Ying Li, Shih-Hung Lee, Chia-Hung Yeh, and C.-C. Jay Kuo

Techniques for Movie Content Analysis and Skimming

Tutorial and overview on video abstraction techniques

The use of video abstraction techniques for movie content analysis with applications in fast content browsing, skimming, transmission, and retrieval is examined in this article. First, we provide a tutorial on abstraction techniques for generic videos, followed by an extensive survey on state-of-the-art techniques for feature film skimming. Then we present our recent work on the development of a movie skimming system using audiovisual tempo analysis and specific cinematic rules. Some preliminary experimental results are reported to validate the proposed skimming system.

Video abstraction is a technique that abstracts video content and represents it in a compact manner. There are basically two types of video abstraction: video summarization and video skimming. Video summarization is a process that selects a set of salient images called *key frames* to represent the video content. Video skimming represents the original video in the form of a short video clip. Video abstraction forms a key ingredient

1053-5888/06/$20.00©2006IEEE

in a practical video content management system, as generated key frames and skims provide users an efficient way to browse or search video content. With the proliferation of digital video, this process will become an indispensable component to any practical content management system. A video summary can be displayed without the worry of timing issues. Moreover, extracted key frames could be used for content indexing and retrieval. However, from the viewpoint of user, a video skim may provide a more attractive choice since it contains audio and motion information that makes the abstraction more natural, interesting and informative.

Film is an art form that offers a practical, environmental, pictorial, dramatic, narrative, and musical medium to convey a story [1]. Although it can be viewed as a type of generic video, complex film editing techniques, such as the selecting, ordering, and timing of shots; the rate of cutting; and the editing of soundtracks, are required to produce a successful movie. Consequently, all of these special features need to be taken into account for better content analysis, understanding, and management. There has been recent work on movie content abstraction, which produces a static storyboard, a summary sequence, or a highlight [2]. A summary sequence provides users a small taste of the entire video, while a highlight contains only the content that may appear interesting to viewers such as the movie trailer. Despite the large amount of research on generic video abstraction, it remains a challenge to generate meaningful movie abstracts due to the special filming and editing characteristics.

SURVEY ON VIDEO ABSTRACTION

VIDEO SUMMARIZATION

Based on the way a key frame is extracted, existing work in this area can be categorized into three classes: sampling based, shot based, and segment based. Most of the earlier summarization work belongs to the sampling-based class, where key frames were either randomly chosen or uniformly sampled from the original video. The video magnifier [3] and the MiniVideo [4] systems are two examples. This approach is the simplest way to extract key frames, yet such an arrangement may fail to capture the real video content, especially when it is highly dynamic.

More sophisticated work has been done to extract key frames by adapting to dynamic video content. Since a shot is defined as a video segment taken from a continuous period, a natural and straightforward way is to extract one or more key frames from each shot using low-level features such as color and motion. A typical approach was proposed in [5], where key frames were extracted in a sequential fashion via thresholding. More sophisticated schemes based on color clustering, global motion, or gesture analysis could be found in [6]–[8]. Realizing that regular key frames cannot represent the underlying video dynamics effectively, researchers have looked for an alternative way to represent the shot content using a synthesized panoramic image called the *mosaic*. Along this direction, various types of mosaics such as static background mosaics and synopsis mosaics have been proposed in [9] and [10]. An interchangeable use of regular key frames and mosaic images has also been studied in [11]. Some other work applied mathematical tools to the summarization process. For instance, a video content could be represented by a feature curve in a high-dimensional feature space with key frames corresponding to the curvature points [12]. One drawback of the shot-based key frame extraction approach is that it does not scale up well for long video.

More recently, efforts have been made in extracting key frames at a higher unit level, referred to as the segment level. Various clustering-based extraction schemes have been proposed. In these schemes, segments are first generated from frame clustering and then the frames that are closest to the centroid of each qualified segment are chosen as key frames [13], [14]. Yeung and Yeo [15] reported their work on video summarization at the scene level. Based on a detected shot structure, they classified all shots into a group of clusters using a time-constrained clustering algorithm, and then extracted meaningful story units (or scenes) such as dialogue and action. Next, representative images (R-images) were selected for each story unit to represent its component shot clusters. All extracted R-images of a story unit were resized and organized into a single regular-sized image following a predefined visual layout called the *poster*. Other schemes based on sophisticated temporal frame sampling [16], hierarchical frame clustering [17], fuzzy classification [18], singular value decomposition, and principle component analysis techniques have been tried with some encouraging results.

VIDEO SKIMMING

A three-layer system diagram for video skimming is shown in Figure 1. In this system, low-level features are extracted and preprocessing tasks (such as commercial break and/or shot detection) are performed at the first layer. At the second layer, mid- to high-level semantic features are derived, which can be accomplished using techniques such as face detection, audio classification, video text recognition, and scene or event detection. The third layer assembles clips that possess user-desired length and content into the final abstract.

Previous work on video skimming can be classified into two categories: summary oriented and highlight oriented. A summary-oriented skim keeps the essential part of the original video and provides users a summarized version [19]. In contrast, the highlight-oriented skim only comprises a few interesting parts of the original video. Movie trailers and highlights of sports are examples of this skim type [20].

Defining which video segments to be highlighted is a subjective and difficult process. It is also challenging to map human perception into an automated abstraction process. Hence, most current video skimming work is summary-oriented. One straightforward approach is to compress the original video by speeding up the playback. As pointed out by Omoigui et al. [21], a video program could be watched in a fast playback mode without distinct pitch distortion using a time compression technology. Similar work was also reported by Amir et al. [22], where an

audio time-scale modification scheme was applied. However, these techniques only allow a maximum time compression of 1.5–2.5 depending on the rate of speech. Once the compression factor goes beyond this range, the speech quality becomes quite poor. Targeting at generating short synopses for generic video, the Informedia Project [23] concatenates audio and video segments that contain preextracted text keywords to form the skim. Special attention was also given to predefined events such as the presence of human faces and camera motion. Without relying on text cues, Nam and Tewfik [24] generated skims based on a dynamic sampling scheme. Specifically, a video source was first decomposed into a sequence of subshots. Each subshot was then assigned a motion intensity index. Next, all indices were quantized into predefined bins, where each bin possessed a unique sampling rate. Finally, key frames were sampled from each subshot based on the assigned rate. During the skim playback, linear interpolation was performed to provide users a moving storyboard. Similar constructions of skims based on pregenerated key frames was also presented in [25] and [26].

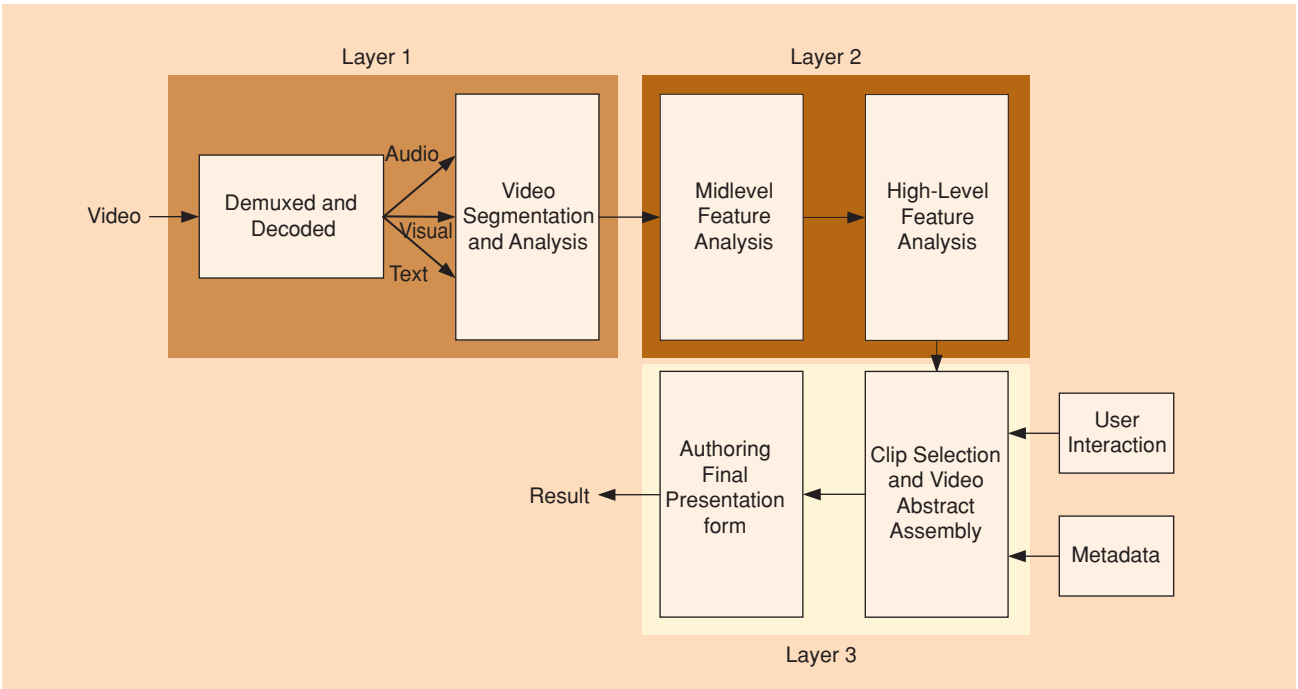
More recently, research on generating skims for domain-specific video data has been reported using some special features. For example, the VidSum project [27] applied a presentation structure, which was designed for their regular weekly forum, to assist in mapping low-level signal events to semantically meaningful events. These events were then assembled to form the summary. He et al. [28] reported their summarization work on talks. Special knowledge of the presentation was utilized, which included the pitch and the pause information, the slide transition points, as well as the information on access patterns of previous users. A detailed user study showed that most of informative parts of original presentations have been well preserved although computer-generated summaries were less coherent than manually generated ones. In [19], a skimming system for broadcast news was presented by exploiting its hierarchical content structure. Given a news video, it filtered out commercials using audio cues, and then detected anchor persons using Gaussian mixture models. Next, since a news story is usually led and summarized by an anchor person, the skim can be constructed by gluing all video parts that contain anchor persons. Finally, there have been some research efforts on generating skims for sports videos based on the identification of exciting highlights such as football touchdowns and soccer goals.

RECENT DEVELOPMENT ON MOVIE SKIMMING TECHNIQUES

Since a video skim appears more natural and attractive to viewers, most recent work on movie abstraction focuses on the generation of a short synopsis of a long feature film. Nevertheless, most existing skimming systems that are built upon key frame-based summarization have two drawbacks: the discontinuity of embedded semantics and the lack of audio content. Moreover, the movie structural information and its distinct storyline have not been well exploited in most previous work. This important context information should help generate more meaningful skims. In this section, a brief introduction to the fundamental story structure of movies is given. The state-of-the-art work in movie content skimming will be described as well.

FUNDAMENTAL STRUCTURE OF MOVIES

Most feature films consist of three distinct parts: the beginning (exposition), the middle (conflict), and the end (resolution). The beginning introduces basic facts such as main characters and their relationships, which are needed to establish the story. Then, the



[FIG1] A three-layer system diagram for video abstraction.

[TABLE1] TYPICAL STORY UNITS OF A MOVIE.

UNIT	TYPICAL LENGTH	DEFINITION
BROAD SCENE	10–15 MIN	COMPRISES SHOTS WITHIN THE SAME THEME (LARGE STORY ELEMENT)
NARROW SCENE	2–5 MIN	COMPRISES SHOTS THAT ARE TEMPORALLY ADJACENT OR UNDER THE SAME PHYSICAL SETTING (SMALL STORY UNIT)
SHOT	3–10 S	A CONTINUOUS RECORD
FRAME	1/30 S	ONE STILL IMAGE

story develops when conflicts among main characters occur, which are usually expressed in the form of an external action or an internal emotional struggle. A conflict could be plotted through several larger narrative elements, where each element creates a local climax to maintain audience's attention. Finally, the resolution resolves the conflict by wrapping up all incomplete story elements.

A movie story could be structured narratively via several large narrative elements. Correspondingly, directors will design appropriate shooting scripts that define necessary scenes (a smaller story unit) to present these elements. As a result, a typical movie could be sectioned at different resolution levels corresponding to different story units. To present a story well, the director, the cinematographer, and the editor need to carefully structure the building blocks of the following four filming elements: 1) the plots, characters, and dialogues in a script; 2) the instruments, notes, and melody in music; 3) the volume, bass, tremble, and sound effects; and 4) the basic visual components (visual effects). It is thus possible to represent a movie in a tree structure that comprises its building components (units) at different resolution levels. Table 1 shows a list of such units, which are constantly referred to by various researchers, with their typical lengths and definitions.

UTILITY FUNCTION-BASED SKIMMING

The VAbstract system [29] is probably the earliest movie skimming system that identifies characteristic video segments such as those containing leading actors, dialog, gunfire, and explosions to form a movie trailer. In this system, a movie was first partitioned into segments of equal length and then one scene (such as the one with dialog, high motion, or high contrast) was extracted from each segment except for the last part of the movie. Finally, all selected scenes were organized in their original temporal order to reduce the possibility of a misleading context change.

Luo and Fan [30] proposed a method that first identified salient objects and mapped principal video shots to certain medical concepts. Next, each principal video shot was given a weight based on its structure (the elements of the shot), assigned medical concept (e.g., lecture or surgery), contained salient objects, and length. Finally, the skim was formed by selecting shots with the highest weight to the one with the lowest weight until their total length reaches the expected length.

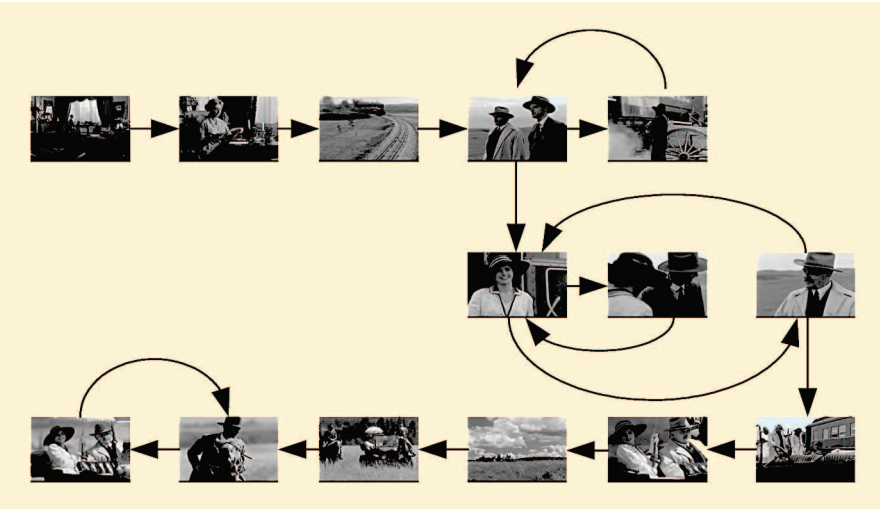
One way to determine the important part of a video object is to exploit user attention models as done in [31]. Various visual attention models were built to capture features such as motion, face, and camera attention. For instance, the motion attention model was used to capture human motion, while the static attention model was for measuring the attention on a static background region. Several patterns involving camera motion were also investigated and used to build the camera attention model. Furthermore, two audio attention models (i.e., the audio saliency model and the speech/music model) were adopted. The extracted attention information was then exploited to identify important shots that form the final video summary.

Another idea, presented in [32], was to segment the video into computable scenes that exhibited consistency in chromaticity, lighting, and sound. The Kolmogorov complexity of a shot that gave the minimum time required for its comprehension was measured. Finally, the beginning and the ending parts of a scene were selected to form the skim as they contained most of its essential information according to the film grammar.

STRUCTURE-BASED SKIMMING

Although a skim could be easily constructed using a predefined utility function, its result is unpredictable. To address this issue, another skimming approach exploring the hierarchical story structure of movies has been investigated. The resulting methods can be categorized into two classes: skimming by the editing style and skimming by multilevel tempo analysis.

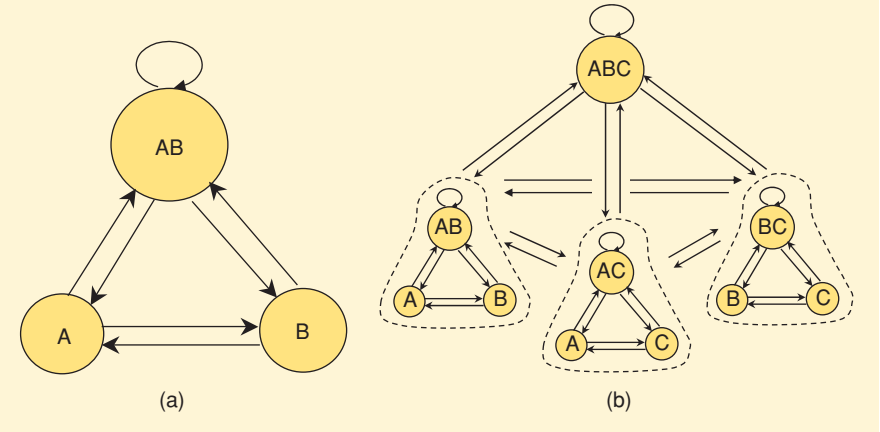
The hierarchical story structure of a movie in the form of frames, shots, scenes, acts, or events can be extracted to serve as a basis for skimming. Shot detection has been extensively studied for more than a decade, and many algorithms have been reported [33]. The concept of a scene refers to a relatively complete video paragraph with coherent semantic meaning. Scene detection demands the integration of multiple media sources. For example, visually



[FIG2] Representing a movie segment using the scene transition graph. "Legends of the Fall, 1994 TriStar Pictures Inc. All rights reserved. Courtesy of Sony Pictures Entertainment.

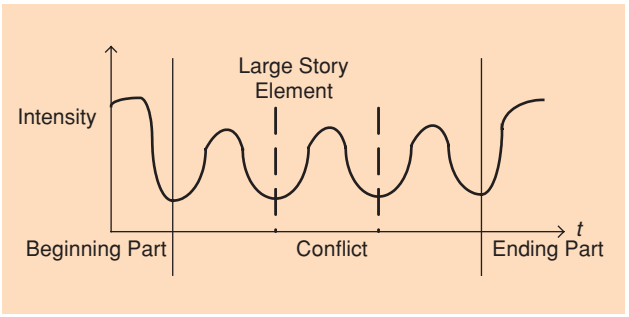
similar and temporally adjacent shots were first clustered into scenes, and then organized into a scene transition graph (STG) in [34]. An example of STG is shown in Figure 2, which is constructed from a partial segment of the movie *The Legends of the Fall*.

A set of heuristic rules was developed by Li and Kuo [2] to detect movie scenes using audio classification and visual analysis tools. For example, temporally adjacent shots should be grouped into one scene if they share the same background music or have the same level of background noise. Certain film editing rules learned from [35] were adopted to identify dialog scenes. Two dialog models are shown in Figure 3, where each node represents a shot containing the indicated speakers, and arrows are used to denote the transitions between shots. These models can be used to recognize the two-speaker and multispeaker scenes by identifying specific shot repetition patterns.



[FIG3] **(a) The two-speaker dialog model (speakers A and B) and (b) the multiple-speaker dialog model (speakers A, B, and C).**

Similar work employing audio and visual cues to detect scenes was reported in [36], where two types of computable scenes (i.e., N-type and M-type) were extracted. The N-type scenes were characterized by consistency of both audio and visual information while the M-type scenes were characterized by consistency in audio but dynamics in visual information. The N-type scenes can be further classified into pure dialog, progressive and hybrid categories.



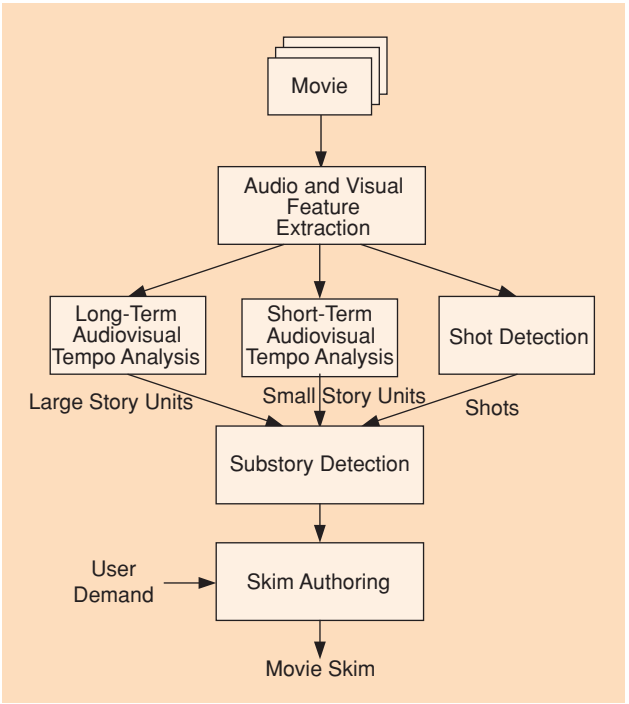
[FIG4] **A typical story intensity map of a movie.**

PROPOSED MOVIE SKIMMING SYSTEM

SYSTEM OVERVIEW

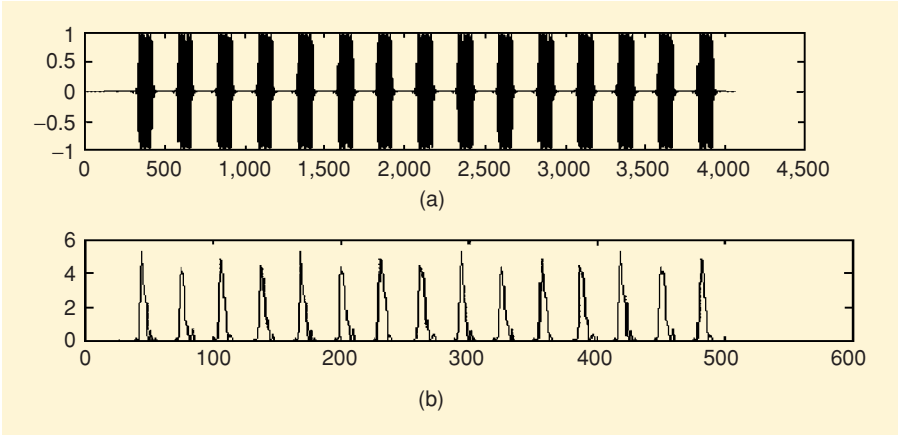
According to [37], the audiovisual structure of a movie is related to its story structure through its intensity map owing to the principle of contrast and affinity. That is, the greater the contrast/affinity in a visual component, the more the visual intensity increases or decreases. Consequently, the director carefully constructs the intensity map to match the storyline. Because this principle applies to different levels of a movie structure, it can be used to detect story units at different scales based on tempo analysis. A typical story intensity map of a movie is shown in Figure 4.

Adams et al. [38] proposed a scheme to analyze the scene content such as the dialog or the chase by extracting film rhythms based on the extracted video tempo. The tempo was measured in terms of the motion information and the shot length [39]. This scheme was however too simple to be generally applicable. A more sophisticated scheme should employ multiple media cues for tempo extraction, subsequently identify all story units at various resolution levels, and organize important or user-desired ones into a skim. This idea has motivated our research on an intelligent movie skimming system as depicted in Figure 5. This movie skimming system is able to generate a

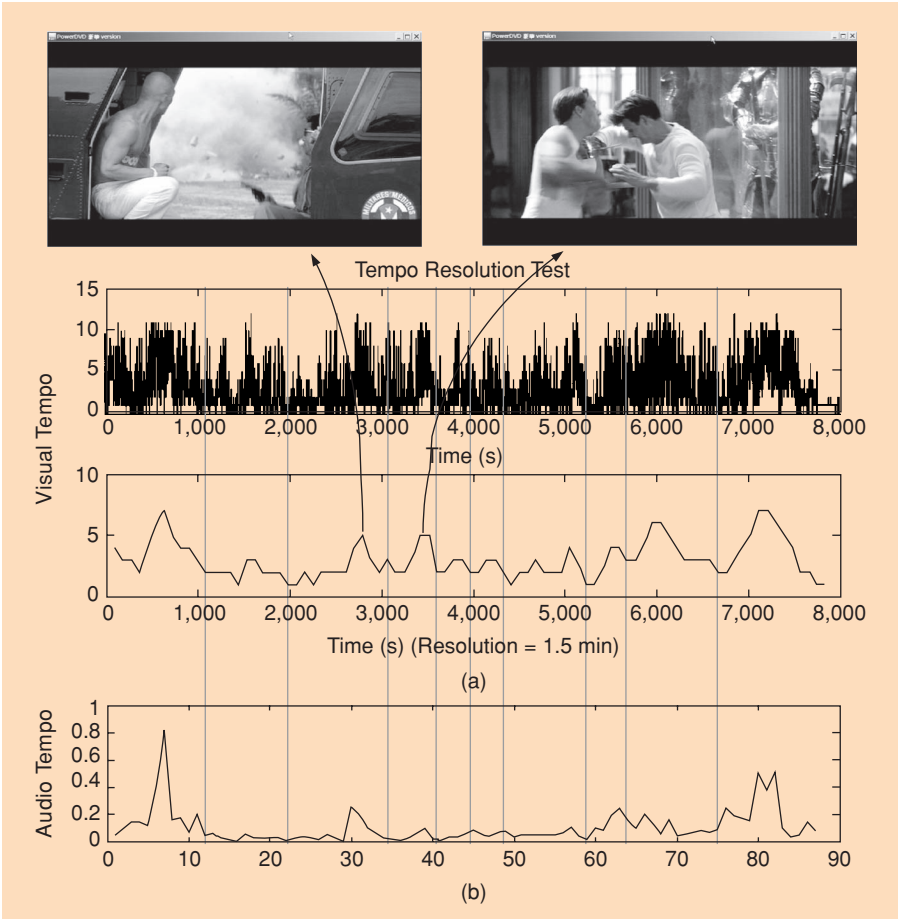


[FIG5] **The proposed movie skimming system.**

user-preferred skim based on extracted movie story structures via audiovisual tempo analysis. With any given feature film, long-term and short-term video tempo analysis tasks are performed to extract large and small story units. Substories are identified from each small story unit based on the scene transition graph. Finally, the skim is generated based on a set of authoring criteria as well as user requests.



[FIG6] (a) A music signal and (b) its detected onsets.



[FIG7] (a) The long-term visual tempo and (b) the long-term audio tempo for movie *Die Another Day*.

AUDIO AND VISUAL FEATURE EXTRACTION

Two audio features, namely, short-term energy and frequency of onset, which indicates the arrival of a note or syllable, are extracted to measure a movie's audio tempo. Figure 6 gives an example of detected onsets for a music signal. Since a movie consists of various types of sounds from different sources, the proposed audio tempo analysis is first performed in each individual frequency band. The extracted features are then normalized and averaged among all frequency bands to obtain the desired audio tempo feature for each audio frame.

Three motion features are used to measure a movie's visual tempo: camera motion, object motion, and motion variance. Object motion is the average magnitude of motion vectors after compensating the camera motion while motion variance indicates the complexity of an object's activity. Moreover, higher weights are assigned to regions of interest, which correspond to extracted human faces and bodies, for object motion calculation. All three features are normalized and averaged to form the visual tempo feature for each video frame.

LONG-TERM AUDIOVISUAL TEMPO ANALYSIS

As discussed previously, a typical movie story is structured into three parts: exposition, conflict, and resolution. Under this framework, a movie can be decomposed into three large story elements corresponding to each of these three parts, or multiple yet smaller story units at a finer granularity since each element may have its own local climax as well. A large story unit usually covers a relatively complete thematic topic and spans 10–15 min in length. Large story units could be extracted via long-term audiovisual tempo analysis.

The long-term visual tempo is captured by analyzing the tempo histogram of a group of pictures (GOP). Every 15 consecutive frames is organized into one GOP, and each group's visual tempo is set at 80% of its highest histogram peak. The visual tempo curve obtained from a test video sequence is shown in the upper subfigure of Figure 7(a). This curve is then carefully smoothed via

a three-minute long window and plotted in its lower subfigure. Finally, a median filter is applied to remove local instantaneous fluctuations, and all local minima of the curve are identified as large story boundaries. For this example, all detected story boundaries have been marked by vertical lines that are superimposed on the figure.

To analyze the long-term audio tempo, four audio features are first extracted from every 3-min window: energy, energy dynamic, onsets, and silence ratio. They are then fused into one single tempo feature. For the example above, the audio tempo feature curve is shown in Figure 7(b). We see that most of the story boundaries prederived from visual cues match with its local minima. This confirms the claim that a movie producer employs both audio and visual elements to achieve the desired story structure.

SHORT-TERM AUDIOVISUAL TEMPO ANALYSIS

The short-term tempo analysis module extracts small story units that constitute large video story elements. To achieve this task, short-term audio tempo analysis is first conducted to extract audio tempo features and generates a smooth tempo curve for the entire video. The story boundaries are then identified as the local minima of the tempo curve. One example is shown in Figure 8, where the original audio tempo curve for a test video is given in Figure 8(a), its smoothed version in Figure 8(b), and all located valleys, from which the story boundaries are identified and subsequently marked by vertical lines, are illustrated in Figure 8(c).

The visual tempo analysis is performed in the next step. Due to the editing effect, the visual tempo may fluctuate from one shot to another even when both shots are within the same story unit. However, the variance of the tempo remains relatively stable. To characterize the short-term visual tempo, we define the motion variance as

$$\sigma_M^2(i) = \sum_{n=1}^{N} W(n-i)[M(n) - \mu(n)]^2, \qquad (1)$$

where $W(n)$ is a window of length N, $M(n)$ is the basic visual tempo for each shot, n is shot index and

$$\mu(n) = \frac{1}{N}\sum_{n=1}^{N} W(n-i) * M(n) \qquad (2)$$

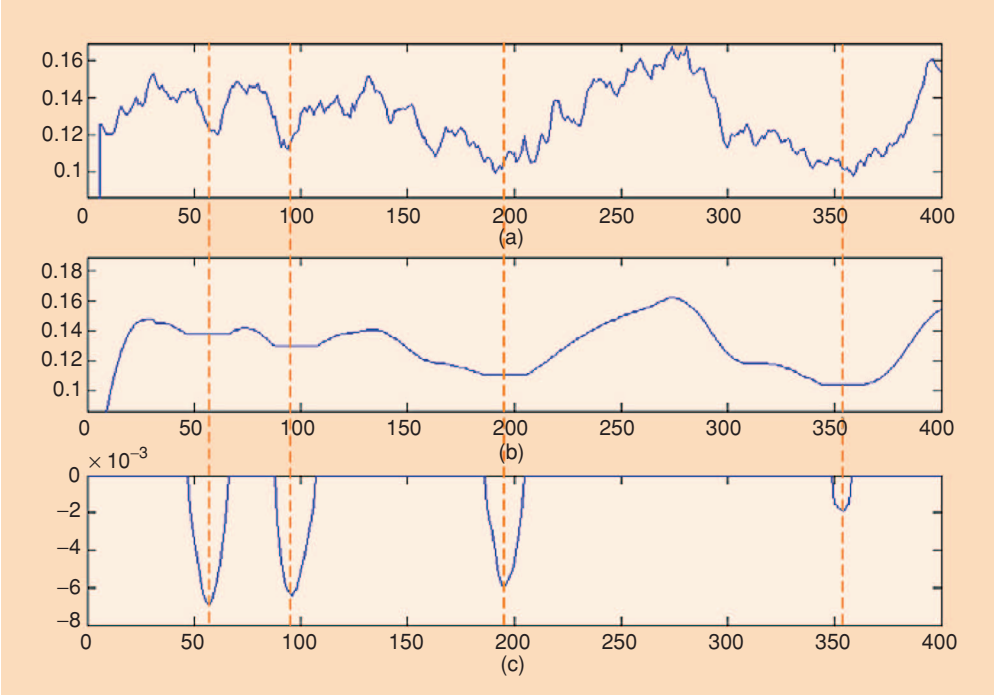
is the mean of the tempo. The story boundaries are identified from the smoothed feature curve by locating its local minima.

This process is demonstrated in Figure 9. The visual tempo curve for the same video clip as the one used in Figure 8 is shown in Figure 9(a). The smoothed curve after the use of morphological filtering, which is used to detect valleys and implemented as a maximize operation followed by a minimize operation, is given in Figure 9(b). The gradient curve whose edge points divide stories into smaller units is plotted in Figure 9(c). Again, vertical lines are used to mark potential story boundaries. Finally, the two boundary sets, obtained from the audio and visual tempo analysis, respectively, are integrated to obtain the final story list.

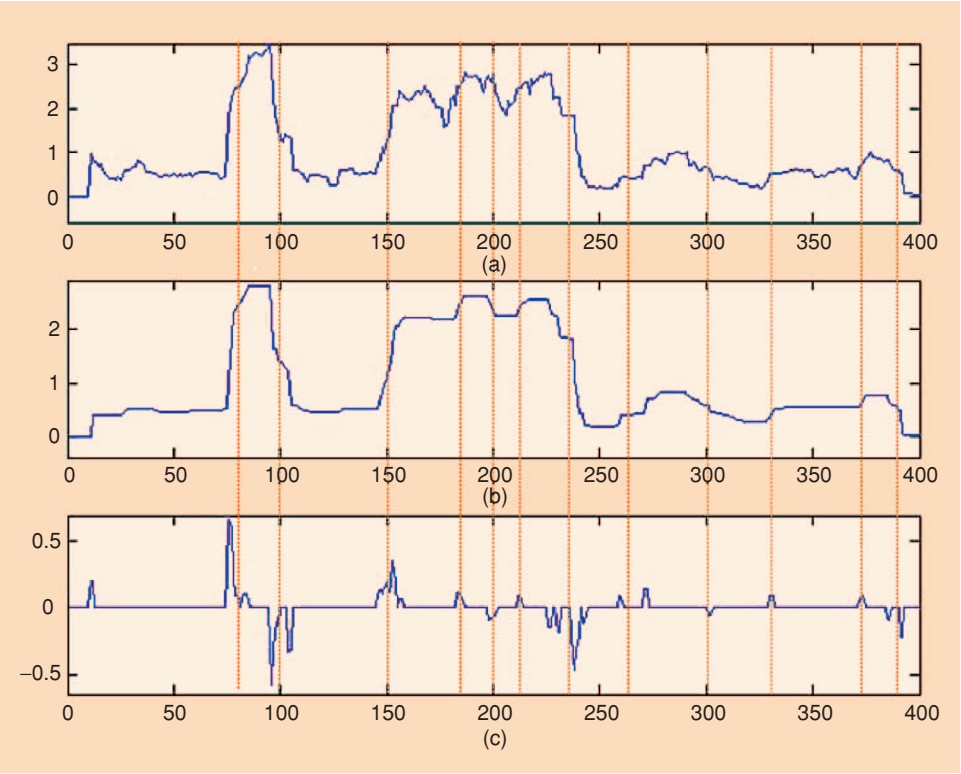
SKIM AUTHORING

In this module, the user-desired skim is generated according to the extracted story structure. Following the criteria defined above, a skim should preserve the continuous flow of the original video in a succinct manner. Moreover, it should appear natural while complying with the strict time constraint. The proposed skim authoring scheme contains the following four steps.

1) Construct a STG [40] for every small story unit of each large story element and extract its independent substories. Specifically, an STG graph describes the story flow within a narrowly defined scene, where each node represents one representative shot. A scene may further consist of



[FIG8] **Short-term audio tempo analysis for a test video.**



[FIG9] **Short-term visual tempo analysis for a test video.**

multiple substories where each substory focuses on a relatively independent topic and involves a set of specific casts. Substory boundaries could be detected from STG by identifying the transitions that switch from a set of repetitive nodes to another.

2) Distribute the user-desired skim duration among all large, small, and substory units in proportion with their respective length.

3) Choose important shots from substory units to abstract the story. These shots must meet the following three criteria: a) temporal consecutiveness, b) covering as many nodes in STG as possible, and c) compliance with the respective time budget. Moreover, to ensure that the skim covers as many contents as possible, the display time of each shot is limited to 1.5 s. When a shot has a longer duration, a frame subsampling scheme will be adopted.

4) Choose a few progressive shots to bridge individual substories into a semantically coherent scene and construct the final skim by gluing them together with previously selected shots.

[TABLE2] PERFORMANCE OF THE PROPOSED SMALL STORY EXTRACTION METHOD.

MOVIE	HIT	MISS	FALSE ALARM	PRECISION	RECALL
MOVIE I	59	11	37	61.5%	84.3%
MOVIE II	31	20	10	75.6%	60.8%

EXPERIMENTAL RESULTS

To evaluate the performance of the proposed movie skimming system, a preliminary experiment was conducted on two movies, each of which is of two-hour duration. The first movie is an action movie called *Die Another Day,* and the second one is a romance called *The Legends of the Fall*. The shot detection algorithm used in this experiment is described in [2]. It applies an adaptive color histogram differencing scheme to detect abrupt as well as gradual content changes.

Figure 7 shows the results of the long-term visual and audio tempo analysis for the first movie. This movie opens up the plot with a secret mission and ends up with a fierce fight scene. This flow coincides with the analysis curve where the first and last segment corresponds to the beginning and the end. Moreover, every segment in the second curve of Figure 7(a) corresponds to a large story unit. Since the large story unit is vaguely defined, it is not easy to evaluate the detection performance. Generally speaking, the detected results are consistent with our experience in watching the movie.

To evaluate the performance of the proposed small story extraction approach, we adopt the precision and recall rates as defined below:

$$\text{recall} = \frac{\text{hits}}{\text{hits} + \text{misses}} \times 100\%, \tag{3}$$

$$\text{precision} = \frac{\text{hits}}{\text{hits} + \text{false alarms}} \times 100\%. \tag{4}$$

They are measured against the manually collected ground truth and tabulated in Table 2. It is worthwhile to mention that the current system is not optimized, and the performance is expected to improve if the implementation details are fine-tuned. The main purpose of showing these preliminary results is to demonstrate that the proposed system offers a promising direction and worths further exploration.

The substory units extracted from the beginning part of *The Legends of the Fall*, which corresponds to the first large story element, is shown in Figure 10(a). It contains 55 shots in total, and each shot is represented by a key frame in this figure. Totally, three substories are detected and shown in the figure. The generated skim of this large story unit is shown in Figure 10(b). All three

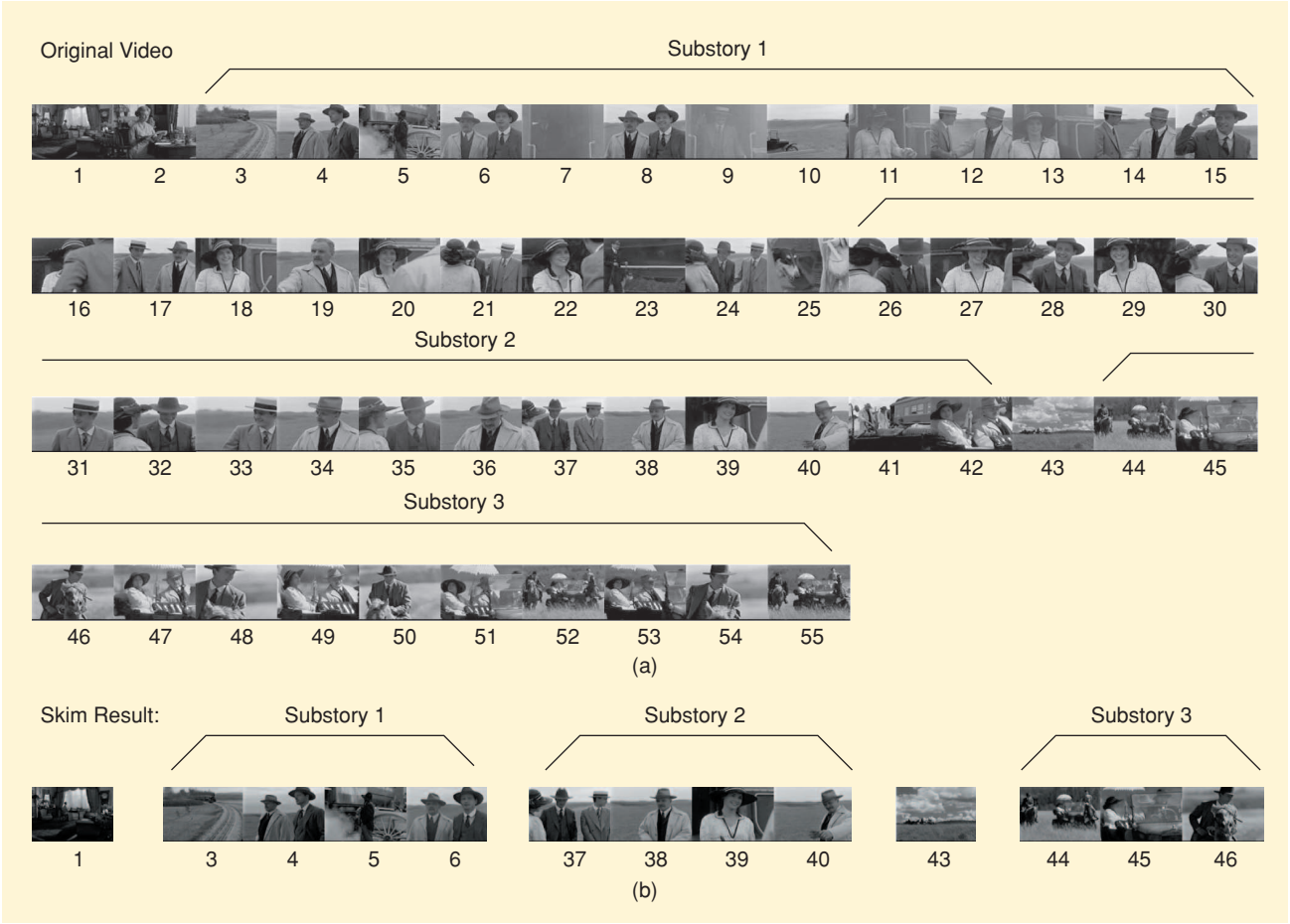
substories have been included, and each of them is represented by a set of consecutive shots. This provides viewers a continuous story flow. Moreover, two progressive shots are selected in the skim so as to make the story synopsis complete. The skimming ratio in this example is 12.5:1.

FUTURE WORK

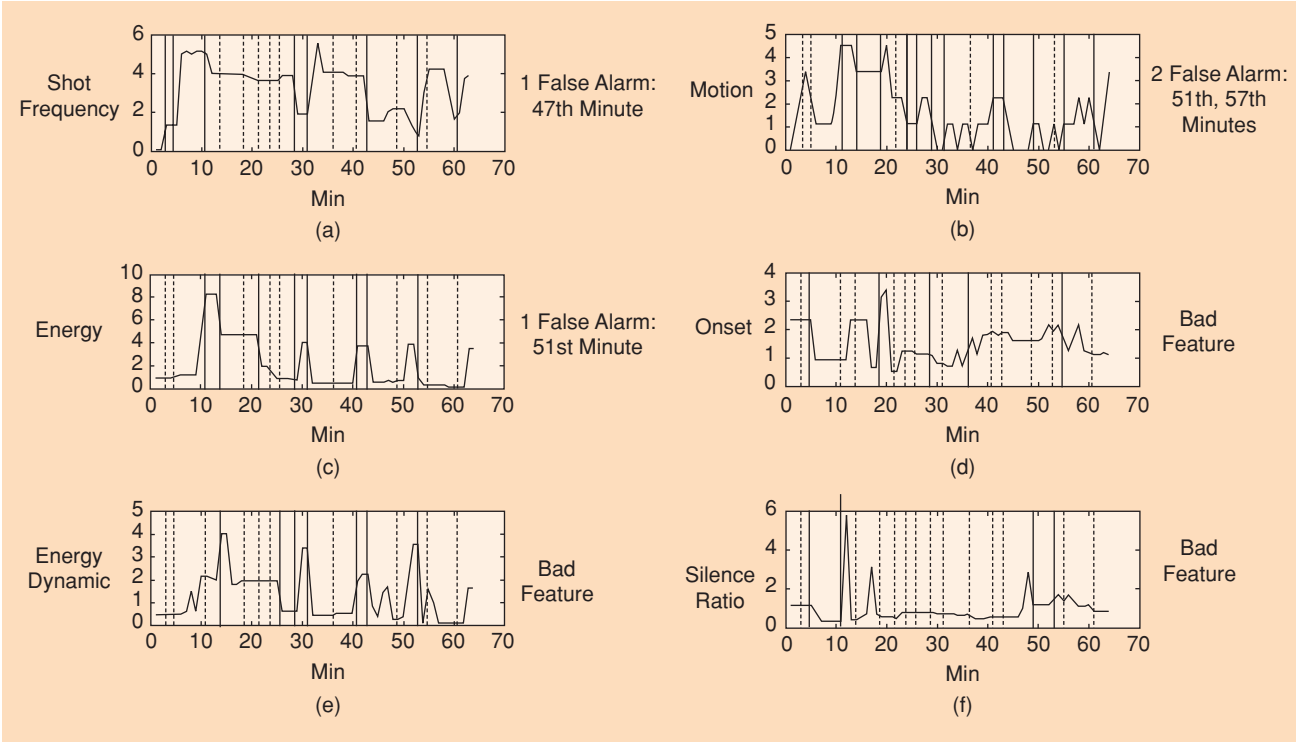
Currently, we are working towards finding optimal features for multilevel movie tempo analysis. This is critical since the effectiveness of selected features often varies with movie genres. For instance, in a violence movie with intensive motion, the motion feature may not serve as an useful choice for distinguishing different story units. To take this into account, some existing work proposes to assign different weights to different features. However, the selection of proper feature weights is still an open problem. It is desired that the weights can be automatically determined based on the movie genre, which is our current research focus. This idea is intuitive since different film categories possess different art forms, and certain features may work the best for a particular genre type. Along this direction, we have obtained some interesting results as shown in Figure 11, where the temporal variations of six different features (shot frequency, motion, energy, onset, energy dynamic, and silence ratio) for a musical movie called *Moulin Rouge* are given. Each subfigure gives an individual feature analysis curve.

VIDEO ABSTRACTION IS A TECHNIQUE THAT ABSTRACTS VIDEO CONTENT AND REPRESENTS IT IN A COMPACT MANNER.

For this example, the mean-shift algorithm [41] was applied to each feature curve to segment the movie into multiple story elements, whose boundaries are marked by vertical lines. Specifically, a solid line indicates that it coincides with the ground truth while a dashed line means that it is a miss. False alarms are not marked. We see from this figure that some features such as the shot change frequency and motion have demonstrated much better performance than onset and silence ratio features. Since *Moulin Rouge* is a musical, it contains many MTV-style scenes filled with loud background music and singing. Thus, the silence ratio and the onset frequency are not suitable in separating different story elements. The best feature in this case is the shot change frequency, which helps separate most of MTV-style scenes from other scenes such as the dialog. Using the proper feature, we can achieve precision and recall rates of 83% and 88%, respectively.



[FIG10] (a) The three substory units and (b) the generated story skim for "The Legends of the Fall." 1994 TriStar Pictures Inc. All Rights Reserved. Courtesy of Sony Pictures Entertainment.



[FIG11] **The long-term feature analysis of *Moulin Rouge*.**

Another movie under our test is a romantic comedy. It does not have intensive action and/or frequent shot cuts. Instead, the content is relaxing and joyful with occasional background music. As a result, the energy dynamic and the silence ratio provide desired features for content analysis. Since the movie genre information might not always be available, automatic movie classification is needed as a preprocessing task. Recently, there has been some ongoing research along this direction [42]. Meanwhile, the affective content analysis approach proposed by Hanjalic [43] in this special issue can also be applied for this purpose.

CONCLUSION

With the proliferation of digital video, video summarization and skimming has become an indispensable tool of any practical video content management system. This article provided a tutorial on the existing abstraction work for generic videos and presented state-of-the-art techniques for feature film skimming. Moreover, our recent work on movie skimming using audiovisual tempo analysis and specific cinematic rules was described. It is our belief that, with the maturity of movie genre classification, content understanding and video abstraction techniques, an automatic movie content analysis system that facilitates navigation, browsing, and search of desired movie content will be arriving in the near future.

AUTHORS

Ying Li received the B.S. and M.S. degrees in computer science and engineering from Wuhan University, China, in 1993 and 1996, respectively, and the Ph.D. in electrical engineering from the University of Southern California in 2003. Since March 2003, she has been with IBM T.J. Watson Research Center as a research staff member. Her research interests include image processing and analysis, multimodal-based video content analysis, e-learning, and computer vision and pattern recognition. She is the author of one book, three book chapters, and numerous technical papers. She also holds five U.S. patents. She is a Member of the IEEE and SPIE.

Shih-Hung Lee received the B.S. and M.S. degree in electrical engineering from the National Tsing-Hua University, Hsin-Chu, Taiwan, in 1996 and 1998, respectively. He is currently a Ph.D. student in the Department of Electrical Engineering at the University of Southern California. He is a research assistant in the Integrated Media System Center and a member of the Multimedia Processing subgroup in Prof. Kuo's Multimedia Research Lab. His research interests include speech and video processing, motion estimation, and computer vision.

Chia-Hung Yeh received the B.S. and Ph.D. degrees from National Chung Cheng University, Taiwan, in 1997 and 2002, respectively, both in electrical engineering. He received a postdoctoral research fellowship award from Prof. C.-C. Jay Kuo's group at the Department of Electrical Engineering-Systems, University of Southern California from August 2002 to December 2004. In the summer of 2004, he joined MAVs Lab. Inc. and has been the technology vice president since. He also serves as an adjunct assistant professor at National Chung-Hsing University since 2005. His research interests are

bioinformatics, multimedia database management, and optical information processing. He has served as a reviewer for international journals and conferences and was an invited speaker at conferences. He is the co-author of a book chapter and more than 50 technical publications in international conferences and journals. He received the Outstanding Student Award from NCCU in 2002.

C.-C. Jay Kuo received the B.S. degree from the National Taiwan University, Taipei, in 1980 and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1985 and 1987, respectively, all in electrical engineering. He is currently professor of Electrical Engineering, Computer Science an Mathematics at USC. His research interests are in the areas of digital signal and image processing, audio and video coding, multimedia communication technologies and delivery protocols, and embedded system design. He has guided about 60 students to their Ph.D. degrees and supervised 15 postdoctoral research fellows. He is coauthor of seven books and more than 700 technical publications in international conferences and journals. He is a Fellow of the IEEE and SPIE and a member of ACM. He received the National Science Foundation Young Investigator Award (NYI) and Presidential Faculty Fellow (PFF) Award in 1992 and 1993, respectively.

REFERENCES

[1] J. Monaco, *How to Read a Film: The Art, Technology, Language, History and Theory of Film and Media*. New York: Oxford Univ. Press, 1982.

[2] Y. Li and C.-C.J. Kuo, *Video Content Analysis Using Multimodal in Formation: For Movie Content Extraction, Indexing and Representation*. Norwell, MA: Kluwer, 2003.

[3] M. Mills, "A magnifier tool for video data," in *Proc. ACM Human Computer Interface*, May 1992, pp. 93–98.

[4] Y. Taniguchi, "An intuitive and efficient access interface to real-time incoming video based on automatic indexing," in *Proc. ACM Multimedia'95*, Nov. 1995, pp. 25–33.

[5] H.J. Zhang, J. Wu, D. Zhong, and S.W. Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recognit.*, vol. 30, no. 4 pp. 643–658, Apr. 1997.

[6] Y.T. Zhuang, Y. Rui, T.S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. ICIP'98*, Oct. 1998, pp. 866–870.

[7] S.X. Ju, M.J. Black, S. Minneman, and D. Kimber, "Summarization of videotaped presentations: Automatic analysis of motion and gestures," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 686–696, Sept. 1998.

[8] C. Toklu and S.P. Liou, "Automatic keyframe selection for content-based video indexing and access," *Proc. SPIE*, vol. 3972, pp. 554–563, Jan. 2000.

[9] M. Iran and P. Anandan, "Video indexing based on mosaic representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 86, no. 5, pp. 905–921, May 1998.

[10] N. Vasconcelos and A. Lippman, "A spatiotemporal motion model for video summarization," in *Proc. IEEE Computer Soc. Conf. Computer Vision Pattern Recognition*, June 1998, pp. 361–366.

[11] Y. Taniguchi, A. Akutsu, and Y. Tonomura, "Panorama Excerpts: Extracting and packing panoramas for video browsing," in *Proc. ACM Multimedia'97*, Nov. 1997, pp. 427–436.

[12] A.D. Doulamis, N.D. Doulamis, and S.D. Kollias, "Non-sequential video content representation using temporal variation of feature vectors," *IEEE Trans. Consumer Electron.*, vol. 46, no. 3, pp. 758–768, Aug. 2000.

[13] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky, "Video manga: Generating semantically meaningful video summaries," in *Proc. ACM Multimedia'99*, Oct. 1999, pp. 383–392.

[14] A. Girgensohn and J. Boreczky, "Time-constrained keyframe selection technique," in *Proc. ICMCS'99*, June 1999, pp. 756–761.

[15] M.M. Yeung and B.L. Yeo, "Video visualization for compact presentation and fast browsing of pictorial content," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 5, pp. 771–785, Oct. 1997.

[16] X.D. Sun and M.S. Kankanhalli, "Video summarization using R-sequences," *Real-Time Imaging*, vol. 6, no. 6, pp. 449–459, Dec. 2000.

[17] K. Ratakonda, M.I. Sezan, and R. Crinon, "Hierarchical video summarization," *Proc. SPIE*, vol. 3653, pp. 1531–1541, Jan. 1999.

[18] A.D. Doulamis, N.D. Doulamis, and S.D. Kollias, "A fuzzy video content representation for video summarization and content-based retrieval," *Signal Process.*, vol. 80, no. 6, pp. 1049–1067, June 2000.

[19] Q. Huang, Z. Lou, A. Rosenberg, D. Gibbon, and B. Shahraray, "Automated generation of news content hierarchy by integrating audio, video, and text information," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Mar. 1999, vol. 6, pp. 3025–3028.

[20] Z. Xiong, R. Radhakrishnan, and A. Divakaran, "Generation of sports highlights using motion activity in combination with a common audio feature extraction framework," in *Proc. IEEE Int. Conf. Image Processing*, Sept. 2003, vol. 1, pp. I-5–I-8.

[21] N. Omoigui, L. He, A. Gupta, J. Grudin, and E. Sanocki, "Time-compression: System concerns, usage and benefits," in *Proc. ACM Conf. Computer Human Interaction*, May 1999, pp. 136–143.

[22] A. Amir, D. Ponceleon, B. Blanchard, D. Petkovic, S. Srinivasan, and G. Cohen, "Using audio time scale modification for video browsing," in *Proc. 33rd Hawaii Int. Conf. System Sciences*, Jan. 2000, vol. 3, pp. 3046–3055.

[23] M. Smith and T. Kanade, "Video skimming and characterization through the combination of image and language understanding," in *Proc. IEEE Int. Workshop Content-Based Access Image Video Database*, Jan. 1998, pp. 61–70.

[24] J. Nam and A.H. Tewfik, "Video abstract of video," in *Proc. IEEE 3rd Workshop Multimedia Signal Processing*, Sept. 1999, pp. 117–122.

[25] A. Hanjalic and H.J. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 8, pp. 1280–1289, Dec. 1999.

[26] X. Zhu, J. Fan, A.K. Elmagarmid, and W.G. Aref, "Hierarchical video summarization for medical data," *Proc. SPIE*, vol. 4674, pp. 395–406, Jan. 2002.

[27] D.D. Russell, "A design pattern-based video summarization technique: moving from low-level signals to high-level structure," in *Proc. 33rd Hawaii Int. Conf. System Sciences*, Jan. 2000, vol. 3, pp. 1137–1141.

[28] L. He, E. Sanocki, A. Gupta, and J. Grudin, "Auto-summarization of audio-video presentations," in *Proc. ACM Multimedia*, Oct. 1999, pp. 489–498.

[29] S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg, "Abstracting digital movies automatically," *J. Visual Commun. Image Represent.*, vol. 7, no. 4, pp. 345–353, Dec. 1996.

[30] H. Luo and J. Fan, "Concept-oriented video skimming and adaptation via semantic classification," in *Proc. 6th ACM SIGMM Int. Workshop Multimedia Information Retrieval*, Oct. 2004, pp. 213–220.

[31] Y. Ma, L. Lu, H. Zhang, and M. Li, "A user attention model for video summarization," in *Proc. ACM Multimedia*, Dec. 2002, pp. 533–542.

[32] H. Sundaram and S.-F. Chang, "Condensing computable scenes using visual complexity and film syntax analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, Aug. 2001, pp. 389–392.

[33] U. Gargi, R. Kasturi, and S.H. Strayer, "Performance characterization of video-shot-change detection methods," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 1, pp. 1–13, Feb. 2000.

[34] M. Yeung and B. Yeo, "Time-constrained clustering for segmentation of video into story units," in *Proc. Int. Conf. Pattern Recognition*, Aug. 1996, vol. 3, pp. 375–380.

[35] K. Reisz and G. Millar, *The Technique of Film Editing*. New York: Hastings House, 1968.

[36] H. Sundaram and S.F. Chang, "Determining computable scenes in films and their structures using audio-visual memory models," in *Proc. ACM Multimedia'00*, Marina Del Rey, Oct. 2000, pp. 95–104.

[37] B. Block, *The Visual Story: Seeing the Structure of Film, TV, and New Media*. Boston, MA: Focal Press, 2001.

[38] B. Adams, C. Dorai, and S. Venkatesh, "Automated film rhythm extraction for scene analysis," in *Proc. ICME'01*, Aug. 2001, pp. 849–852.

[39] B. Adams, C. Dorai, and S. Venkatesh, "Towards automatic extraction of expressive elements from motion pictures: Tempo," in *Proc. ICME'00*, July 2000, pp. 641–644.

[40] M. Yeung, B. Yeo, and B. Liu, "Extracting story units from long programs for video browsing and navigation," in *Proc. IEEE Multimedia Computing Systems*, 1996, pp. 296–305.

[41] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[42] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 52–64, Jan. 2005.

[43] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized TV," *IEEE Signal Processing Mag.*, vol. 23, no. 2, pp. 90–100, 2006. SP