2-2007

Lecture video enhancement and editing by integrating posture, gesture, and text

Feng WANG

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Ting-Chuen PONG

Citation

1

Lecture Video Enhancement and Editing by Integrating Posture, Gesture and Text

Feng Wang & Chong-Wah Ngo & Ting-Chuen Pong

***Abstract*— This paper describes a novel framework for automatic lecture video editing by gesture, posture and video text recognition. In content analysis, the trajectory of hand movement is tracked and the intentional gestures are automatically extracted for recognition. In addition, head pose is estimated through overcoming the difficulties due to the complex lighting conditions in classrooms. The aim of recognition is to characterize the flow of lecturing with a series of regional focuses depicted by human postures and gestures. The regions of interest (ROIs) in videos are semantically structured with text recognition and the aid of external documents. By tracing the flow of lecturing, a finite state machine (FSM) which incorporates the gestures, postures, ROIs, general editing rules and constraints, is proposed to edit videos with novel views. The FSM is designed to generate appropriate simulated camera motion and cutting effects that suit the pace of a presenter's gestures and postures. To remedy the undesirable visual effects due to poor lighting conditions, we also propose approaches to automatically enhance the visibility and readability of slides and whiteboard images in the edited videos.**

***Keywords* : Lecture video editing, gesture, posture and video text recognition**

I. Introduction

Due to the popularity of distance education and e-learning, recorded lectures and presentations are becoming more and more widely used. To produce high quality videos, expert cameramen and professional editors are usually required to handle the capture and editing work. This process is impractical in most cases due to the associated costs and labourious work. The advances in content-based video analysis, nevertheless, have brought new opportunities for the automatic indexing and editing of lecture videos due to two facts. Firstly, the classroom environment is structured and this makes it easier to detect the dynamic changes such as moving objects and handwritten annotations. Secondly, the captured videos are usually associated with external textual documents (*e.g.,* PowerPoint). The *linking* of videos and documents could be accomplished by exploiting the relationship between visual, audio and texts.

In the past few years, numerous issues have been addressed for the content analysis of lecture or instructional videos. These issues include topical detection, synchronization, summarization and editing. Typical demonstrated systems include Classroom 2000 [1] and BMRC lecture browser [24]. In topical detection, a lecture video is structured according to the topics of discussion by audio [15], visual [12], [15], [19] or cinematic expressive cues [23]. The detected topics are synchronized (or linked) with external documents for effective indexing, retrieval and editing [5], [12], [19], [27]. To facilitate browsing and summarization, keyframes [11], mosaics [13] and statistical highlights [3], [8] are also extracted.

This paper addresses the issues of video editing based on the analysis of poses, gestures and texts in lectures. Previous related works include [14], [16], [19] and [30]. In these approaches, a stationary overview camera is used to capture the overview of a presentation, while another tracking camera is used to track the lecturer ([14], [16] use another camera to track and capture the audience). The editing is achieved by switching shots among cameras based on a set of predefined rules. Due to the lack of content such as gesture and pose analysis, the edited videos are usually not natural enough. In particular, the interaction between a presenter and the projected slides cannot be easily emphasized and realized. Among these approaches, except [19], no preprocessing step (*e.g.,* topical detection and linking) is performed prior to editing. The preprocessing, which structures and links video segments to slides, facilitates the analysis of interaction between the presenter and the slides.

We address three editing problems in the domain of lecture videos: (i) What to show at any given moment? (ii) How to display the rhythm of a lecture in an aesthetic way? (iii) How to improve the readability of texts in the edited videos? The first problem involves multi-modality content analysis, with the aim to understand the flow and focus of lecturing. We explore gesture, posture and text of videos to discover the interaction between a presenter and the targeted focus. The list of tracked interactions, associated with their focus instances, form the observations to determine the rhythm of the edited video with aesthetic considerations for the second problem. To this end, we propose a finite state machine (FSM) to show the interactions by encoding the editing rules to constrain the selection of focal length, cutting and camera motion for the edited video. By awareness of the underlying gestures, poses and focus of lecturing, FSM can simulate appropriate motion like zooming in on a particular region to emphasize the interaction between a presenter and the concept under explanation.

Capturing high quality videos with a simple camera setup is usually a difficult task. Normally, the qualities of lecture videos are unsatisfactory due to environmental conditions (*e.g.,* lighting effects), low resolution of video cameras, and video compression which usually affects the visualization of texts in videos. For the third problem, we propose two approaches to enhance the visual quality of LCD projected slides and whiteboard images. The first approach, similar to [12] but in a more efficient way, utilizes the information available in the external documents to produce high-resolution slides in the edited video. The second approach improves the quality of the handwritings on the whiteboard by video text detection and color contrast enhancement.

F. Wang is with the Department of Computer Science, Hong Kong University of Science & Technology, Clear Water Bay, Kowloon, HK. Email: wfeng@ust.hk

C. W. Ngo is with the Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong. Tel: (852)2784-4390. Fax:(852)2788-8614. Email: cwngo@cs.cityu.edu.hk

T. C. Pong is with the Department of Computer Science, Hong Kong University of Science & Technology, Clear Water Bay, Kowloon, HK. Email: tcpong@ust.hk

The main contribution of this paper is to propose a gesture and posture driven editing approach to trace the flow of lecturing, by attending to the focus of lecturing at any moment. Meanwhile, the aesthetic elements, which outline the general and basic rules of selecting and adjoining various views of focuses, are taken into account to generate the appropriate rhythm for showing the dynamic interactions between the presenter and the focuses. To improve the visual readabilities of the projected and handwritten words in the edited video, two approaches are also proposed to enhance the visibility of texts on the LCD projected screen and the whiteboard respectively. The remainder of this paper is organized as follows. Section II discusses and compares existing capturing and editing systems. Section III describes the camera setting, preprocessing and major components of our framework. Gesture recognition, head pose estimation and text analysis are described in Section IV to address the problem of focus estimation. Under asethetic considerations, a finite state machine (FSM) is proposed to display the focus of lecturing with a set of editing rules in Section V. Two algorithms are proposed in Section VI to enhance the visibilities of slide texts and whiteboard handwritings. The experiment results and usability studies are given in Section VII. Finally, we conclude the paper in Section VIII.

II. Related Works

A. Lecture Capture

Video camera is perhaps the most popular device in capturing live lectures due to its cost and flexibility in capturing various multimedia information. An associated problem, nevertheless, is the low resolution and quality of the produced videos. As a result, besides video camera, several other devices are used by different systems to capture data with higher quality. In [1], a structured high-tech classroom (Classroom 2000) is introduced to acquire data from both the teacher and the students. Besides placing several cameras in the corners of the classroom, an electronic whiteboard is equipped to capture the teacher's handwritten annotations, while electronic notebooks are given to students to record the notes made during the class. Because each classroom needs to be installed with the required hardware and software, the system is not easily portable. The expenditure due to the hardware and software costs is considerably high. A certain amount of manual work is usually required in order to manage and synchronize the multimedia information for browsing.

In [24], an RGB capture device is installed to directly acquire the high-quality video stream projected to the screen from the computer. The recorded information, nevertheless, is usually limited to the slide images being projected onto the screen. Because the high-resolution external documents of lecture notes are usually available, the use of RGB capture device may be not necessary. In our previous work [27], we utilize the reconstructed super-resolution video texts to synchronize the LCD projected video and the compound documents. In this paper, we further propose an efficient approach to register the video and the documents with video texts. Because the mapping from video frames to external documents is known, we can directly enhance the visual quality of the original video without the aid of RGB capture devices. A major advantage of this approach is that the interaction between a presenter and slide images is still preserved after projecting the high-resolution symbolic documents to the video.

B. Video Editing

Relatively few works have addressed the issue of lecture video editing ([6], [19], [30], [14], [25], [21]). In [30], by detecting the changes of slide images on the screen, editing is carried out by switching shots between the screen and the presenter. Several simple editing rules, *e.g.*, the duration of each shot should not be too long or too short, are applied to the editing. The main drawbacks of the system are: i) the lack of content analysis to highlight the lecture focus; ii) the quality of the produced video remains as low as the original one. In [14] and [25], various video editing rules suggested by videographers are adopted and automatically applied in a system similar to [30]. The visual changes computed based on the video frame difference from a static wide-angle camera is used to guide an active camera to pan, zoom or tilt. Although no precise gesture detection is performed, some gesture changes may also be captured and highlighted by the active camera. In [21], frame difference and skin color are employed to simply estimate a presenter's head poses. The systems in [14], [25], [30] and [21] attempt to automatically manage several cameras to capture live presentations for online audiences. Even though the qualities of the edited videos may not be as good as those produced by professional videographers, they are much better than physically presenting the raw videos without the videographers' aid. In these systems, to optimize speed for real-time broadcasts, simple features such as frame difference [14], [25], [21], screen changes [30] and skin color [21], instead of more sophisticated techniques, are adopted to detect possible events in a presentation for video editing.

While [14], [25], [30] focus on real-time broadcasting, the issues in offline editing of lecture videos have also been addressed in [6] and [19]. Due to the exemption of real-time constraint, offline video editing has the capacity of performing more detailed content analysis. For instance, by permitting delay for understanding the contextual flow of lecturing, better editing decisions such as the appropriate selection of focal length and camera motion could be determined.

In [19], computer vision techniques are applied for the detection of presentation topics and the synchronization of slides with videos. By taking into account the topic boundaries and the constraints on shot duration, a single video stream is automatically produced by picking video segments from a tracking camera and an overview camera. In [6], a framework for virtual videography is proposed. In this framework, gesture analysis is utilized to guide the editing, while camera panning and zooming are inserted to pinpoint the regions of interest. Nevertheless, the editing is done offline in a manual operation which is a laborious and time-consuming process. Table I gives a brief comparison of various systems. Our current system is not real-time since offline processing is required for automatic content analysis and visual quality enhancement.

In most presentation authoring systems [1], [19], the resulting multimedia documents contain multiple streams including videos, slides, the teacher's and students' annotations. While

TABLE I
COMPARISONS OF EDITING TECHNIQUES

	[19]	[6], [7]	[30]	[14], [25]	[21]	Ours
Room setting	LCD projector	Chalk board	LCD projector	LCD projector	Chalk board	LCD Projector, whiteboard
Content analysis	Shot detection & synchronization	Gesture	Screen changes	Visual changes	Gesture, pose	Shot detection & synchronization, gesture and pose
Visual enhancement	No	Yes	No	No	No	Yes
Real-time	No	No	Yes	Yes	Yes	No
Automatic	Yes	No	Yes	Yes	Yes	Yes

browsing these documents, the users have to switch between different streams. Our system aims to produce a single video stream that combines most of the useful information so that the users can concentrate on just one stream. One essential condition is that the produced video should be in high resolution and of good quality so that the use of other streams is not necessary. In [6] and [7], one method is proposed to create high-resolution chalkboard images from different views. Basically once an ROI (Region Of Interest) is detected, a close-up view is created by filling in the edited frame with the ROI extracted from another still camera with better visibility. In this paper, we further propose two algorithms to automatically enhance the visibilities of texts on the projected screen and the whiteboard.

III. Overview of the Proposed Framework

Figure 1 illustrates an overview of our framework for lecture video editing. We consider the classroom setting with an LCD projected screen and a whiteboard being placed side by side. Two stationary cameras are pointed to the screen and the whiteboard respectively. This camera setting can be easily amended for classrooms with only a screen or a whiteboard. Before lecture capture, external documents are uploaded for the presentation. The videos, together with the documents, are then fed into our system after class. The two videos are automatically synchronized by comparing their audio tracks.

As shown in Figure 1, the video-taped lectures are initially divided into shots and further synchronized with the external documents. The shot detection is based on our previous work in [20] and [27]. The synchronization depends upon the matching of texts in videos and documents [27]. The linking of shots and slides facilitates focus analysis and visual enhancement. The focus at any given moment is estimated by finding the presenter-slide interaction through gesture, posture and text recognition. The shot-slide registration is done to locate the exact ROIs under interactions for visual enhancement and editing.

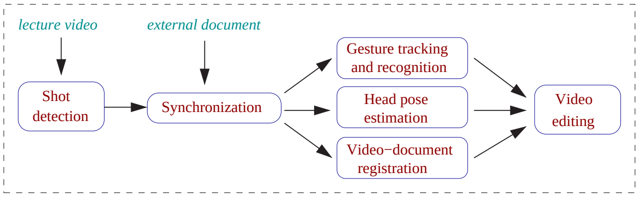


Fig. 1. A framework for lecture video editing

IV. Posture, Gesture and Text Recognition for Focus Estimation

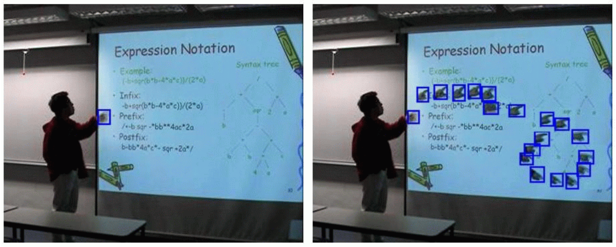
Posture, gesture and video text are three major visual cues that describe the activities in a classroom. The recognition of these cues has been intensively studied in computer vision research communities. In this section, we employ appropriate techniques to extract useful cues for lecture focus estimation. Previously we have proposed techniques for gesture tracking and recognition [28], and video slide enhancement through video-slide synchronization [27], [28]. These techniques are applied to lecture videos with an LCD projected screen. This paper extends techniques in [27], [28] for videos with both screen and whiteboard. In addition, we consider posture tracking and recognition. In contrast to [30] which utilizes gestures to decide camera zoom, three cues (gesture, posture and text) are jointly explored to track the flow of lecturing and mark the focuses to allow more complex editing decisions.

A. Gesture

Gestures are used by almost all presenters. Most gestures in lectures are deictic, which are used to direct the students' attentions to something that the lecturer is talking about. Gesture is therefore a reliable cue to estimate the focus of the lecture. However, due to the lack of salient features, the robust detection of hand gestures appears as a difficult problem. In our case, we deal with this difficulty by restricting the search region of gestures within and surrounding the detected text regions. In other words, the texts and figures in both screen and whiteboard are partitioned into various regions (see Section IV-C for details and Figure 6(a) for example). Our approach keeps track of the interaction between gestures and the regions.

In [4], frame difference is used to detect a presenter's gestures. In this paper, we utilize skin color, besides frame difference, for more robust gesture detection. To rapidly locate potential candidates, we adopt the rule-based classifier in [22] to efficiently detect skin color pixels. The classified pixels are then spatially grouped as disjoint skin regions by density-based clustering. By combining frame difference and skin color, most gestures can be detected correctly. Figure 2(a) shows an example where a gesture is detected. Once a gesture enters the surveillant region, our approach tracks the gesture and logs its trajectory over frames. Figure 2(b) superimposes the tracked trajectory on the slide for illustration. Unintentional gestures can be easily discarded since they usually appear and vanish in a short period of time.

During a lecture, a sequence of hand gestures is continuously and dynamically changing and mixing with some non-gesture movements. As observed, a trajectory can consist of multiple meaningful gestures smoothed by intermediate non-gesture movements. Figure 3(d) depicts a gesture path with three meaningful gestures, and figures 3(a)-(c) show three frames along the path. A subsequent problem after tracking is to segment and extract useful gestures from a trajectory.

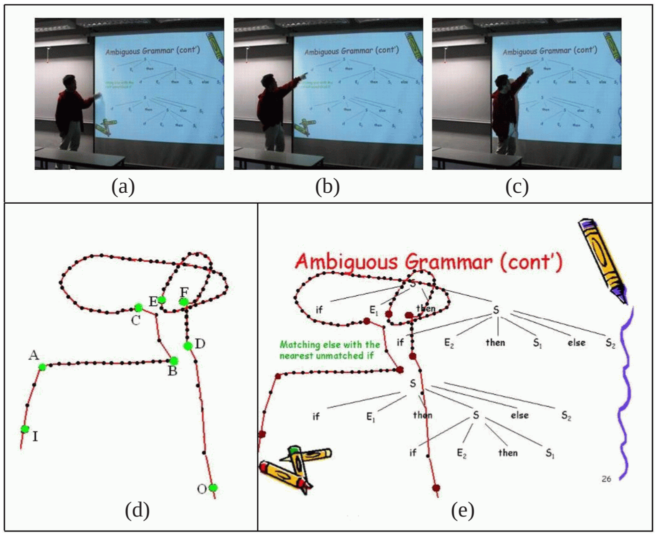


(a) (b)

Fig. 2. Gesture detection and tracking. (a) Detected gesture; (b) Tracked gesture.

The purpose of gesture segmentation lies in two aspects: i) to estimate different focuses when several gestures are connected together; ii) to reject some meaningless hand movement.

To extract individual gestures, we employ the heuristic breakpoint detection algorithm in [28]. This algorithm utilizes the hypothesis that immediate movements are fast and span insignificant time intervals, which complies with our observation of typical gesture paths. By this algorithm, points A to F in Figure 3(d) are identified as breakpoints, while points I and O are regarded as entrance and exit respectively. Consequently, the segments AB, CE and EF are extracted for gesture recognition based on the hypothesis.

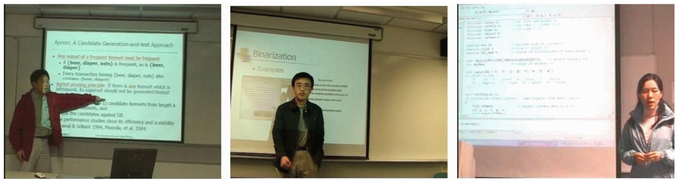

Ambiguous Grammar (cont')
E
F
C
D
A
B
I
O
Ambiguous Grammar (cont')
if
E1
then
S
E2
then
S1
else
S2
Matching else with the nearest unmatched if
26

(a) (b) (c)

(d) (e)

Fig. 3. Generating gesture trajectory by tracking. (a) at point A; (b) on circle CE; (c) at point E; (d) gesture path; (e) gesture path on slide.

We identify three typical gesture classes: *pointing*, *lining* and *circling* that are generally considered as useful cues for editing. A total of twenty points are uniformly sampled from each segmented gesture for feature extraction. The number of points being sampled is not a critical issue as long as these points provide distinctive features to describe the evolution of its gesture stroke, and most importantly, are tolerant to noise and jerky movement. Given a sampled point v_i, we compute its relative distance d_i and angle ϕ_i as features. Denote v_0 as the starting point of a segment s, and v_m as a point in s that has the longest distance from v_0. The features $d_i = |v_0 v_i|/D$ and ϕ is the angle between the line $v_{i-1}v_i$ and the horizontal axis, where D is a normalizing factor which denotes the distance from v_o to v_m, *i.e.*, $D = |v_0 v_m|$. Based on the extracted features, we employ the discrete Hidden Markov Models (HMM) for gesture training and recognition [17].

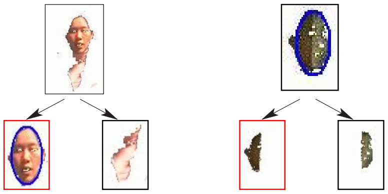


(a) (b) (c)

Fig. 4. Challenges of posture recognition in lecture videos.

B. Posture

Besides gesture, a presenter's posture is another cue that can direct students' attentions during a lecture. Presentation capture has posed several new technical challenges for posture recognition. The task is difficult when the low-level multi-modal features are coupled with complex lighting conditions in classrooms and the low-resolution quality of videos. Figure 4 shows a few examples to illustrate the challenges of recognition. In Figure 4(a), the front lights are turned off in order to make the text on the screen visible. When the presenter stands in front of the screen, half of the face looks dark while the other half is illuminated by the light from the LCD projector. In Figure 4(b), the face is overlaid with a slide image emitted from the projector. As seen in figures 4(a)-(c), effective posture detection and recognition is even more challenging considering the fact that a face merely occupies approximately 1% of a video frame, when only one camera is used to capture the overview of a projected slide.

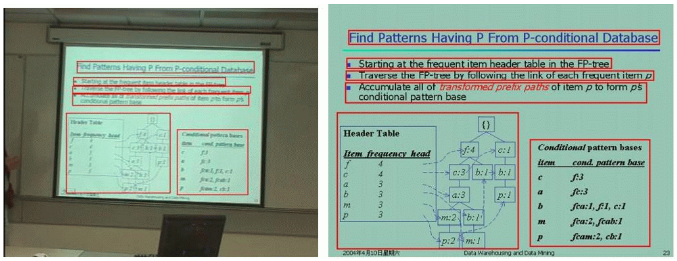


(a) (b)

Fig. 5. Hierarchical clustering of skin color pixels. (a) Skin colors of Fig. 4(a); (b) Skin colors of Fig. 4(c).

Like gesture detection, we utilize skin color, which has been demonstrated as a reliable cue in [10], for face detection. Skin color detection is vulnerable to noise due to varying lighting conditions and skin-like colors. Our skin color classifier indeed occupies a rather large region in color space by considering different races and brightness. Most skin pixels can be correctly detected if the color of the skin does not change significantly when being projected by LCD light. Some noises, nevertheless, are included which can ultimately affect the skin pixel clustering.

Figure 5 shows two types of noise that are difficult to deal with in lecture videos. In (a), the face appears to be split into two with different illuminations. In (b), the shirt has a skin-like color. To robustly handle noise, we propose a two-level hierarchical clustering algorithm. Skin density is considered at the first level while the difference of skin color is utilized for further decomposition at the second level. Initially, all the detected skin color pixels are spatially grouped by a set of



Find Patterns Having P From P-conditional Database

- Starting at the frequent item header table in the FP-tree
- Traverse the FP-tree by following the link of each frequent item *p*
- Accumulate all of *transformed prefix paths* of item *p* to form *p*'s conditional pattern base

Header Table

Item	*frequency*	*head*
f	4	
c	4	
a	3	
b	3	
m	3	
p	3	

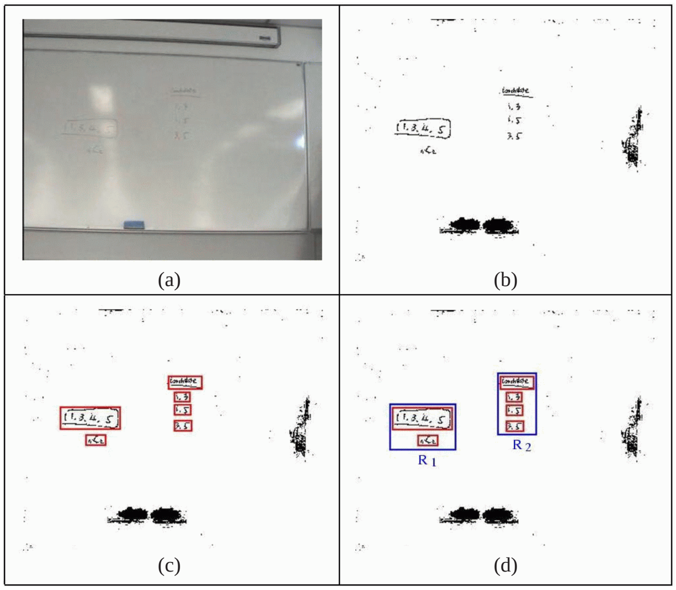
***Conditional* pattern bases**

item	*cond. pattern base*
c	*f:3*
a	*fc:3*
b	*fca:1, f:1, c:1*
m	*fca:2, fcab:1*
p	*fcam:2, cb:1*

2004年4月10日星期六 Data Warehousing and Data Mining 23

(a) (b)

Fig. 6. Structuring video content (a) with the text layout of its electronic slide (b).



(a) (b) (c) (d)

Fig. 7. Text detection in whiteboard. (a) Original frame; (b) Detected difference; (c) Detected textboxes; (d) Grouped text regions.

rectangular blocks. Each block is then segmented into several clusters according to its color difference. The blocks in both levels comprise the candidate faces for further face detection in pose recognition. Figure 5 shows the results of hierarchical clustering. In (a), the face remains as a whole at the top level although it has been segmented into two parts at the lower level. In (b), the face and shirt are decomposed as two parts in the second level by color difference.

Based on the hierarchical representation of skin clusters, we detect and track the facial features by a two-phase verification procedure. The details are given in Appendix. A feature vector of length 24 is extracted from the face template for pose estimation. The vector basically encodes the relative position and direction of facial features, which are generated directly from the internal parameters of the template (details in Appendix). We define three poses (*left*, *frontal*, *right*) and construct a neural network for recognition. The smoothness of pose transitions is exploited to improve pose estimation. A window function is used to remedy false estimation. For instance, one *left* pose surrounded by consecutive *frontal* poses is smoothed as *frontal* pose. While the algorithm here is appropriate for offline pose estimation, we also propose another algorithm in [29] which is efficient enough for real-time applications.

C. Video Text

While posture and gesture characterize the dynamic changes of focus during a lecture, video text structures the candidate ROIs. The ultimate aim of text recognition is to semantically organize the text layout and seamlessly improve the visibility of texts in videos with the aid of external documents. Video texts, for instance, suffer from poor visibility and cannot be fully recognized even by human without the aid of external documents. We achieve both tasks through the reconstruction of geometric transformation between video and external documents, by utilizing the recognized texts in videos.

During video capture, the slide images are projected to a camera plane that is usually not parallel to the projected screen. To estimate the projection, we compute the homography [28] by corresponding points between videos and slides. The points are extracted through text detection and recognition [27]. To ensure the robustness of estimation, we select twenty pairs of matching titles with high confidence based on the similarity measure proposed in [27]. The positions of titles form the matching points for homography computation.

By registration, the relationship between a video and its external document can be easily realized. With a PowerPoint slide as example, the paragraphs and figures are semantically grouped into separate objects with symbolic markers. By extracting the markers and constructing the one-to-one mapping between the instances in slides and videos through homography projection, we can easily organize and structure the layout of videos. Figure 6(a) shows the layout of a video frame where texts and figures are semantically organized with the aid of an external document in Figure 6(b). With layout structuring, gesture detection and tracking can be effectively performed as described in Section IV-A. Meanwhile, the synchronization of gestures and external documents is also feasible. Figure 3(d) shows a moving path of gestures superimposed on top of an electronic slide. The gesture path marks the flow of lecturing, and indicates the interactions between gestures and semantic instances over time. By awareness of the interactions, the focus of lecturing can be estimated, while the visibility of focus can be enhanced with the aid of external documents.

It is not easy to capture a whiteboard video with high visual quality and resolution due to the lighting conditions. Different bright spots as shown in Figure 7(a) can usually be observed on a whiteboard. The handwriting usually has a rather low contrast to the whiteboard and does not show strong edges. As seen in Figure 7(a), the handwriting is difficult to detect directly in the original frame. To solve this problem, we maintain an empty whiteboard image initially and update the handwriting added to the whiteboard over time. The handwriting is detected by the difference between the current and empty whiteboard images. The procedure is illustrated in Figure 7. In Figure 7(b), the pixel difference between the original frame in (a) and the empty whiteboard is shown. We employ the video text detector in [27] to extract the textboxes from the whiteboard while removing the non-text regions. The result is shown in Figure 7(c). After text detection, the whiteboard is segmented into several regions by grouping neighboring textboxes into the same region. Figure 7(d) shows two resultant regions on the whiteboard.

D. Focus Estimation

The first problem in lecture video editing is to decide what should be shown at any given moment. In our approach, this is addressed by estimating the focus of lecturing which emphasizes the intention of a lecturer and the attentions of the students. Physically, the focuses may include a presenter, an overview or a specific region of the screen or the whiteboard.

The focus estimation is determined based on the recognized postures and gestures. A higher priority, nevertheless, is given to gesture than posture. Generally speaking, gesture shows to be a more reliable feature. Posture is changing from time to time depending on the presenter's lecturing style. Gesture provides vivid cue to mark the flow of lecturing. The exact ROIs can indeed be located by analyzing the interaction between the gesture and the screen or the whiteboard, without awareness of the postures. However, postures are useful when gestures are occluded or absent. When only posture is present, we define three types of focus: *presenter*, *screen* and *whiteboard*, corresponding to the three kinds of postures recognized in Section IV-B. In general, postures indicate the overview of a focus, while gestures zoom in on the ROIs.

One interesting aspect of our focus estimation is that the intentional gestures are extracted from free-hand movement, together with the semantic structuring of lecture content for recognition. A gesture is intentional if it is recognized as one of the defined gestures and interacts with one or few ROIs semantically segmented in the slide or whiteboard. To the best our knowledge, this work has not been previously addressed.

V. Video Editing

In video editing, we need to decide not only *what* is to be shown, but also *how* to show it. The former is determined by the focus estimation in Section IV, while the latter is based on aesthetic considerations. In this section, we first decompose a shot into subshots associated with actions and focuses to facilitate content representation. These subshots form the basic units for editing decisions. The task of editing is then to composite the subshots so as to generate the pace and rhythm that suit the gestures, poses and focuses. We propose a finite state machine which integrates a finite number of actions and focuses with editing constraints to automatically simulate camera motion and transitions with aesthetic considerations.

A. Subshot Representation

Through Section IV, we monitor the content of a shot as a series of recognized gestures and postures interacting with a variety of focuses over frames. Each focus is an instance obtained after layout structuring described in Section IV-C. Based on this information, a shot is readily partitioned into smaller segments, called subshots. Each subshot consists of a single unit of action where an action can be a gesture, a pose or a combination of gesture and pose. In principle, a subshot is associated with at least one focus depending on the underlying action.

B. Aesthetic Considerations

Our original videos are captured by two stationary cameras. Watching a video with a fixed focal length can be dull, and more importantly, the focus of lecturing cannot be appropriately delivered, particularly if the focused content is too small to read. To produce a focus-oriented video with the proper rhythm of showing, the aesthetic elements, *i.e.*, the rules and idioms that a videographer usually practises [2], [18] need to be carefully considered in editing. These elements include focal length, view transition and subshot duration. The focal length is to emphasize the degree of interaction, while the selection of focal length is mainly dependent on the underlying action jointly governed by gesture, posture and focus. To avoid abrupt change of focal length, the view transition and the duration of a subshot can be determined directly based upon the general cinematic rules [2], [18] which outline the basic regulations of placing and connecting subshots for almost all video genres. In addition, to guarantee the smoothness of transitions, various transitional effects including camera motion and cut can be simulated to connect adjacent subshots of different views. Overall, the rule of thumb is to deliver focuses with appropriate views and camera motion while keeping the coherency and momentum of storytelling. In principle, the proper way of echoing the focus of a subshot is jointly determined by the previous status and current intention of showing. Table II summarizes the rules we used for lecture video editing.

In focal length selection, we consider three views: loose view (LS), medium view (MS) and tight view (TS). LS captures the overview of the screen or the whiteboard. MS captures the medium view and emphasizes the interactions between actions and focuses. TS is a close-up view to highlight focuses. The focal length specifies the range of view and the intensity of the interaction to be expressed in storytelling. In our design, the selection of focal length is mainly based upon the observation of gestures and postures. Table II describes how to determine the focal length of a subshot. When no gesture is detected, an LS of the screen or the whiteboard is shown, depending on which side the presenter is facing, to give an overview picture. When a frontal pose is recognized, an MS is displayed to have a closer view of the presenter's emphasis. If a gesture is present, either an MS or a TS of the focus is shown. In general, if a set of focuses is circled or a particular figure is pointed, a close-up view will be generated to highlight the region of interest (ROI). When neither gesture nor face is observed, the focus is the whole screen or whiteboard determined by the position of the presenter.

Based on the general cinematic rules [2], [18], the changes of focal length from one subshot to another need to be coherent in order not to generate abrupt view transitions. Two subshots with large scale difference in focal length should not be adjoined directly. For instance, LS should not be adjacent to TS and vice versa. These rules are enforced in Table II by prohibiting the transitions from LS to TS or TS to LS. There are various ways of connecting adjacent subshots of different focal lengths during transition, for instance, by camera cut, dissolve and zoom. The choice of transitional effects is determined

TABLE II

EDITING RULES (SCR: LCD PROJECTED SCREEN; WB: WHITEBOARD, "-": DON'T CARE CONDITION)

Current subshot			**Inputs**			**Next subshot**	
Duration	*Focus*	*Focal len*	*Gesture*	*Posture*	*Focus*	*Focal len*	*Transition*
≥ 120 frames	Scr (Wb)	LS	No	left/right	Wb (Scr)	LS	Cut
			No	frontal	Presenter	MS	Cut
			Yes	-	ROI	MS	Zoom or Cut
			No	No	Scr (Wb)	LS	No or Cut
	Lecturer	MS	Yes	-	ROI	MS	Cut
			No	left/right	Wb (Scr)	LS	Cut or Translate
			No	N̲o	Scr (Wb)	LS	Cut
	ROI	MS or TS	lining or pointing		Textline(s)	MS	Translate
			circling	-	Figure or Textlines	TS	Zoom or Translate
			pointing	-	Figure	TS	Zoom
			No	left/right	Wb (Scr)	LS or MS	Translate or Cut
			No	frontal	Lecturer	MS	Cut
			No	No	Scr (Wb)	MS or LS	Zoom or Cut
< 120 frames	-	-	-	-	-	Extend current subshot	

based on the focus of a subshot. In our domain, the elements of focus are LCD projected slide, whiteboard, presenter and ROI. In principle, a cut is inserted when switching among different elements to indicate the change of space. For instance, when the focus is switched from the whiteboard to the screen or vice versa, a camera cut is inserted during switch. Similarly, when a frontal pose is recognized, a cut to the presenter is introduced to imply the change of pace and focus. Camera zoom in (or out) may be used when transiting an element from (or to) ROI so as to lessen (or intensify) the degree of impression on the ROI. In addition, camera translation is used to resemble eye movement by following the presenter's intention when the recognized posture is to turn left or right. When switching among different ROIs, camera motion such as zoom and translation is also used to smooth the delivery of focuses. For instance, when a presenter interacts with several focuses on the screen (or whiteboard) with different but coherent and continuous gestures, different types of motion are simulated depending on the recognized gestures and the actual content of ROIs. Basically zoom is used to emphasize the interaction if a circling or pointing gesture is found with an ROI of figure, while translation is use to hint the flow of explanation when the gesture moves from one ROI to another. In addition to cut and camera motion, the gradual changes such as wipe and dissolve are also inserted between shots when slides are flipped to hint the change of topics. The list of possible transitional effects between subshots are outlined in Table II. To prevent the excessive use of transition, the duration of a subshot should not be too short. As seen in Table II, we enforce each subshot to contain at least 120 frames (5 seconds) so that people have enough time to recognize the content.

C. Finite State Machine

We propose to employ FSM (Finite State Machine) for lecture video editing, by interpreting and encoding the aforementioned aesthetic elements. Physically, an FSM is composed of a list of connecting states where each state represents an editing decision. The switching of states, which is determined based upon the input (gesture, posture and focus) of a subshot and the most recent edited subshot, gears the changes of focal length with various transitional effects. Mathematically, an FSM is described as $\mathcal{M} = (s_0, \mathbf{S}, \mathbf{I}, \delta, \mathbf{T})$, where s_0 is the initial state, δ is a function, and $\mathbf{S}$, $\mathbf{I}$, $\mathbf{T}$ are respectively the sets of states, inputs and transitions. The set $\mathbf{S}$ describes the focal length and duration of showing the presenter, whiteboard, screen or ROI. The set $\mathbf{I}$ is composed of the gestures, postures, focuses and time spans of subshots, while the set $\mathbf{T}$ includes the transitional effects such as camera zoom and cut between subshots. The function $\delta(s_i, x) = (s_j, t_k)$ determines the next state $s_j \in \mathbf{S}$ and its associated transitional effect $t_k \in \mathbf{T}$, by taking the current state s_i and the input x of s_j as parameters. The current state s_i encodes the information such as the focal length and the duration of current subshot being shown.

To efficiently encode the editing rules in Table II, our FSM is organized into $\mathbf{S}$ of 14 states. $\mathbf{S}$ is composed of three major components, respectively, the screen (6 states), whiteboard (6 states) and presenter (2 states). We use six states respectively for the screen and the whietboard to rerepresent the combination of three focal lengths and two types of subshot duration (i.e., ≥ 120 frames and < 120 frames). For the presenter state, due to the fact that a presenter is normally shown together with the screen or the whiteboard, we only need two states to represent the combination of MS and duration types. The fourteen states are interconnetced based on the rules imposed on view transitions. For instance, it is impossible to have an edge beween two states with LS and TS respectively. The edges of states specify the set of allowable transitional effects in $\mathbf{T}$ according to Table II. Depending on the input $\mathbf{I}$, the function δ switches the current state s_i to s_j while exhibiting transitional effects when traversing their edge. In brief, during editing, a new video is novelly synthesized by emiting one subshot each time, when stopping at a state to generate an appropriate focal length, and passing through an edge to simulate transitional effects depending on the focus.

Note that although we adopt look-ahead strategy through offline editing, FSM has no capability of looking ahead or generating the current view based on the future subshots. The look-ahead feature indeed comes from the fact that no editing decision is made until a gesture or posture is completed. In real-time applications, an editing decision is made on-the-fly before a gesture is completed. The decision may be noisy if the incomplete gesture is simply unintentional. By offline processing, FSM resembles the ability of look-ahead to advance temporally and take action until a gesture is recognized and the target ROI is identified. A better decision for next subshot is then made depending on the input and current edited subshot.

FSM is also employed in some other systems such as [14] for lecture video editing. In [14], three states are defined to represent the speaker-tracking, audience-tracking and overview cameras respectively. Transiting from one state to another is triggered by events and governed by the transitional probabilities. Compared with [14], our FSM is novel in its expressive ability to trace and show as closely as possible the flow of lecturing, by considering various aspects of subjects and constraints under the camera setup. One interesting note is that FSM can integrate both multimedia content and film aesthetic to emphasize the rhythm of interactions for realistic editing, which, to the best of our knowledge, has not yet been seriously attempted.

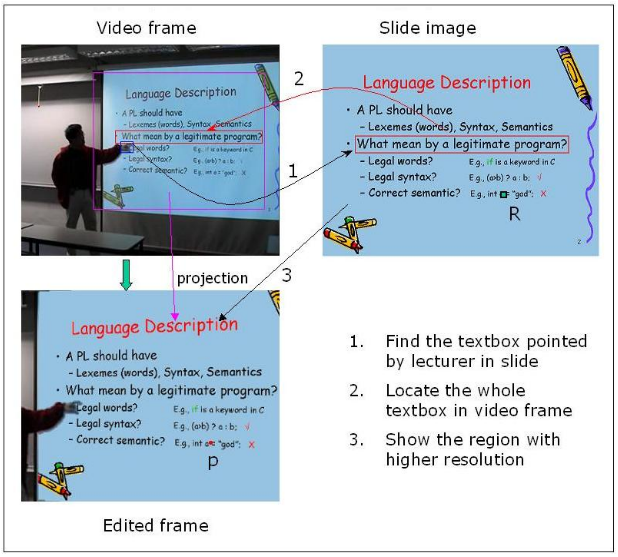
VI. Visibility Enhancement

The visibility of lecture videos, especially the readability of texts is usually the most concerned by the audience. In a typical classroom environment, even if the camera parameters are carefully set, the content of the slides and the handwritings is not easily recognized in the captured videos. The enhancement of the slide and whiteboard images in the edited video is not only important for improving visual quality, but also necessary when camera motion like zoom is performed. In this section, we present our approaches for enhancing the visibilities of video slides and handwritings on the whiteboard. The former is accomplished by the aid of external documents, and the latter is achieved through the color contrast enhancement of text regions in the whiteboard.

A. Visibility Enhancement in Video Slide

Consider the process of video capturing, where the value of a pixel is determined by the sensed energy from a small region in the slide screen. This process is seriously affected by the lighting conditions and the resolution of the camera. The visual quality of the slide image is usually distorted since the light reflected from the screen is normally unequally distributed. As a result, some parts in a slide image may be over-illuminated while the other parts may be under-illuminated. A straightforward way to enhance the visual quality of video slides is to project the content of the external documents to the slide images. In our case, this approach is feasible since each shot is linked to its corresponding electronic slide, and furthermore, the transformation between the projected and the real slides can be computed as described in Section IV-C.

Based on the estimated focuses, we project the instances in the external documents onto the focuses in the edited subshots. The edited ROIs will be displayed with higher resolution while the undesirable effects caused by lighting conditions can be removed. Figure 8 illustrates the detailed procedure when a pointing gesture is recognized. The pointed textbox is initially extracted from the registered slide, by which we can precisely locate the corresponding pointed textbox in the video frame. Because the pointed textbox is aimed to be displayed at the center of the edited video, we can easily compute the transformation to zoom. The slide image to be projected is automatically extracted from the external document. The resolution of the slide image is normally higher than a video frame. In Figure 8, the value of a pixel p in the edited frame is the calculus of a small region R in the slide image. To augment the presenter in front of the edited frame, object segmentation is done prior to the composition. Since we adopt static camera setting, the presenter can be easily detected and tracked over frames by motion segmentation.



1. Find the textbox pointed by lecturer in slide
2. Locate the whole textbox in video frame
3. Show the region with higher resolution

Fig. 8. Projecting electronic slide to edited video.

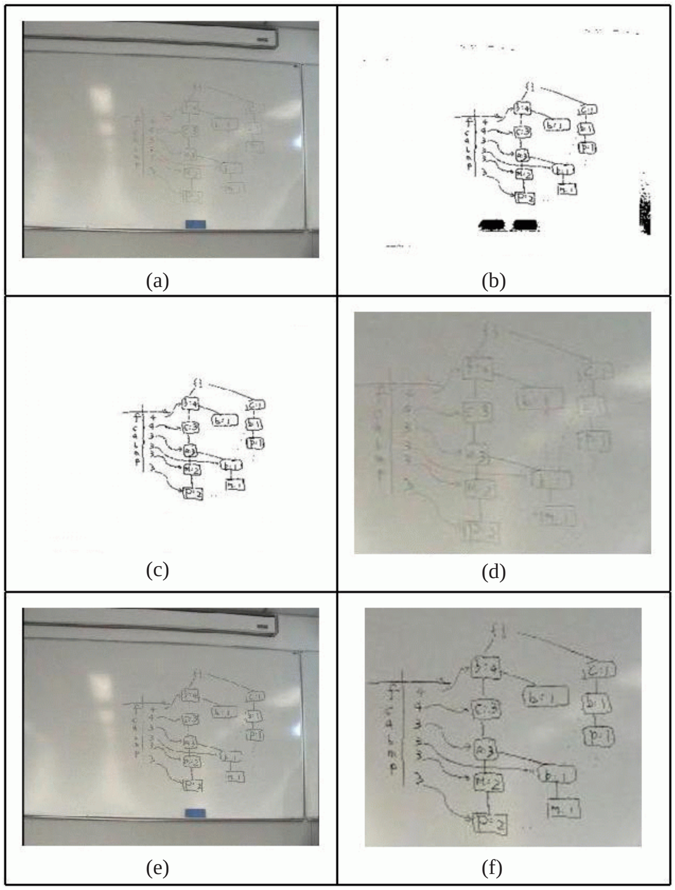
B. Handwriting Enhancement in Whiteboard

As discussed in Section IV-D, due to the lighting condition and reflection from the whiteboard, the handwritten words and diagrams are not easy to detect or recognize. To enhance the visibility, the difference map between the current and the original frames is first computed. The map basically captures the skeletons of handwritings that facilitate text detection. Figures 9(b) and 9(c) show the difference map and the detected text region respectively. The visual quality of handwritings is then enhanced by increasing the color contrast between the text region and the background. Figure 9(f) demonstrates an enhanced and zoomed whiteboard image. For comparison purpose, we show the effects of zooming whiteboard image without contrast enhancement (Figure 9(d)), and of contrast enhancement but without zooming (Figure 9(e)). Comparing (d)-(f) with the original frame (a), we can find (d) is, if not worse, no better than (a). The texts in (e) and (f), however, are easier to recognize. Basically, (e) is still too small to read, and (f) is easier but the characters are slightly blurred after zoom. In the final produced video, we choose between (e) and (f) depending on the editing rhythm.

During whiteboard enhancement, no special step is performed when a presenter is writing on the whiteboard. To prevent occlusion, the handwritten texts are detected and enhanced only after the presenter has completed the writing. The original characters are replaced with the enhanced handwritings. This is possible since editing is conducted in an offline manner.

VII. Experiments and Usability Studies

We conduct experiments on 9-hour videos consisting of 15 presentations given by 10 lecturers and tutors. The presenters



(a) (b) (c) (d) (e) (f)

Fig. 9. Edited whiteboard video by enhancing handwriting visual quality. (a) Original frame; (b) Detected difference; (c) Detected text region; (d) Zoomed image; (e) Contrast-enhanced image; (f) Edited whiteboard image.

include 5 males and 5 females. The presentations are given in the classrooms and seminar rooms of different sizes, layouts and lighting designs. Basically two overview cameras are stationarily mounted. One captures the scene containing the LCD projected screen and the other points toward a whiteboard. This camera setting can be easily amended for classrooms with only a screen or a whiteboard. The external document is not limited to the PowerPoint slide and can be presented in other forms.

A. Recognition Accuracy

For gesture recognition, we use 200 samples for each gesture class to train HMMs. For head pose estimation, we use 300 face images for neural network training. Table III and Table IV show the experimental results of gesture and pose recognition. As seen in the table, most gestures and poses are correctly recognized.

For video text recognition, we achieve approximately 95% of accuracy in recognizing the titles of external documents. The recognition accuracy is indeed not as critical as posture and gesture, since we only need a subset of titles with high similarity (as computed in [27]) to estimate the geometry transformation between video and external document. The transformations in all the 15 videos are correctly estimated, based on the results that the text layouts of videos are seamlessly structured with the aid of external documents.

TABLE III

RESULTS OF GESTURE RECOGNITION(N_g: THE NUMBER OF EACH GESTURE USED IN THE VIDEOS; N_c: THE NUMBER OF EACH GESTURE CORRECTLY RECOGNIZED; RECOGNITION RATE: $acc = \frac{N_c}{N_g}$)

Gesture	Circling	Lining	Pointing
Number of gestures (N_g)	452	637	971
Correctly recognized (N_c)	430	614	909
Recognition rate (acc)	0.951	0.964	0.936
Overall Performance	0.948		

TABLE IV

RESULTS OF HEAD POSE ESTIMATION (N_f: THE NUMBER OF FACES FOR POSE ESTIMATION; N_c: THE NUMBER OF FACES THAT ARE CORRECTLY ESTIMATED FOR EACH POSE; PRECISION $pre = \frac{N_c}{N_f}$)

Pose	Face left	Face right	Face front
Number of faces (N_f)	6670	7039	4985
Correctly estimated (N_c)	6056	6345	4622
Precision (pre)	0.908	0.901	0.927
Overall	0.911		

B. Focus Estimation

Based on the recognized postures and gestures, we conduct experiments to estimate the focus of lecturing. Table V shows the accuracy of focus estimation on the tested videos. We recognize 3 kinds of focuses: *Audience*, *Screen* and *Whiteboard* by gesture and posture. The ground-true focuses are manually labelled according to whether the lecturer is talking about the slide, the whiteboard or to the audience. In Table V, the second and third rows show the accuracies when only gesture or posture is used for focus estimation, and the last row shows the accuracy when gesture and posture are both considered. In general, posture is especially useful when gesture is occluded or absent, while gesture is useful when posture is ambiguous or not seen. When posture is integrated with gesture, the accuracy of estimation is significantly improved as shown in the table. Indeed, when gestures are present, we can estimate not only the simple focuses, but also synchronize the underlying actions with the semantic text layouts extracted from videos. In Table V, the numbers inside the brackets indicate the accuracy of estimating the region of interest (ROI) in the videos. The ground-truth ROIs (e.g., textline, table, figure) are marked manually by watching the unedited videos. The manual judgement is based on gesture and speech. In the experiment, a correct detection means there is a match between the detected and ground-truth ROIs. As shown in this table, when gestures are present, the ROIs can be correctly located most of the time. Posture helps when no gesture is detected, although the ROI estimation is not so exact as gesture.

TABLE V

ACCURACY OF FOCUS ESTIMATION

Focus	Audience	Screen	Whiteboard
Gesture	-	64% (95%)	75% (89%)
Pose	92%	77%	71%
Pose + Gesture	92%	94% (86%)	95% (82%)

C. Usability Evaluation

To evaluate the usability of the proposed system, we conduct a subjective study to compare different video capturing and automatic editing methods. We show five different versions of

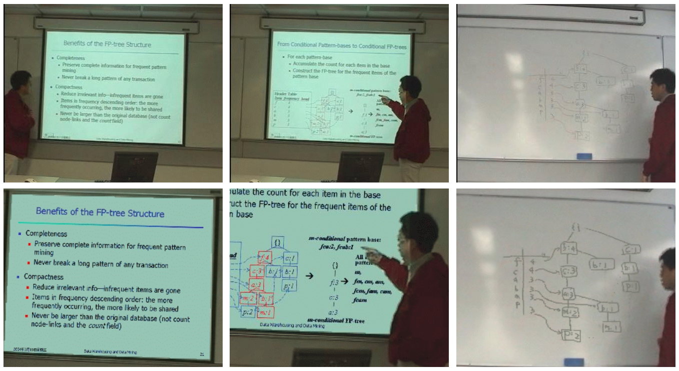


Fig. 10. First row: some snapshots of the original videos; Second row: the corresponding ones in the edited video.

a lecture video to evaluators for assessment (see Table VI). The 1^{st} version is the original videos captured by two static cameras. The next three versions are the videos automatically edited based on the 1^{st} version. The 2^{nd} version is based on the results of lecturer tracking. Cuts are inserted to the video whenever the presenter moves from the screen to the whiteboard or vice versa. The 3^{rd} version is similar to the 2^{nd} except that the visual qualities of slide screen and whiteboard are enhanced. The 4^{th} version is based on our proposed approach presented in this paper. The 5^{th} version is captured by a moving video camera. The camera is operated by a student who has experience in video capturing, and has good knowledge about the content of presentation. Figure 10 shows some frames from the original videos and the corresponding ones from the edited video (4^{th} version).

We invite 20 evaluators consisting of students, professors, and movie artists to grade the six different versions of video. The title of presentation is "Association rule mining" and all participants have engineering or science background. Each video is graded based on five criterions: quality (or readability) of slide and whiteboard images, cinematic effect, concept understanding and enjoyability. The criterion "cinematic" judges the effect and suitability of camera cuttings and motions (the 1^{st} version is not rated since no cinematic effect is involved). The criterion "concept understanding" and "enjoyability" test which kinds of editing styles can make learning and teaching more comprehensive and enjoyable. For each criterion, the evaluators are requested to give a satisfaction score $[0 \sim 10]$, where 0 is the worst, and 10 indicates the most satisfactory score. Each version of video is randomly renamed so that the participants do not know the exact technique we use. The participants can give comments to explain their rating.

Table VI shows the means and standard deviations of the subjective evaluation. In general, almost all participants agree that the 1^{st} version (unedited) is unsatisfactory since the slide screen is small and the handwriting on the whiteboard is unclear. Most participants feel that, by alternating slides and whiteboard, the effect of the 2^{nd} version is better. When the qualities of slides and handwriting on whiteboard are enhanced in the 3^{rd} version, the scores for "understanding" and "enjoyability" are improved as well. Compared with other versions, the visual quality of projected slides in the 4^{th} version is significantly improved, particularly when the *focus* of lecturing is zoomed and the high resolution slide images are shown. Most evaluators agree that the movement and cutting in 4^{th} version make the video less dull and more enjoyable. In particular, the selective focus of slides and whiteboard makes them feel more comfortable and enhances their understanding. For the 5^{th} version, some evaluators comment that it is tiring to watch a video with a camera chasing the presenter throughout the lecture without any cut. The video is less enjoyable and hence affects the understanding of concept. The cameraman had to listen and pay attention to the lecturer, think about what is being and will be talked about, and determine where and how to capture at any moment.

We also conduct a subject evaluation to compare the video edited by our system and a manually edited one. The lecture video is the same as in Table VI. One more video (the 6^{th} version) is manually edited by making shot selection and cuts from the 1^{st}, 2^{nd} and 5^{th} versions. Another group of 10 people are invited to evaluate the 4^{th} and 6^{th} versions and the results are shown in Table VII. As seen in the table, the 6^{th} version gets higher scores for "cinematic" and "enjoyability". At the same time, the qualities of projected slide and whiteboard are acceptable when a manually controlled camera is used. For "cinematic" and "enjoyability", most participants agree that the 4^{th} is still comparable with the 6^{th} version. The 4^{th} version attains better scores for "concept understanding" than the 6^{th} version because the readabilities of the projected slide and whiteboard are thought of as rather important factors to understand a lecture video.

D. Discussions

1) Practical Concerns in Using FSM: Based on the editing rules interpreted by the FSM, a presenter is expected to stand or move in front of a classroom. When a presenter goes a step away and is not captured by the cameras, the whole screen or whiteboard is shown depending on where the presenter moves away from.

In our lecture videos consisting of 10 different presenters who are not given any guidelines when delivering presentations, more than 94% of the focuses can be correctly estimated based on our strategy. In the other 3% of the cases, the focuses shown by the presenters are ambiguous. For example, the presenters do not move the hand away from the screen when turning back to the students. Most of these cases do not cause serious problems since showing either focus (presenter or slide) is acceptable, although the former is definitely better. For the remaining 3%, the focuses estimated from gestures and postures are not the intention of the presenters. When an error occurs in focus estimation, a subshot with inapppopriate content and view could be inserted. Since the content to be shown is determined individually based on the information inside each subshot, the error does not affect the content to be shown in the next subshot. However, the way the next subshot is shown (e.g., focal length selection) may be affected in order to be coherent with the editing rules when adjoining the two subshots. the error may affect the way the next subshot is shown (e.g., focal length selection), but not the content.

2) Close vs. Distant Interaction: Our approach offers accurate focus estimation when the defined gestures interact on top

TABLE VI
SUBJECTIVE EVALUATION OF EDITED AND UNEDITED VIDEOS

Method		Quality of projected slide		Quality of whiteboard		Cinematic		Concept understanding		Enjoyability	
		Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv
1	Original	4.89	1.79	3.47	1.26	-	-	5.00	1.34	4.63	1.74
2	Motion	5.00	1.80	3.74	1.52	5.63	1.57	5.42	1.39	4.84	1.83
3	Motion + visual	5.84	1.80	5.47	1.90	5.74	1.56	6.11	1.33	5.68	1.57
4	Gesture + pose + visual	**8.05**	1.27	**6.37**	1.21	**6.53**	1.43	**7.32**	1.01	**6.84**	1.57
5	Manually moving camera	6.18	1.97	3.37	2.27	5.07	1.66	5.47	1.54	5.16	1.42

TABLE VII
SUBJECTIVE EVALUATION OF VIDEOS EDITED BY OUR SYSTEM AND EDITED MANUALLY

Method		Quality of projected slide		Quality of whiteboard		Cinematic		Concept understanding		Enjoyability	
		Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv	Mean	Stdv
4	Our approach	8.21	1.27	6.34	1.26	6.61	1.50	7.52	1.19	6.78	1.53
6	Manually edited	6.65	1.61	5.46	1.71	6.90	1.52	6.27	1.91	6.88	1.64

of the slide or whiteboard. Since no 3D gesture information is analyzed, unexpected cases may happen if a presenter points from a distance far away from the targeted ROI. For instance, when a presenter stands on the whiteboard side and points to the projected slide, a false positive ROI may be selected if the gesture happens to interact with an ROI on the whiteboard. By allowing a gesture to interact close enough to an ROI, part of the ROI could be occluded. The partial occlusion, nevertheless, is not a serious problem since our cameras are faced directly to the screen and the whiteboard, and it is expected that a presenter stands beside the ROI when making gestures. In view of the aforementioned issues, certain guidelines could be useful if provided to the presenters, but we do not request so for the experiments presented in this paper.

3) Focal Length Selection and Visual Enhancement: There are two factors for the selection of focal length. Firstly, the resolution should be high enough so that textual elements are readable. Secondly, the complete scene of the presenter, slide and their interaction should be displayed to enhance understanding and impression. However, limited by the resolution, there is a tradeoff between these two factors. In other words, it is not feasible to show everything in one frame. In our work, to smooth the rhythm of display, when a gesture is identified, we begin by showing a medium view which includes the presenter, the slide and their interaction. A tight view then follows by highlighting the ROI with higher resolution. At this view, although the presenter may not be seen, the gesture which represents the interaction is still visible. In view selection, basically we process ROIs to guarantee good readability before utilizing editing rules to drive the rhythm of display.

4) Camera Setting: The number of required cameras largely depends on the requirement and classroom setting. For instance, a fixed overview camera can be set up to provide establishing shot, while a tracking camera can be used to follow the presenter. Our current approach adopts the simplest setting of two static cameras in order not to overload a classroom with hardware. The setting indeed fits well for most classrooms with one screen and one whiteboard. More cameras can be accommodated on top of our current setting, by adding a few more editing rules to the FSM.

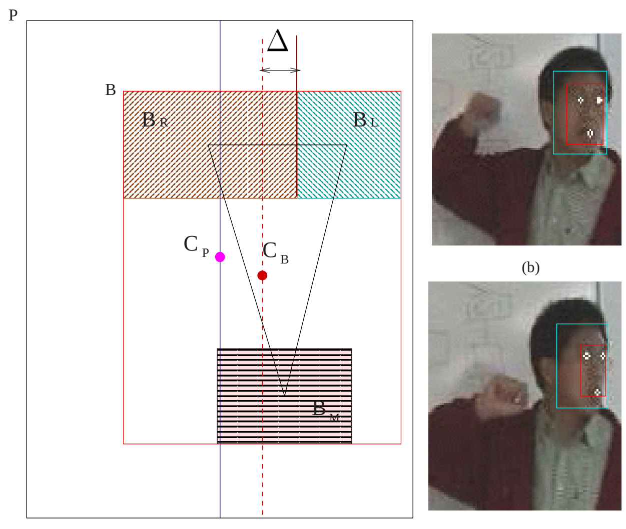
5) Editing with Content Awareness: While our idea of estimating the focus at any given moment for making wise editing decision appears interesting, there are other systems such as AutoAuditorium [30] that produce videos with good quality, although less detailed content analysis is involved. The existing systems [14], [16], [21], [30], which are designed for making prompt editing decision for real-time broadcast, can still perform satisfactorily with the setting of active cameras despite simple visual analysis. Although the interaction between a presenter and slide may not be properly displayed due to the lack of content understanding, the quality of videos is still generally acceptable. Our work in this paper indeed is similar in spirit to the research efforts in [14], [16], [21] and [30], but we assume a simpler and convenient setting with two static cameras being pointed towards a whiteboard and an LCD projected screen. With this set up, the chance of producing visually engaging videos is indeed limited without the detailed content analysis. As studied in our user evaluation, the videos edited by simply switching shots between whiteboard and slide are not better than the videos capable of following lecturing focus. Due to the fact that most systems assume different classroom and camera settings, it is not flexible to compare our approach with other existing systems. Nevertheless, through our subjective evaluation on the various versions of videos edited based on static cameras, the evaluators commonly agree that the videos which emphasize interaction and focus are the best option.

VIII. CONCLUSION

We have presented our approach to tracing the focus and flow of lecturing for video editing, by integrating three visual cues: postures, gestures and texts. A finite state machine is proposed by integrating these cues with cinematic rules and idioms in an automatic and systematical manner. Our contribution in terms of editing aspect lies in the exploitation of dynamic and static feature interaction for more realistic attention based editing, under the consideration of aesthetic elements. The dynamic features refer to the gradual changes of postures and gestures, while the static features refer to the semantic instances that are automatically structured for focus representation. Although visual cue recognition has been intensively investigated in the current literature, specialized content analysis is still required for specific video domain such as in modern classroom environment. In this paper, we have identified several important challenges in recognition, and correspondingly, proposed feasible techniques to tackle the difficulties. Overall, encouraging results are obtained through experiments and subjective evaluation. Several challenges remain for our current work, for instance, the super-resolution reconstruction of handwritten words on the whiteboard, and the modeling of the contextual relationship of different gestures for more effective editing. Besides posture and gesture, audio cues, particularly speech, are also important factors to be considered in editing. The analysis and fusion of audio-visual cues for lecture video editing can be another crucial issue that needs to be addressed in future.

Appendix

Based on the hierarchical representation of skin clusters as shown in Figure 5, we detect and validate frontal faces by a two-phase verification procedure. In the first phase, an ellipse is fitted to each cluster to detect frontal face. By considering the camera setting and the general shape of the human face, we heuristically and statistically exclude false candidates based on the fitness confidence, ellipse size, and density of skin pixels. Figure 5 shows that two frontal faces are successfully fitted by ellipses, while the remainder are rejected as false matches. In the second phase, facial features (eyes and mouths) are located for further verification. Initially, the skin pixels of potential frontal faces are normalized by compensating for the unevenness and variety of illumination. Then morphological filters (*open* and *close*) are applied to highlight the facial features [26].



(a)

(b)

(c)

Fig. 11. Facial feature tracking: (a) adaptive face template; (b) and (c) located facial features.

We use a face template (see Figure 11(a)) to locate the facial features on the filtered image. A face template is adaptively generated based on the given candidate region and its morphological filtered image $\mathcal{R}_{pv}$. The template is basically formed by two bounding boxes P (Figure 11(a)) and B (Figure 11(b)) that minimally enclose a face region and the peaks and valleys of $\mathcal{R}_{pv}$ respectively. Three search regions B_R, B_L and B_M are adaptively defined, based on the orientation of a face induced by P and B, to effectively locate the eyes and mouth. The search regions are determined by the centers C_P and C_B of P and B respectively. Let $C_P = (C_Px, C_Py)$ and $C_B = (C_Bx, C_By)$, we define a deviation term Δ as $\Delta = \alpha \frac{C_Bx - C_Px}{W_P} W_B$, where W_P and W_B are respectively the width of P and B. The parameter $\alpha = 1.1$ is an empirical constant estimated from a face database of 300 images in different head poses. The term Δ estimates the degree of deviation from a frontal face. Ideally, $\Delta = 0$ if a lecturer directly faces to the front. Based on Δ, the width of B_R and B_L are respectively $\frac{W_B}{2} + \Delta$ and $\frac{W_B}{2} - \Delta$ as shown in Figure 11(a). When a face turns to left, for instance, the width of B_L will be relatively narrower than B_R. B_M lies between the centers of B_L and B_R, and its width is $\frac{W_B}{2}$. B_R, B_L and B_M are the regions where the facial features (eyes and mouth) are expected to lie in.

In the filtered image $\mathcal{F}$, facial features are highlighted and their centers show the highest values. We get the three regions (B_R, B_L and B_M) in $\mathcal{F}$ by fitting it with the face template. In each search region, five local maximum points are selected from $\mathcal{F}$ as candidate facial features. Three among these points, one from each region, that fit the triangle formed by the centers of B_R, B_L and B_M best (see Figure 11(a)) are selected as the locations of facial features. Figures 11(b) and (c) show two examples of locating facial features. In the detection phase, all the candidate skin clusters are tested and the clusters not showing salient facial features are excluded.

Once a face is detected, its template is used to continuously track the facial features in the following frames. The tracking is based on skin-color detection and the continuous update of face template by P, B and $\mathcal{F}$. A smoothness constraint inferred from the previous feature locations is imposed on the face template to ensure the robustness of tracking.

We extract parameters from the face template (Figure 11(a)) for head pose estimation by neural network. Let E_L, E_R and M be the centers of the detected facial features. The relative positions among them and the whole face are used for pose estimation. The parameters are generated from the set $\{C_P, C_B, E_L, E_R, M\}$. Any pair from the set gives two features: the length and direction of the line connecting them. Four additional parameters are the ratio of widths and heights of two rectangles P, B: $\frac{W_P}{H_P}$, $\frac{W_B}{H_B}$, $\frac{W_P}{W_B}$, $\frac{H_P}{H_B}$, where H_P and H_B are the height of P and B respectively. In total, there are 24 features altogether. We use these features to train a neural network for classification.

Acknowledgement

The work described in this paper was supported in part by the grants SSRI99/00.EG11, DAG01/02.EG16, HIA01/02.EG04, and a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 118906).

References

[1] G. D. Abowd, C. G. Atkeson, A. Feinstein, and C. Hmelo, "Teaching and Learning as Multimedia Authoring: The Classroom 2000 Project," *ACM Multimedia Conf.*, pp. 187-198, 2000.

[2] D. Bordwell and K. Thompson, *Film Art: An Introduction*, Random House, 1986.

[3] M. Chen, "Visualizing the Pulse of a Classroom," *ACM Multimedia Conf.*, 2003.

[4] M. Chen, "Achieving Effective Floor Control with a Low-Bandwidth Gesture-Sensitive Videoconferencing System," *ACM Multimedia Conf.*,2002.

[5] B. Erol, J. J. Hull, and D. S. Lee, "Linking Multimedia Presentations with their Symbolic Source Documents: Algorithm and Applications," *ACM Multimedia*, 2003.

[6] M. Gleicher and J. Masanz, “Towards Virtual Videography”, *ACM Multimedia Conf.*, 2000.
[7] M. Gleicher, R. M. Heck, and M. N. Wallick, “A Framework for Virtual Videography”, *Int. Symp. on Smart Graphics*, 2002.
[8] L. He, E. Sanocki, A. Gupta, and J. Grudin, “Auto-Summarization of Audio-Video Presentations,” *ACM Multimedia Conf.*, 1999.
[9] L. He and Zhengyou Zhang. “Note-Taking with a Camera: Whiteboard Scanning and Image Enhancement,” *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2004.
[10] R. L. Hsu, M. Abdel-Mottaleb, and A. Jain, “Face Detection in Color Images,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696-706, May, 2002.
[11] S. X. Ju, M. J. Black, S. Minneman, and D. Kimber, “Summarization of Videotaped Presentations: Automatic Analysis of Motion and Gesture,” *IEEE Transactions on Circuits and Systems for Video Technology*, 1998.
[12] T. Liu, R. Hjelsvold, and J. R. Kender, “Analysis and Enhancement of Videos of Electronic Slide Presentations,” *Int. Conf. on Multimedia & Expo*, 2002.
[13] T. Liu and J. R. Kender, “Spatio-temporal Semantic Grouping of Instructional Video Content,” *Int. Conf. on Image and Video Retrieval*, 2003.
[14] Q. Liu, Y. Rui, A. Gupta, and J. J. Cadiz, “Automatic Camera Management for Lecture Room Environment,” *Int. Conf. on Human Factors in Computing Systems*, 2001.
[15] T. F. S. -Mahmood and S. Srinivasan, “Detecting Topical events in digital video,” *ACM Multimedia Conf.*, 2000.
[16] E. Machnicki and L. Rowe, “Virtual Director: Automating a Webcast,” *Multimedia Computing and Networking*, 2002.
[17] J. Martin and J. B. Durand, “Automatic Gesture Recognition Using Hidden Markov Models,” *Int. Conf. Automatic Face and Gesture Recognition*, 2000.
[18] Y. Matsuo, M. Amano, and K. Uehara, “Mining Video Editing Rules in Video Streams,” *ACM Multimedia Conf.*, 2002.
[19] S. Mukhopadhyay and B. Smith, “Passive Capture and Structuring of Lectures,” *ACM Multimedia Conf.*, 1999.
[20] C. W. Ngo, T. C. Pong, and T. S. Huang, “Detection of Slide Transition for Topic Indexing,” *Int. Conf. on Multimedia & Expo*, 2002.
[21] Masaki Onishi and Kunio Fukunaga, “Shooting the Lecture Scene Using Computer-controlled Cameras Based on Situation Understanding and Evaluation of Video Images,” *Int. Conf. on Pattern Recognition*, 2004.
[22] J. Kovac, P. Peer, and F. Solina, “Human skin colour clustering for face detection,” *Int. Conf. on Computer as a Tool*, 2003.
[23] D. Q. Phung, S. Venkatesh, and C. Dorai, “Hierarchical Topical Segmentation in Instructional Films based on Cinematic Expressive Functions,” *ACM Multimedia Conf.*, 2003.
[24] L. A. Rowe and J. M. Gonzlez,“BMRC Lecture Browser,” http://bmrc.berkeley.edu/frame/projects/lb/index.html
[25] Y. Rui, A. Gupta, and J. Grudin,“Videography for Telepresentations”, *Int. Conf. on Human Factors in Computing Systems*, 2003.
[26] Luis. J, “Active Face and Feature Tracking,” *Int. Conf. on Image Analysis and Processing*, 1999.
[27] F. Wang, C. W. Ngo, and T. C. Pong, “Synchronization of Lecture Videos and Electronic Slides by Video Text Analysis,” *ACM Multimedia Conf.*, 2003.
[28] F. Wang, C. W. Ngo, and T. C. Pong, “Gesture Tracking and Recognition for Lecture Video Editing,” *Int. Conf. on Pattern Recognition*, 2004.
[29] F. Wang, C. W. Ngo, and T. C. Pong, “Exploiting Self-Adaptive Posture-based Focus Estimation for Lecture Video Editing”, *ACM Multimedia Conf.*, 2005.
[30] *IBM Auto Auditorium System*, www.autoauditorium.com.

Feng Wang received his BSc in Computer Science from Fudan University, Shanghai, China, in 2001. Then he enrolled in the Hong Kong University of Science and Technology, where he is currently a PhD student in the Dept. of Computer Science and Engineering. Mr. Wang’s PhD thesis topic is video content analysis and its applications for multimedia authoring of presentations. His research interests include multimedia computing, pattern recognition and IT in education.

Chong-Wah Ngo (M’02) received his Ph.D in Computer Science from the Hong Kong University of Science & Technology (HKUST) in 2000. He received his MSc and BSc, both in Computer Engineering, from Nanyang Technological University of Singapore in 1996 and 1994 respectively.

Before joining City University of Hong Kong as assistant professor in Computer Science department in 2002, he was a postdoctoral scholar in Beckman Institute of University of Illinois in Urbana- Champaign (UIUC). He was also a visiting researcher of Microsoft Research Asia in 2002. CW Ngo’s research interests include video computing, multimedia information retrieval, data mining and pattern recognition.

Ting-Chuen Pong received his Ph.D. in Computer Science from Virginia Polytechnic Institute and State University, USA in 1984. He joined the University of Minnesota - Minneapolis in the US as an Assistant Professor of Computer Science in 1984 and was promoted to Associate Professor in 1990. In 1991, he joined the Hong Kong University of Science & Technology, where he is currently a Professor of Computer Science and Associate Vice-President for Academic Affairs. He was an Associate Dean of Engineering at HKUST from 1999 to 2002, Director of the Sino Software Research Institute from 1995 to 2000, and Head of the W3C Office in Hong Kong from 2000 to 2003. Dr. Pong is a recipient of the HKUST Excellence in Teaching Innovation Award in 2001.

Dr. Pong’s research interests include computer vision, image processing, pattern recognition, multimedia computer, and IT in Education. He is a recipient of the Annual Pattern Recognition Society Award in 1990 and Honorable Mention Award in 1986. He has served as Program Co-Chair of the Web and Education Track of the Tenth International World Wide Web Conference in 2001, the Third Asian Conference on Computer Vision in 1998, and the Third International Computer Science Conference in 1995.