in a video, with its clearness guaranteed.

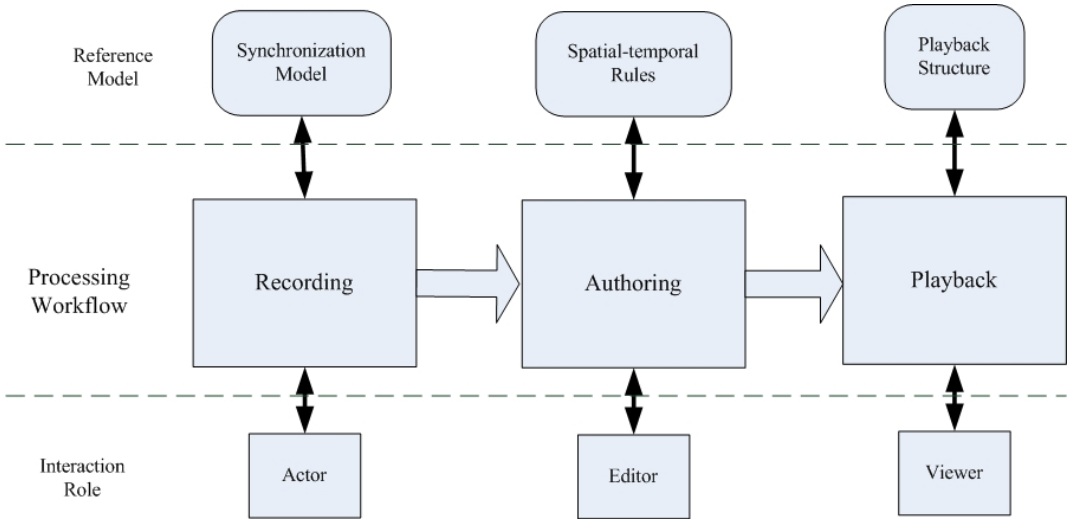


Figure 3. System Workflow.

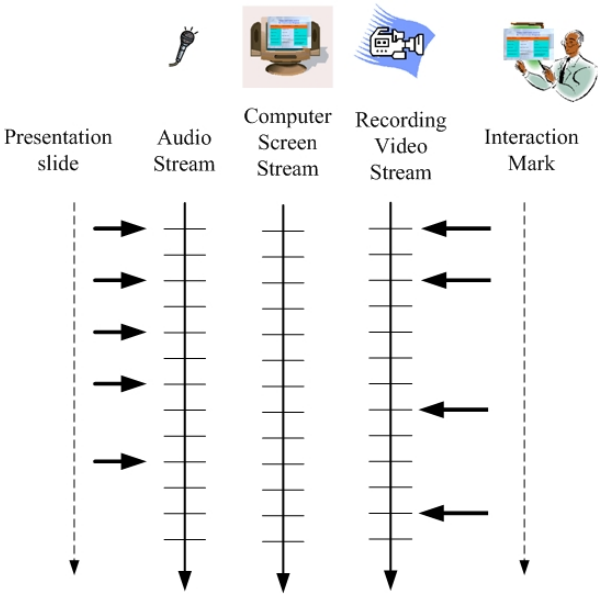
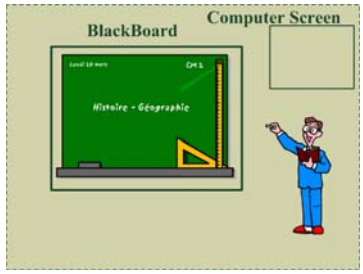


Figure 4. Interaction Marks

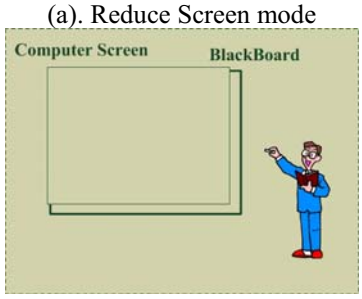
For interactive using, we provide four recording modes for the instructor as follows (in figure 5):

- **Normal mode:** This mode is default when you initial the system with setup functions.
- **Reduce Screen mode:** You can reduce screen size in this mode.
- **Full Actor mode:** In some cases, it's necessary to keep the original pictures. For example, we use zoon-in or zoon-out to capture a special scene.
- **Full Screen mode:** We provide full screen mode to record the screen like the screen recorder software.

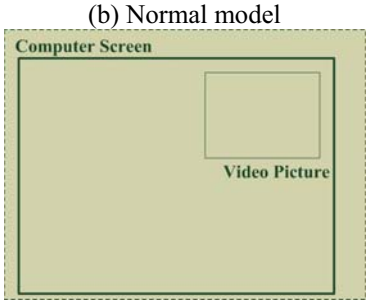
In addition, these modes can be changed in real-time. We consider that the speaker has personal presentation style which he wants to show. So the user can change the recording mode to point out the focus to students.



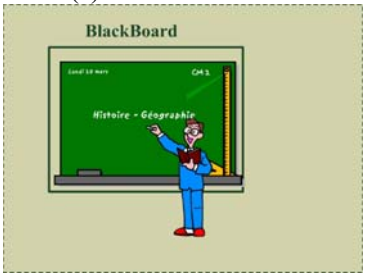
(a). Reduce Screen mode



(b) Normal model



(c) Full Screen mode



(d) Full Actor

Figure 5. Layouts of Capturing System

B. Teacher-Oriented Segmentation

In a classroom, a teacher controls the progress of a course. We find that there are three major roles in a classroom. We can classify a classroom scene as following three types (in figure 6):

- **A teacher and the blackboard**: the teacher is writing something on the blackboard at this time. Or the teacher is explaining something.
- **A teacher and students:** the teacher is talking with students.
- **Students and students:** this scene shows that the time is not course time or not a teaching time.

(a)A teacher and the blackboard

(b)A teacher and students

(c)Students and students

Figure 6. Roles of a Classroom

And then the audio information will be used as an important component of our segmentation method. Generally speaking, the features of voice can be divided into two types: frame-level features and clip-level features [19, 20]. The definition of a frame is a group of the adjacent samples. A clip is a composition by frame. Usually clip-level characteristics indicate that these frame-level features.

A scene of the classroom, not a general case, we focus on that the relations between the voices of the course role and background noise.

- **A teacher and the blackboard:** the voice is unique and continuous. The background noise is a few. The voice comes from the instructor.
- **A teacher and students:** the voice is distinct and alternating, and the background noise is a few. The voice is not continuous because of the talk between the instructor and the student.
- **Students and students:** the voices are irregular. The voices almost become background noises.

The prototype system uses four modes (as shown in figure 7). Each mode has a different screen-video layout combination. The default video size is 1024 by 768. The selection of screen size is set by an interactive interface, which allows a user to set the position and the size of screen and video. These are off-the-shelf standard components installed on a Windowing system. The selections available depend on individual setup of a PC.

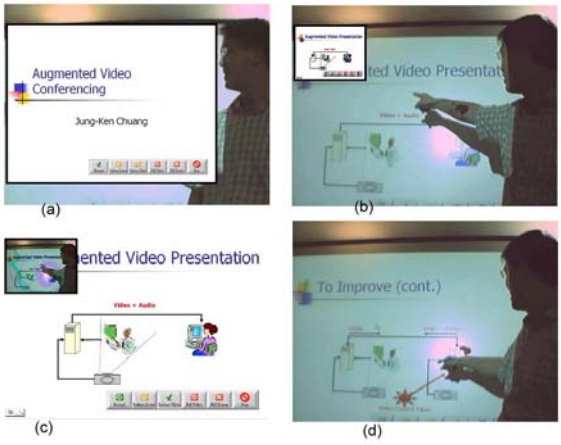


Figure 7. Interactive Recording Modes (a. Normal mode. b. Reduce screen c. Full screen d. Full actor)

IV. Story-Based Authoring

The concept of hypervideo is extended into Detail-on-Demand hypervideo by including the function of the video summarization [21].

Hypervideo is a new video processing technique similar to hypertext. This approach differs from the current system and it is a necessary in a semantic viewpoint.

For the proposed authoring tool, the following terms need to be described again:

- **Object**: An object is defined by users by using the editor to highlight. An object can be viewed as an anchor.
- **Node:** We use the node to construct our tree in Figure 8. A node could be a video clip, an image, text or a web page.
- **Scene**: It differs from the term in the tradition video processing tool. The duration of the scene is determined by an editor user. Actually, we do not focus on how to detect a scene completely and exactly. A scene is a unit while we would like to weave the different video story script.
- **HyperLink**: Link structure describes the connection between major video material and other media.

When we talk about hypertext, multimedia, and hypermedia, there is a relation between. The nonlinear

information link is a major property of hypermedia and hypertext and gives the media viewer an opportunity to decide his/her reading path.

One of the issues is that a user can insert the any kind of material into a video if it is necessary for him/her. In Figure 8, we take a video clip B as an example for constructing a hypervideo tree. Clip B includes additional information for two video clips-a resource R1 and a hyperlink L1. Video M includes two materials and a hyperlink. One of these materials can be used to link to another video that is placed at higher level. To limit the complexity of the navigating tree, all auxiliary materials of the auxiliary media will be disable. That is, for example, if we trigger Clip M to explain Clip B, we can not see the other auxiliary materials R1 and Video3 to explain M. For some reasons, we skip some video clips to view Clip M immediately by using a hyperlink L1 of Clip B.

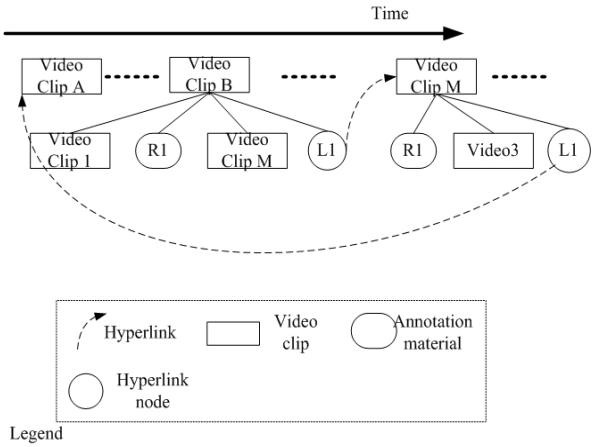


Figure 8. An example of Hypervideo tree

A. Spatial and Temporal Rules

a. Temporal Rules

In such a case, when we watch a scene, how long the presentation of an object will attract the eyes of a human? The major problem is that a not-long-enough period cannot attract the audience's attention and its function.

- **Rule1**: the duration of a defined object should be more than a threshold.

b. Spatial Rules

In designing such a small screen size of 320x240 or 640x480, we found restrictions that come from the audience and users in our experience as following:

1. The size of the marked area that a user uses to make hyperlink should be limited to a proper size.
2. Two objects can not be close to each other in a frame.

Here we define the rules to solve the spatial problem:

- **Rule2**: the size of the marking area should be large than a minimum area.
- **Rule3**: if the marked areas cross over each other and overlapped area is more than a value, the foreground is enabled and the background is disabled.

c. Spatial and Temporal Rules

Now we consider a case that there are two marked areas in Frame k and the two areas are not overlapped; in frame k+n, the area size is more than a threshold.

Combining the above-mentioned rules, we get spatial-temporal rules for HyperVideo Authoring:

Rule4: the early marked object is the foreground object if the two objects cross over each other

d. Story Board Authoring

Figure 9 shows the user interface of the video authoring. The details are described in the following.

- **Common File Format:** In our system, the format of the annotated video is not restricted.
- **Interactive Story:** Like a DVD movie, you can have a seamless presentation. In a DVD format disk, the audience has some selection buttons on a menu image when a DVD movie is playing.
- **Object annotation**: The user just loads a video file and plays it in this area like using a video player. Then using the "mark in" and "mark out" buttons to log the duration that annotated object will present.
- **Constructing the story:** We provide a video story board for the user to construct and browse the hierarchy of video. When a user wants to make a new story tree, he can pick a piece of video or cut one from the original file.

How can the user connect these pieces? The branch point shall be defined in those pieces, and the user shall determine the duration that includes a region or an object to be a "jump point". Our story board is not simply a collection of one frame, because it is unreasonable to use a frame as a branch point. It is impossible that an audience can catch a frame and select it while a video is playing.

V. Video Playing and Presentation

A. Viewer-based Presentation Player

As shown in Figure 10, in the proposed system, the key issue is how the annotation data can be played into video sequence exactly. We design a hypervideo presentation engine in a hypervideo player. The presentation engine comprises three components:

- Navigation manager: The major component of presentation engine is the navigation manager that lets user browse the video sequence and receives the user input (mouse click for video link). And this component also can control the process of the video decode while the user forward or backward the video.

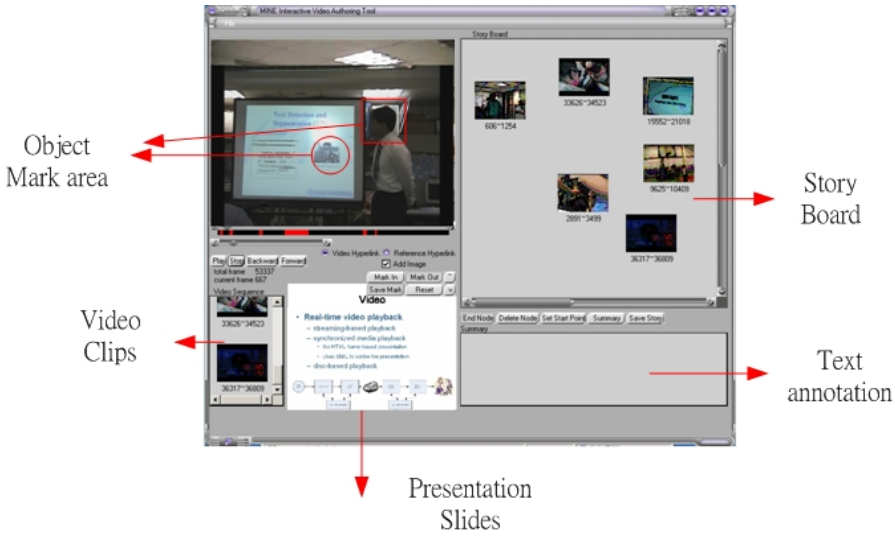


Figure 9. Story Board Authoring

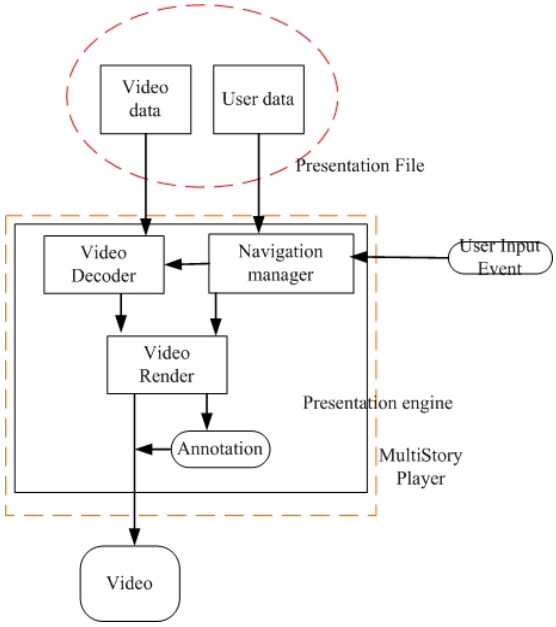


Figure 10. Architecture of Proposed Player.

- Video decoder: This component is responsible for decoding video like a MPEG decoder.
- Video Render: We use this component to render a video that comes from the output of the navigation manager and the video decoder. So if one of the inputs is interrupted, the video render will output the general video sequence without hypervideo functions.

VI. Conclusions and Future Works

A few issues that we need to solve before the system can be used on the industrial market. The synthesized video lecture has a disadvantage. We are working on a fast object tracking technology, to separate the foreground from the background. As such, the potion of discussion persons in the video can be synthesized to the video. Such a system will be more realistic.

This paper proposes novel techniques for video recording, hypervideo authoring and interactive video playback multimedia knowledge including techniques for discovering perceptual and semantic knowledge for e-learning. The interaction can be discovered when video be recorded, edited, and playback.

For material creation, we reveal a common problem and propose a solution. A recording model is also given to enhance instructors' interactions into the material. An instructor can be an actor to show the different actions in the recorded video.

References

[1] Rong Zhao and W.I. Grosky, "A novel video shot detection technique using color anglogram and latent semantic indexing", *23rd International Conference on Distributed Computing Systems Workshops (ICDCSW'03)*, pp.550, 19-22 May, 2003.

[2] Huamin Feng, Wei Fang, Sen Liu, and Yong Fang, "A new general framework for shot boundary detection and key-frame extraction", *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, November 10-11, 2005.

[3]O. Chum ,James Philbin ,M. Isard , Silicon Valley ,and Andrew Zisserman, "Scalable near identical image and shot detection", *Proceedings of the 6th ACM international conference on Image and video retrieval*, Amsterdam, pp. 549–556, 2007.

[4]Konrad Schindler and David Suter, "Object detection by global contour shape", *Pattern Recognition*, pp. 3736-3748, vol. 41, Issue 12, December 2008.

[5] Mei Han Sethi, A. Wei Hua, and Yihong Gong, "A detection-based multiple object tracking method", *Proceedings of 2004 International Conference on*

Image Processing, ICIP '04, pp.3065-3068, vol. 5, October 2004.

[6] Hui-Huang Hsu, Shih, T.K., Chia-Tong Tang and Yi-Chun Liao, "Real-time multiple tracking using a combined technique", *Proceedings of 2005 Advanced Information Networking and Applications, AINA'05*, pp.111-116, vol.1, March 2005.

[7]Jinguk Jeong, "lay segmentation for the play---break based sports video using a local adaptive model", *Journal of Multimedia* Tools and Applications, pp.149-167, vol. 39, Issue 2, September 2008.

[8] Massimo De Santo, Gennaro Percannella, Carlo Sansone, and Mario Vento, "Segmentation of news videos based on audio-video information", *Pattern Analysis & Applications,* vol.10, Issue 2, April 2007.

[9] Wendy E. Mackay and Glorianna Davenport, "Virtual Video Editing in Interactive multimedia Applications", *Communications of the ACM*, vol.32, Num 7, July 1989.

[10]Chitra Dorai, Oria, V., and Neelavalli, V., "Structuralizing educational videos based on presentation content", *Image Processing, 2003 International Conference on*, vol. 2, pp.1029-32, 14-17 Sept. 2003.

[11] Byunghee Jung, Junehwa Song, and Yoonjoon Lee, "A narrative-based abstraction framework for story-oriented video", *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 3, Issue 2, May 2007.

[12] A. Ekin, A.M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization", *IEEE Transactions on Image Processing, pp.796-807,* vol. 12, Issue: 7, 2003.

[13] Yoshiaki Hada, Hiroaki Ogata, and Yoneo Yano, "XML-based Video Annotation system for Language Learning Environment", *Proceedings of the Second International Conference on Web Information Systems Engineering*, vol. 1, 3-6 Dec. 2001.

[14] Phillip Heeler and Carolyn Hardy, "A preliminary report on the use of video technology in online courses", *Journal of Computing Sciences in Colleges*, vol. 20, Issue 4, April 2005.

[15] Kuo-Yu Liu, Natalius Huang, Bo-Hung Wu, Wei-Ta Chu, and Herng-Yow Chen, "The WSML system: web-based synchronization multimedia lecture system", Proceedings of the tenth ACM international conference on Multimedia, pp. 662 – 663, 2002.

[16] Lawrence Y. Deng, Timothy K. Shih, Sheng-Hua Shiau, Wen-Chih Chang, and Yi-Jen Liu, "Implementing a Distributed Lecture-on-Demand Multimedia Presentation System", *Proceedings of the 22nd International Conference on Distributed Computing Systems*, pp-111-115, 2002.

[17]Streaming author: http://www.streamauthor.com.tw/products/index.html

[18] Herng-Yow Chen and Kuo-Yu Liu, "Web-based synchronized multimedia lecture system design for teaching/learning Chinese as second language", *Journal of Computers & Education,* pp.693-702, vol. 50, Issue 3, April 2008.

[19] Li-Qun Xu, Yongmin Li, "Video classification using spatial-temporal features and PCA", *Multimedia and Expo, Proceedings of the ICME '03*, vol. 3, 6-9 July 2003.

[20] Chen-Hsiu Huang, Chi-Hao Wu, Jin-Hau Kuo, Ja-Ling Wu, "A musical-driven video summarization system using content-aware mechanisms", *Circuits and Systems,* 2005(ISCAS'2005), vol. 3, May 2005.

[21] Sawhney, N., Balcom, and D. Smith, "I. HyperCafe: narrative and aesthetic properties of hypervideo", *Proceedings of the seventh ACM conference on Hypertext*, pp. 1-10, 1996.

Yi-Chun Liao is an Assistant Professor of the Department of Computer Science and Information Engineering at China University of Technology, Taiwan. He received the PhD degree from Tamkang University, Taipei, Taiwan in 2005, and received his BS and MS degrees in Computer Engineering from Tamkang University in 1998 and 2001, respectively. His current research interests include Multimedia Computing, e-Learning, and Video Processing.

Chun-Hong Huang is an Assistant Professor of the Department of Computer Information and Network Engineering at Lunghwa University *of Science and Technology*. His current research interests are in the areas of multimedia processing, 3D information analysis and retrieval. His contact email is ch.huang@mail.lhu.edu.tw. (Corresponding author)