Editing by Viewing: Automatic Home Video Summarization by Viewing Behavior Analysis

Wei-Ting Peng, Wei-Ta Chu, Chia-Han Chang, Chien-Nan Chou, Wei-Jia Huang, Wen-Yan Chang, and Yi-Ping Hung, *Member, IEEE*

***Abstract*—In this paper, we propose the Interest Meter (IM), a system making the computer conscious of user's reactions to measure user's interest and thus use it to conduct video summarization. The IM takes account of users' spontaneous reactions when they view videos. To estimate user's viewing interest, quantitative interest measures are devised based on the perspectives of attention and emotion. For estimating attention states, variations of user's eye movement, blink, and head motion are considered. For estimating emotion states, facial expression is recognized as positive or neural emotion. By combining characteristics of attention and emotion by a fuzzy fusion scheme, we transform users' viewing behaviors into quantitative interest scores, determine interesting parts of videos, and finally concatenate them as video summaries. Experimental results show that the proposed concept "editing by viewing" works well and may provide a promising direction to consider the human factor in video summarization.**

***Index Terms*—Attention detection, editing by viewing, emotion recognition, Interest Meter (IM), video summarization.**

I. Introduction

ALTHOUGH shooting a home video is often enjoyable, editing videos is often tedious and troublesome. To conduct good editing, in addition to choosing appropriate and convenient software, the user's prior knowledge of media aesthetics, editing theory, and computer skills is also essential [3]–[5]. Commercial video editing software such as Adobe Premier,[1] , Sony Vegas,[2] or Apple iMovie[3] is equipped with a variety of editing tools. However, for novice home users who are not fluent in filmmaking and editing, these tools can be more confusing than helpful.

Argyle [1] indicated that users have the following reactions when they are interested in something: laughing, more fixations, fewer blinks, and lively movements of shoulders and head-nods. Eye gaze plays an important role in attention because a listener usually pays attention to the speaker by looking at him. Moreover, an intuitive and obvious clue to show emotion is facial expression. It has been demonstrated that emotion influences people's attitude to take various actions, and there is evidence that it plays an essential role in rational decision making, perception, learning, and other cognitive functions [2]. Motivated by the studies described above, we construct a module called Interest Meter (IM) to conduct psychological analysis and explore how the human attention system facilitates video editing and summarization.

The proposed system (IM) conducts blink detection, saccade detection, head motion detection, and facial expression recognition to measure users' interest. IM monitors users' reactions when they view a home video, such as facial expressions, blinks, eye movements, and head motions, and identify which parts of video clips s/he might be interested in. These clips would then be chosen into the final video summary. We would show that the proposed video summarization system can make an appealing video summary with ease. Note that this system works from a significantly different aspect from conventional content-based attention. We detect human's actions and analyze psychological states rather than analyzing visual/aural variations that are widely adopted in previous works. In a word, the proposed system facilitates users to "do video editing by viewing the video," rather than "do video editing via the complex interfaces containing tens of icons."

Contributions of this work are summarized here.

- We present a novel video summarization method that analyzes user's viewing behaviors rather than the visual content itself. The concept of "editing by viewing" is believed to be one of the first works to conduct video summarization directly from human's psychological states.
- In contrast to content-based methods for detecting saliency parts in videos, we analyze viewing behaviors respectively from attention and emotion perspectives and then fuse them by the concept of fuzzy logic. Few works have been proposed before for estimating user's interest from viewing behaviors.

The remainder of this paper is organized as follows. Section II provides literature survey from three aspects. Section III describes construction details of the attention model and the emo-

Manuscript received November 02, 2010; revised February 25, 2011; accepted March 08, 2011. Date of publication March 22, 2011; date of current version May 18, 2011. This work was supported in part by the National Science Council, Taiwan, under Grant NSC 98-2221-E-002-127-MY3 and Grant NSC 98-2221-E-002-128-MY3 and by the Excellent Research Projects of National Taiwan University, under Grant 99R80300. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zicheng Liu.

W.-T. Peng and Y.-P. Hung are with the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 106, Taiwan (e-mail: d93944004@ntu.edu.tw; hung@csie.ntu.edu.tw).

W.-T. Chu is with the Department of Computer Science and Information Engineering, National Cheng Chung University, Chiayi 621, Taiwan (e-mail: wtchu@cs.ccu.edu.tw).

C.-H. Chang, C.-N. Chou, and W.-J. Huang are with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei 621, Taiwan (e-mail: chiahan_chang@yahoo.com.tw; 92502093@cc.ncu.edu.tw; aga3134@gmail.com).

W.-Y. Chang is with the Instititute of Information Science, Academia Sinica, Taipei 115, Taiwan (e-mail: wychang@iis.sinica.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2011.2131638

[1] [Online]. Available: http://www.adobe.com/products/premiere/

[2] [Online]. Available: http://www.adobe.com/products/premiere/

[3] [Online]. Available: http://www.apple.com/ilife/imovie/

tion model and then states a fusion scheme to estimate user's interest. Details of video summarization, including analysis of the accompanying music, are provided in Section IV. Experimental results on each component and video summarization are provided in Section V, followed by conclusion and future work described in Section VI.

II. Related Work

Shooting video is fun but editing has proven to be frustrating. Hence, users usually put video footage on the shelf without further intention to elaborately edit. Therefore, video summarization has been studied for years from various perspectives. From the information analysis aspects, internal information, external information, or integration of both have been widely used in automatic video summarization. For enriching browsing experience, video summaries are often associated with music and video-centric or music-centric methods have been developed. To ease video editing, methodologies of fully automatic, semi-automatic or manual editing with friendly interfaces have been developed.

A. From the Perspective of Information Analysis

Money and Agius [6] provide an extensive literature survey on video summarization. They classify related literature into three categories: 1) internal summarization techniques; 2) external summarization techniques; and 3) hybrid summarization techniques. By definition, internal summarization techniques analyze internal information from video streams, which was produced during the production stage of the video lifecycle. These techniques extract low-level image, audio, and text features to facilitate summarization and are the most common summarization techniques [7]–[11]. External summarization techniques analyze external information during any stage of the video lifecycle. User-based information, which is information directly from users and contextual information, which is information not sourced directly from users or video streams, are two main types of external information. As for hybrid summarization techniques, both internal and external information are analyzed.

External information is collected when users view and interact with video content and then this information is analyzed to develop video summaries. Money and Agius [12] developed a video summarization technique by analyzing user's physiological response, including electro-dermal response (EDR), respiration amplitude (RA), respiration rate (RR), blood volume pulse (BVP), and heart rate (HR). Joho *et al.* [13] presented an approach on affective video summarization based on viewer's facial expressions. Our previous work [14] analyzed variations of viewer's eye movement and facial expression when he or she viewed a home video and transformed these behaviors into clues for determining the important part of each video shot. Compared with other similar works, we propose a framework to explore the impact of user's viewing behaviors on video editing. In our investigation, when viewers watch videos, they do not always have significant facial expression. For this reason, we also employ eye movement to determine important video parts. In this work, we enhance our previous work [14], [41] by exploring various features for attention and emotion evaluation and exploiting fuzzy theory to integrate different information as the final interest measurement.

B. From the Perspective of Audio-Visual Synthesis

For an enriching browsing experience, video summaries are often associated with music, and musical video (MV) generation has become popular in recent years. Generally speaking, there are two methods to synthesize music with a video: video-centric and music-centric. In a video-centric method, the music is dubbed based on visual features extracted from video. For example, Mulhem *et al.* [7] developed a pivot vector space method that automatically picks the best audio clip from a database to mix with a given video shot. In a music-centric method, various video clips are edited to match with a music piece. This is the approach commonly seen in the music industry to generate MVs. To produce an MV, a song's beat and tempo are first analyzed, and then appropriate video clips are selected to match with music segments to generate a specific rhythm. Foote *et al.* [8] presented methods for automatic creation of music videos. Hua *et al.* [9] proposed another segment-based matching method for home video summarization. Yoon *et al.* [10] used computable characteristics of video and music to promote coherent matching. Wang *et al.* [11] proposed both video-centric and music-centric algorithms to synthesize a musical video.

C. From the Perspective of Computer–Human Interaction

To ease video editing and simultaneously match user's need, sort of human intervention can be added in the cycle of summary generation. From the perspective of computer–human interaction, we can classify them as manual, fully automatic, or semi-automatic. Most commercial editing software facilitates the manual process with powerful or friendly interfaces. Although they provide a wide variety of functions, editing a video using this software can still be difficult even for experts and much more so for a novice. Fully automatic video editing systems such as [8]–[10] and the software PowerDirector [36] can render MVs through their built-in algorithms. Although the fully automatic methods take little time, users are not able to make changes when they are not satisfied with the results.

Wang *et al.* [11] proposed a dynamic-programming-based algorithm for automatic or semi-automatic generation of personalized musical videos. Shipmanet *et al.* [37] proposed the Hyper-Hitchcock program that provides a user interface and various semi-automatic techniques to generate video summaries, but not musical videos. The semi-automatic software MuVee [38] provides a user interface so that users can adjust the results of an automatically generated music video.

III. Interest Meter

Fig. 1 shows the system framework. When a user views the video, we capture his/her upper body (the most important part is face) and detect face and eyes. The IM is then constructed based on the attention model and the emotion model, where attention describes the visual focus of the user and emotion describes the inner state of the user. For the video, we first segment it into shots based on difference of HSV color histograms in two consecutive frames. By fusing information from attention and emotion and with the temporal correspondence of how the user re-

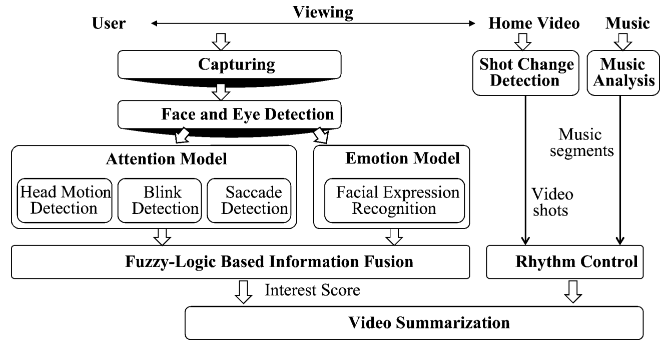

Fig. 1. Framework of video summarization based on the IM.

acts when viewing a specific video shot, we estimate the degree of interest. Based on this measure, a video summary is generated by selecting video clips of higher interest.

To estimate attention states from human's viewing behaviors, we establish head motion detection, blink detection, and saccade detection. To estimate emotion states, facial expression recognition is the main part. In both models, we transform these states into quantitative scores. This information is then fused to determine the interest score. Note that this interest score is not derived from visual content itself, but from the viewer's behavior. The same video segment may draw different levels of interest from different viewers. With this score, interesting video clips are selected to generate a personalized video summary.

A. Face and Eye Detection

Face detection is the first step in the IM system framework, and it is indispensable in eliminating interference from the background while allowing the user to move freely. Current face detection techniques have been developed over a long period and can detect faces efficiently and accurately. In our work, the position of face is estimated by using the boosted cascade face detector proposed by Viola and Jones [15].

One property of the human visual system is that people can identify a human face from a great distance even though the facial details are vague. This means that the symmetric characteristic is sufficient for recognition. A human face is made up of features such as the eyes, nose, mouth, and chin. They are different in shape, size, and structure, but they are always in the right places and have good proportions to each other [16], [17]. The eyes are below the eyebrows, and they are almost symmetrical around the nose. Based on face detection results, we demarcate possible regions for searching eyes. The face region is divided into four equal regions, i.e., left-top, right-top, left-bottom, and right-bottom. The right-eye detector is used to detect the right eye in the left-top region and the left-eye detector is used to detect the left eye in the right-top region. This strategy avoids false eye detection at the positions of the nostrils or mouth. We adopt the boosted cascade eye detector proposed by Castrillón *et al.* [18] to quickly and accurately estimate locations of eyes.

B. Attention Model

1) Head Motion Detection: When viewing an interesting video clip, viewers may frequently move their heads to track objects in videos. Larger head motion subtly means higher attention. Based on face detection results, we calculate displacement of center of face between two consecutive frames and calculate the corresponding score for head motion $S_m(t)$ as

$$S_m(t) = e^{-(m(t))^2/\sigma 1} \tag{1}$$

where $m(t)$ is the displacement (in terms of pixels) from the center of the face at frame t to the center of the face at frame $t-1$ and $\sigma 1$ is a control factor set as 200.

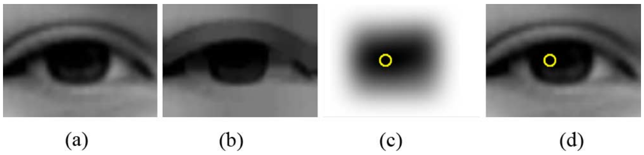

Fig. 2. Detection of eyeball center. (a) Original eye image. (b) Opening operator is applied on (a). (c) Gaussian filter is applied on (b). (d) Result of eyeball center detection.

We have to emphasize that the unit of a *frame* in this work is different from video frames or audio frames in conventional multimedia analysis. The time axis of a video is equally divided into 0.01-s units, that is, the time difference between frame t and $t-1$ is 0.01 s, and we calculate the score for head motion for every 0.01-s frame. The same setting is used for estimating other attention and emotion scores and for music analysis and video summarization in Section IV.

2) Blink Detection and Saccade Detection: Fewer eye blinks would be drawn when viewing an interesting video clip. Similarly, viewers may gaze at important objects and draw fewer saccades. Therefore, this information provides important clues for us to detect attention states from human viewing behavior. For blink and saccade detection, we consider three visual features: the center of the eyeball, two corners of eyes, and the upper eyelids. To find the center of the eyeball, the opening operator is first applied to eliminate the highlight that may be caused by the reflection on the cornea [Fig. 2(b)], and then the iris is estimated by convolving the gray eye image with a Gaussian-shaped filter to find the center of the darker region [Fig. 2(c)]. Vezhnevets *et al.* [19] propose a similar function for the same purpose. We define the function as

$$G(x,y) = Ae^{-(x-x_0)^2+(y-y_0)^2/2(\sigma 2)^2} \tag{2}$$

where the term A is the amplitude, (x_0, y_0) is the center, and $\sigma 2$ controls the width of the Gaussian shape. We rescale the eye image to a fixed size before convolution. The parameter $\sigma 2$ can be chosen according to the expected iris size. After convolution, the pixel with the lowest response is considered to be the approximate iris center [Fig. 2(d)].

The method described above only yields an approximate estimation of the iris center. In order to refine the iris center and estimate the iris radius, we further identify the circular shape of the iris. First, the edge map of the eye image after eliminating the highlight is obtained by using the Canny edge detection method. To find sample points of the iris boundary on the edge map, we virtually draw rays radically from the approximate iris center and then obtain the intersections of the rays and edges. The di-

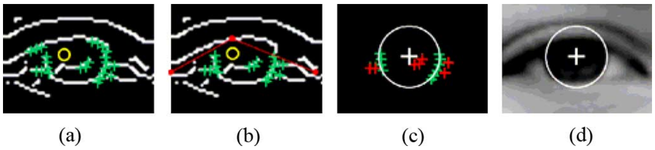

Fig. 3. Iris center detection.

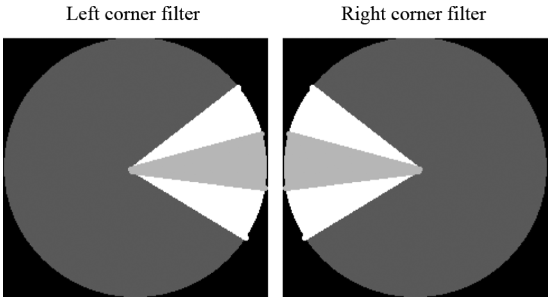

Fig. 4. Illustration of the wedge filter.

ameter of the iris is always smaller than the length between the two corners [20], and thus the length of each ray is limited to half the length between two corners.

An example set of sample points is shown in Fig. 3(a). We restrict the directions of the rays because the iris is likely to be occluded by the eyelids and eyelashes. The range of angles is adjustable to accommodate different users, but it is initially defined to include the ranges -45° to 45° and 135° to 225°. One ray is traced per 5°, resulting in at most 108 candidate sample points. In a real situation, however, there are many outliers due to eye blinks. In order to eliminate these outliers, an upper eyelid point is obtained by tracing a vertical ray from the starting point and finding an intersection, and then those sample points above the two links between the upper eyelid point and the two eye corners are excluded [Fig. 3(b)].

The candidate sample points may still contain outliers. A circle is fit to the candidate sample points using the Random Sample Consensus (RANSAC) paradigm [21]. Unlike a least-squares fitting approach, this paradigm reduces the influence on outliers. We introduce two restrictions on the RANSAC fitting process to increase robustness. First, only candidate circles that include the starting point within the covered areas are considered. Second, based on the structure of the eye, the ratio of the iris diameter to the length between the two eye corners is about 1:3 and only candidate circles with reasonable ratios (about 1:3) are considered. The inliers and outliers are shown as green and red crosses, respectively, in Fig. 3(c), and the final circle fit is shown in Fig. 3(d).

To detect corners of the eyes, we refer to the method exploiting a nonlinear wedge filter in [22]. From [22], "The eyes can be thought of as islands in a sea of flesh tone. They have a distinct spatial pattern at each corner." Therefore, finding the corners of the eyes can be accomplished by finding a wedge of sclera color surrounded by flesh tone. In [22], a "wedge filter" was designed to detect the left eye corner and the right eye corner, respectively, as illustrated in Fig. 4. A brighter region in Fig. 4 means a higher value to be convolved with pixels' intensity values. The largest wedge looks for the flesh tone, and the smallest wedge looks for the sclera tone. At every location on the face, the left corner filter and the right corner filter are respectively convolved with intensity values of the pixels in the masked region. A right or left corner is detected if the average value of the pixels in each wedge satisfies certain criteria. Detailed settings of the wedge filters and the detection criteria please refer to [22].

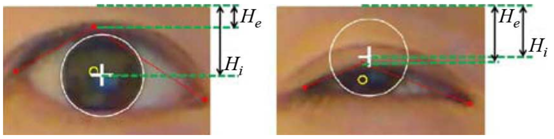

Fig. 5. Two different eye states.

Based on positions of the eyeball and two eye corners, we estimate the degree of eye movement by comparing the relative distances between the eyeball center and the eye corners over time. If the velocity of the eyeball movement between the current frame and the previous frame is larger than a threshold, a saccade is detected.

A blink is defined as a user closing his eyes. Therefore, an eye blink is detected when the iris center is occluded by the upper eyelid. Whether or not the iris center is occluded determines the status of the eye at each frame. Let $\mathrm{Blink}(t)$ represent the status of the eye at frame t as

$$\mathrm{Blink}(t) = \begin{cases} \text{open}, & \text{if } H_i \geq H_e \\ \text{closed}, & \text{otherwise} \end{cases} \tag{3}$$

where H_i and H_e are the distances from the upper boundary of the eye region to the iris center and to the upper eyelid point, respectively. Fig. 5 shows two eyes states. As the eye changes from the open to closed states, we determine that a blink occurs.

3) Blink Score and Saccade Score Calculation: To transform results of blink detection into quantitative score, we first define a blink detection function $b(t)$. If a blink is detected around the frame t, $b(t) = 1$; otherwise $b(t) = 0$. The blink score $S_b(t)$ can be expressed as

$$S_b(t) = \begin{cases} 1, & \text{if } \sum_{t \in W} b(t) \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where W is a 1-s sliding window centered at the frame t, that is, the window covers a 0.5-s segment ahead of frame t and a 0.5-s segment behind frame t. If there is fewer than one blink in this 1-s window, it indicates higher attention in the corresponding duration.

Goldstein *et al.* [23] classify eye movements into three categories: fixations, smooth pursuits, and saccades. They reported that a movement velocity larger than 200°/s corresponds to a saccade. In this work, we take saccades into account because they indicate shifts of viewing attention. More saccades occur in viewing a shot that is less interesting for viewers. Similarly, we analyze saccades around a 1-s sliding window W. If a saccade is detected around the frame t, the saccade detection function

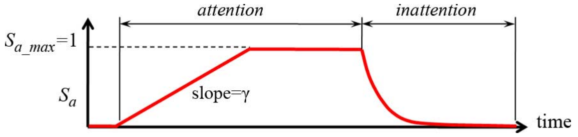

Fig. 6. Evolution of attention score.

$s(t) = 1$; otherwise, $s(t) = 0$. The saccade score $S_s(t)$ can be expressed as

$$S_s(t) = \begin{cases} 1, & \text{if } \sum_{t \in W} s(t) = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

4) Attention Score Calculation: The fuzzy system, proposed by Takagi and Sugeno [24], is a paradigm applied in developing both linear and nonlinear systems for embedded control. The advantage of fuzzy logic is that we can describe systems using simple English-like rules. It does not require system modeling or complex math equations governing the relationship between inputs and outputs. Based on this theory, we use fuzzy logic to integrate three scores described above and calculate the attention score $S_a(t)$ at the frame t. The *fuzzy if–then rule* can be expressed as

$$\begin{cases} \text{IF } (S_m(t) \text{ is high) AND } (S_b(t) \text{ is 1) THEN } (S_a(t) \text{ is } FS_1(t)) \\ \text{IF } (S_m(t) \text{ is high) AND } (S_s(t) \text{ is 1) THEN } (S_a(t) \text{ is } FS_1(t)) \\ \text{IF } (S_b(t) \text{ is high) AND } (S_s(t) \text{ is 1) THEN } (S_a(t) \text{ is } FS_1(t)) \\ \text{otherwise } S_a(t)\ is\ FS_2(t) \end{cases} \tag{6}$$

where S_m is the head motion score, S_b is the blink score, and S_s is the saccade score. The score $S_m(t)$ is determined to be high if $S_m(t) > 0.6$. The notation $S_a(t) = FS_1(t)$ means that the user is attentive to some object and $S_a(t) = FS_2(t)$ means score evolution for inattentive parts. From (6), if more than two of the three scores (corresponding to head motion, blink, and saccade) are high, the part being viewed is viewed as an attentive part. Otherwise, it is viewed as an inattentive part.

In general, attention accumulates gradually over time but may go down immediately. In the attentive situation, the attention score increases stably with a slope of γ. On the contrary, the attention score would decrease by $\alpha (\alpha < 1)$ times the original attention score when the user is inattentive. Fig. 6 illustrates the evolution of attention score. Based on this observation, the value of attention in the present frame should change depending on the value of the previous adjacent frame. Therefore, we can define $FS_1(t)$ and $FS_2(t)$ as follows:

$$S_a(t) = \begin{cases} FS_1(t) = S_a(t-1) + \gamma, \text{(attentive)} \\ FS_2(t) = \alpha \times S_a(t-1), \text{(inattentive)} \end{cases}. \tag{7}$$

C. Emotion Model

When viewing videos, users spontaneously express their feelings through facial expressions. For example, when something funny appears in a video, most users smile or laugh at what they see. Thus, we adopt facial expression analysis to obtain information from such user's reaction. For facial expression recognition, instead of analyzing six-class expressions [25], [26], we classify human expressions into two categories: positive expressions and neutral expressions. A positive expression is defined as a positive human reaction that implies the user is interested in this object, including smiling and laughing. Expressions other than positive expressions are classified as neutral expressions. Although people may move their heads when viewing videos, we reasonably assume that they watch videos by both eyes and focus only on emotion recognition for frontal face.

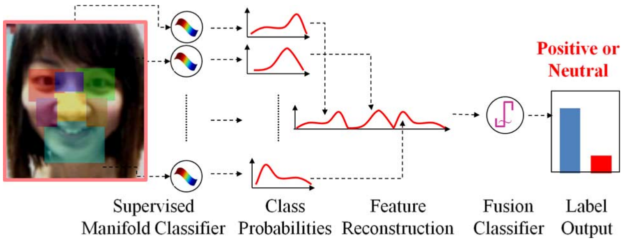

Fig. 7. Classifier learning for facial expression recognition.

We adopt a manifold learning method to integrate multi-component information for facial expression recognition. Our work employs a total of nine facial components to determine expression. Given a face image I, a representative feature is constructed by learning the mapping $M : R^d \times c \longrightarrow R^k$ based on facial components. Essentially, the mapping M encodes the probability of each expression in facial components and can be defined as

$$M(I) = [m_1(I_1), m_2(I_2), \ldots, m_c(I_c)] \tag{8}$$

where c is the number of components, $m_i(\cdot)$ is an embedding function of the component i and I_i is a d-dimensional subimage of the ith component. By learning geometry of the training data, an embedding function $m_i(I_i)$ can be obtained by projecting I_i onto the learned manifold. In our framework, a probabilistic representation of $m_i(I_i)$ can be written as

$$m_i^p(I_i) = \frac{D^p}{D^p + D^n} \quad m_i^n(I_i) = \frac{D^n}{D^p + D^n} \tag{9}$$

where D^p is the shortest distance from I_i to the positive training data and D^n is the shortest distance from I_i to the neutral training data. The value $m_i^p(I_i)$ denotes the probability of positive emotion based on the facial component I_i, and the value $m_i^n(I_i)$ denotes the probability of negative emotion based on the facial component I_i. Based on this formulation, the multicomponent information is then encoded in a k-dimensional feature vector $M(I)$, where k is $2 \times 9 = 18$ in this case. To characterize significance of components from the embedded features, a fusion classifier $F : R^k \longrightarrow \{\text{Positive}, \text{Neutral}\}$ is constructed based on a probabilistic SVM classifier. This method allows our system to recognize users' emotions.

Fig. 7 illustrates the process of learning a classifier based on multiple components. Relative positions and sizes of the nine facial components are defined in advance, according to general layout of face parts including eyes, nose, cheek, and chin. Before emotion recognition, results of eye detection (Section III-A) are used to determine the positions of the top three facial components (left eye, center of two eyes, and right eye). Positions of the remaining facial components are then determined by fitting

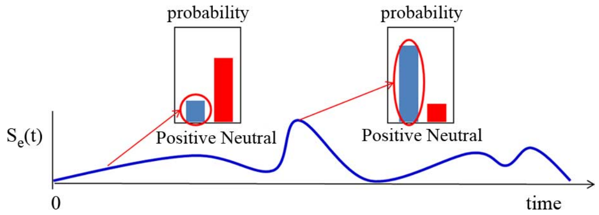

Fig. 8. Calculation of emotion score.

TABLE I
FUZZY IF–THEN RULE OF INTEREST SCORE

$S_r(t)$		$S_a(t)$		
		High $0.7 < S_a(t)$	Medium $0.4 < S_a(t) < 0.75$	Low $S_a(t) < 0.5$
$S_e(t)$	High $0.6 < S_e(t)$	High $S_r(t) = 0.75$	High $S_r(t) = 0.75$	High $S_r(t) = 0.75$
	Medium $0.45 < S_a(t) < 0.65$	High $S_r(t) = 0.75$	Medium $0.45 < S_r(t) < 0.65$	Low $S_r(t) = 0.5$
	Low $S_a(t) < 0.5$	High $S_r(t) = 0.75$	Low $S_r(t) = 0.5$	Low $S_r(t) = 0.5$

the predefined layout based on these three determined components. Intensity values of pixels in these facial components are then extracted for classifier learning.

Based on facial expression recognition results, we again transform them into quantitative scores. We use the probability of a positive emotion as the emotion score $S_e(t)$, which ranges from 0 to 1. A larger emotion score $S_e(t)$ represents that the visual content at frame t is more important for viewers than neutral ones. Fig. 8 illustrates a sample evolution of the emotion score for a video. Note that only the probability of positive emotion over time is considered.

D. *Fuzzy-Logic-Based Information Fusion*

We use fuzzy logic to integrate attention score and emotion score into the interest score $S_r(t)$. The reasons for using fuzzy logic are twofold. First, we can hardly find a generic combination method (either linear or nonlinear) to integrate multiple cues. Different people have different scales of reactions to interesting things, and the parts of interest may be distinct to different people. Training a set of parameters to model various viewing behaviors is not feasible. Second, performance of detecting blink, saccade and emotion is not perfect yet. Various degrees of performance perturbation, especially saccade detection, make defining a hard decision/combination difficult. In this work, we divide these scores into three rough regions (high, medium, and low) and determine final interest scores by fuzzy fusion. Implementation of this task is based on the AForge.NET library.

The *fuzzy if–then rule* can be expressed in Table I. Because different users have different scales of response on attention and emotion, we design these ranges empirically. The idea of fuzzy fusion is based on the relative strength between attention and emotion. For the scores at frame t, when either the emotion score or the attention score falls into the "high" region, the corresponding interest score is set as high as well. If neither score is high and one of them is low, the corresponding interest score is set as low. For the remaining case, the interest score lies in the range from 0.45 to 0.65, in which the exact value is determined by the function implemented in AForgeNet.[4]

IV. VIDEO SUMMARIZATION

Based on the interest scores described above, we are able to select important video segments and concatenate them as a summary. To enrich the browsing experience of video summary, in this work we target on generating a video summary with accompanying music, just like a musical video, in which video shots change as significant music beats strike. Here, we first describe how to divide a user-selected music piece into smaller segments and then select a video clip of appropriate length from each video shot to accompany with a music segment. The final summary is presented in an audiovisual manner and provides richer presentation than a video-only summary.

A. *Music Analysis*

To coordinate visual and aural presentations, we would like to make the rate of shot changes in the generated summary conforms to tempo of the music. Thus, we estimate tempo of the user-selected background music in the following. Onsets are first detected based on energy dynamics, as they generally occur when there are significant energy changes [43]. We apply the Fourier transform with a Hamming window to calculate the frequency bins of each frame. The spectral flux [44] is one prominent feature that is widely used to measure changes of magnitudes between frequency bins. A peak of magnitude change is selected as an onset if it fulfills the peak-picking algorithm in [45]. We define a detection function $peak(n)$, which outputs one if the $n\mathrm{th}$ frame represents a peak and outputs zero otherwise. Note that we also equally divide music into 0.1-s frames, similar to the setting in Sections III-B and III-C.

The tempo value of the nth frame is defined as a sum of peaks over a local window with size w

$$tempo(n) = \sum_{k=(n-w)/2}^{(n+w)/2} peak(k). \tag{10}$$

The red dashed curve in Fig. 9 shows evolution of tempo values for selected music.

The idea of video summarization is to select a shorter but interesting segment from each video shot and concatenate them as the final summary. Suppose that there are K video shots in the original home video. We first divide the music into K segments and find appropriate video segments from K video shots to fit in with these K music segments. The length of the ith music segment is

$$L_{M_i} = L_M \times \frac{L_{V_i}}{L_V} \tag{11}$$

where L_M and L_V denote the total length of the music and video, respectively. The value L_V denotes the length of the ith video shot. Note that $\sum_{i=1}^{K} L_{M_i} = L_M$. The starting time of the ith music segment is therefore

$$\hat{t}_{M_i} = L_M \times \frac{\sum_{j=1}^{i-1} L_{V_i}}{L_V} \tag{12}$$

[4] [Online]. Available: http://www.aforgenet.com/framework/

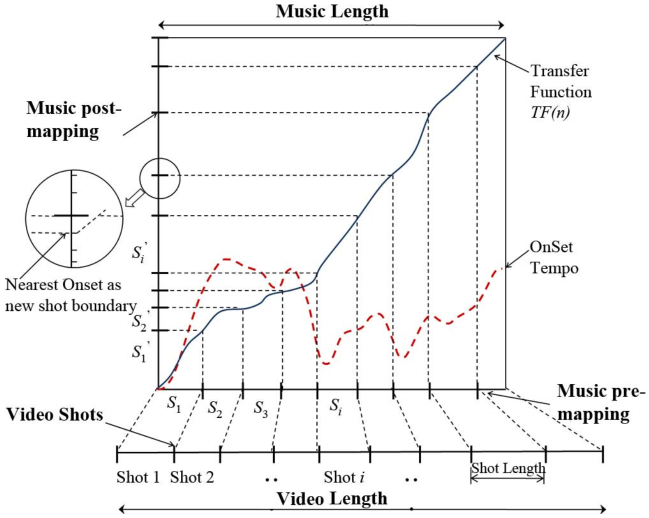

Fig. 9. According to music tempo, we can determine appropriate length of video segments to be selected in the summary.

and the frame number corresponding to this time instant is denoted as $f(\hat{t}_{M_i})$.

The x-axis "music pre-mapping" in Fig. 9 shows division results. This division is dependent solely on the ratio of length of a specific video shot to length of the whole video. However, to generate vivid video summaries, visual rhythm created by shot changes should be correlated with music tempo, that is, music segments of fast tempo should be accompanied with fast shot changes in the generated video summary and vice versa. To accomplish this task, we try to alter the duration of each music segment according to its corresponding tempo information. Motivated by the histogram equalization techniques, a monotonically increasing transformation function $TF(n)$ is designed as

$$TF(n) = \sum_{j=1}^{n} (\text{tempo}_{\max} - \text{tempo}(j) + \delta) \tag{13}$$

where $\text{tempo}_{\max}$ denotes the maximum value of tempo values over the whole music, $\text{tempo}(j)$ denotes the tempo value of the jth frame and δ is a factor that controls the strength of the video rhythm. This transfer function is illustrated as the monotonically increasing curve in Fig. 9. Based on this transformation function, the starting time of the ith music segment is adjusted as

$$\hat{t}_{M_i} = L_M \times \frac{TF(f(\hat{t}_{M_i}))}{TF(f(L_M))}. \tag{14}$$

After transforming starting time of each music segment, the length of a music segment is inversely proportional to the corresponding music tempo. The y-axis "music post-mapping" in Fig. 9 shows the transformed division results. We clearly see that for music segments S_2 and S_3, which have higher tempo values (see the red dashed curve), this lengths of transformed segments S_2' and S_3' are smaller. On the other hand, the music segment S_i with low tempo values is transformed into S_i' to have larger length. For the ith music segment, we would like to find an appropriate clip from the ith video shot. The selected video clips are finally concatenated as the video summary, the length of which is the same as the length of the user's selected music. Note that, with the transformation described in Fig. 9, a shorter video clip would be selected for music segments with high tempo, and a longer video clip would be selected for music segments with low tempo. This makes higher shot change rate for the music segments of high tempo (e.g., S_2 and S_3) and low shot change rate for music segments of low tempo (e.g., S_i). This trend makes video rhythm of the generated video summary matches with tempo of music.

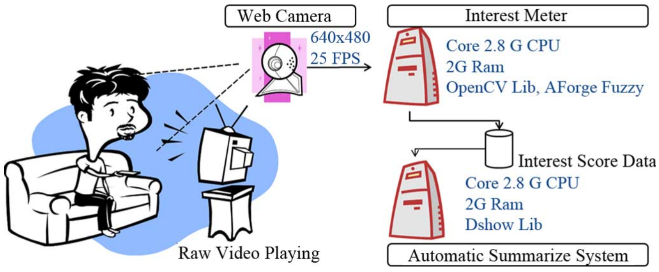

Fig. 10. Environment settings for performance evaluation.

B. Summary Generation

To select the most appropriate clip from the $i\text{th}$ video shot, we would like to find a clip that has the length (in terms of frame number) $\ell_{M_i} = f(\hat{t}_{M_{i+1}}) - f(\hat{t}_{M_i})$ and has the largest sum of interest scores. We apply a sliding window with length ℓ_{M_i} on the interest scores calculated in Section III and calculate the sum of interest score for the jth frame as

$$z_j = \sum_{k=j}^{j+\ell_{M_i}} S_r(k) \tag{15}$$

where $S_r(k)$ is the integrated interest score of the kth frame. The best video clip to match with the ith segment is then determined by finding someone that has the largest score sum. The best video clip starts at the j^*th frame if

$$j^* = \arg\max_j z_j. \tag{16}$$

V. Experimental Result

Here, we describe the performance evaluation from various perspectives. We first evaluate performance of iris detection and blink detection. Then, we verify the idea of the IM by analyzing the user's interest in viewing two pilot video examples. For video summarization, we compare our method with random selection, a novice's manual editing results, and a method based on perceptive analysis. Finally, we present a discussion regarding generality and limitation of our work.

Fig. 10 illustrates the environment setting for performance evaluation. A webcammounted on the monitor is used to capture user's upper body (especially face and shoulder). Resolution of the captured video is 640 × 480 and the frame rate is 25 fps. A PC is used to analyze attention and emotion states of the viewer and output interest scores. According to interest scores, another PC is used to summarize the home video and display summarization results for subjective evaluation.

TABLE II
PERFORMANCE COMPARISON OF IRIS CENTER DETECTION

Method	***Accuracy*** **(*e* ≤ *0.05*)**	***Accuracy*** **(*e* ≤ *0.10*)**	***Accuracy*** **(*e* ≤ *0.25*)**
Our method	80.36%	**97.05%**	**99.65%**
Valenti [31]	**84.10%**	90.85%	98.49%
Tűrkan [32]	19.00%	73.68%	99.46%
Asteriadis [33]	74.00%	81.70%	97.40%
Bai [34]	37.00%	64.00%	96.00%
Campadelli [35]	62.00%	85.20%	96.10%
Hamouz [36]	59.00%	77.00%	93.00%
Cristinacce [37]	56.00%	96.00%	98.00%
Jesorsky [38]	40.00%	79.00%	91.80%

A. Accuracy of Iris Center Location

The BioID database [27] consists of 1521 grayscale images of 23 different subjects with a resolution of 384 × 286 pixels. These facial images were taken during several sessions at different places, i.e., this dataset features uncontrolled illumination and background variations. In addition, the subject may move both in scale and pose. Sometimes their eyes are closed or they turned away from the camera. In many samples, the subjects wear glasses so that the eyes are hidden by the spectacle frames or there is strong highlight on the glasses. The BioID database is usually considered a challenging dataset and the ground truth of left and right iris centers is provided. To evaluate our iris center location method, the normalized error is used, which indicates the error obtained by the worst eye estimation normalized by the distance between two eyes. This measure was proposed by Jesorsky *et al.* [35] as

$$e = \frac{\max(d_{\text{left}}, d_{\text{right}})}{\omega} \tag{17}$$

where d_{left} and d_{right} are the Euclidean distances between the ground truth and the determined locations of the left and right eyes, respectively. These two distances are at most the length of one eye. The value ω is the Euclidean distance between two eyes in the ground truth and its value is at most two times of the length of one eye.

We evaluate accuracy of iris detection with different error tolerance, say $e \leq 0.25$, $e \leq 0.1$, or $e \leq 0.05$. Table II shows comparison of detection performance by different methods. The experimental results of other methods are collected from the literature [27], [35]. Unsurprisingly, the detection accuracy increases as the error tolerance increases. Our method has the best performance when the error tolerance is 0.25 or 0.1. In the case of least error tolerance ($e \leq 0.05$), our method still exceeds others except for the method proposed by Valenti and Gevers [28]. They train a classifier to find the best possible choice out of all candidate iris centers. However, their result is easily influenced by the training set, which is not clearly addressed in [28].

B. Accuracy of Blink Detection

We design an experiment to evaluate accuracy of blink detection. Subjects were invited to sit in front of the computer, approximately 60 cm away from the camera. They were instructed to act naturally but were asked not to turn their heads or move too quickly. False positives and missed blinks are both considered to calculate detection accuracy, which is defined as

$$\text{accuracy} = \frac{\text{true positives in detection results}}{\text{all blinks}}. \tag{18}$$

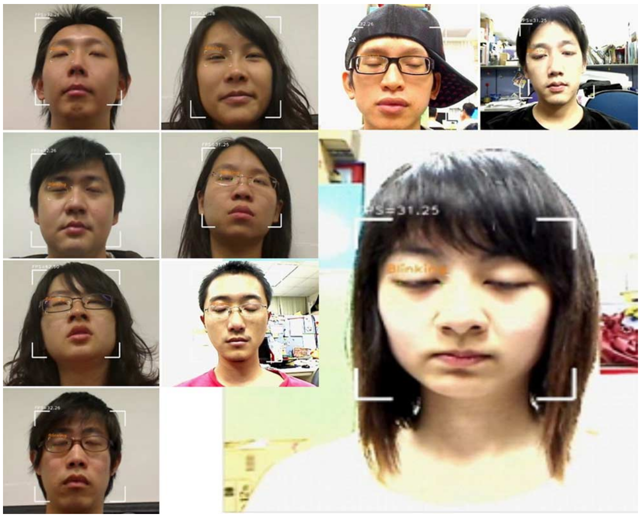

Fig. 11. Sample results of blink detection.

TABLE III
PERFORMANCE COMPARISON OF BLINK DETECTION

Method	**Accuracy**
Our	91%
Chau and Betke [45]	95.3%

We collect 200 blinks in total from ten different subjects to evaluate the blink detection module. In our experiment, there are 13 missed blinks and five false positives, so the accuracy is $182/200 = 91\%$. Sample results of blink detection for these ten subjects are shown in Fig. 11. We clearly see that this method works well even when the subjects are in cluttered background or wear glasses. Table III compares the performance of our method with the one proposed in [42]. Our system achieves an accuracy comparable to Chau's system and provides a solid foundation for attention analysis.

C. Verification of IM

We invited six subjects (four males and two females) to evaluate the IM system. Participants range from 20 to 35 years old. All participants were unaware of the purpose of the experiment. We prepared two testing videos, which were both downloaded from Youtube and were captured by amateur photographers in travel. The length of both videos is about 2 min. Note that both videos have significant motion and unstable exposure due to bad photography skills. Video 1 is composed of normal and funny segments, and Video 2 is composed of normal and attentive segments. The video segments that are generally acknowledged as funny by Youtube users are the so-called *funny* segments. The video segments that obviously attract users (but not necessary funny) are called *attentive* segments. The *normal* segments are generally boring and have no specific topic. These segments are interlaced to form Videos 1 and 2 so that the IM estimation can be clearly verified.

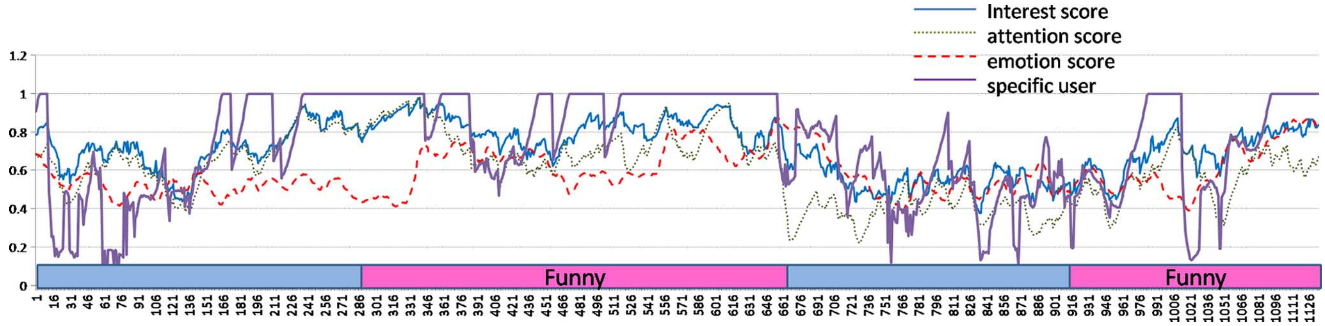

Fig. 12. Average scores of participants in watching Video 1.

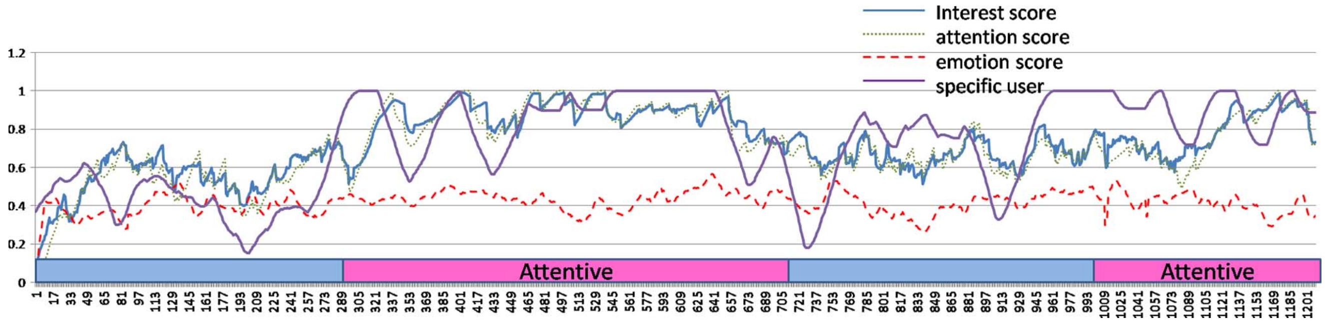

Fig. 13. Average scores of participants in watching Video 2.

When participants watched two videos, the proposed system analyzes their viewing behaviors and calculates attention, emotion and interest scores for each frame. This experiment was designed to verify whether the IM measures the user's interest well. The blue, green, and red curves in Figs. 12 and 13 show the average interest scores, attention scores and emotion scores, respectively. In Fig. 12, we can find that the attention and emotion scores are relatively lower in normal segments and are higher in funny segments. In Fig. 13, although the emotion score has a relatively smooth evolution, the attention score apparently shows how attentive parts attracts human's attention and draws higher scores. Every participant has his own subjective cognition, though they view the same videos. Responses of participants are not necessarily the same for each frame, but we can still examine the difference between funny/attentive and normal segments. In addition to average results, we especially show the evolution of interest score obtained from a specific user as the purple curves. From the purple curves, we see that although response of a specific user seems to tremble all the time, evolution of interest score still matches the trend described above.

TABLE IV
INFORMATION OF THE TEST VIDEOS

Video	Content	Duration	Summary Time
1	Travel	13m 46s	3m 10s
2	Vacation	8m 06s	2m 10s
3	Motor Riding	18m 41s	3m 50s
4	Scenery	10m 58s	2m 10s
5	Wedding	7m 26s	1m 20s

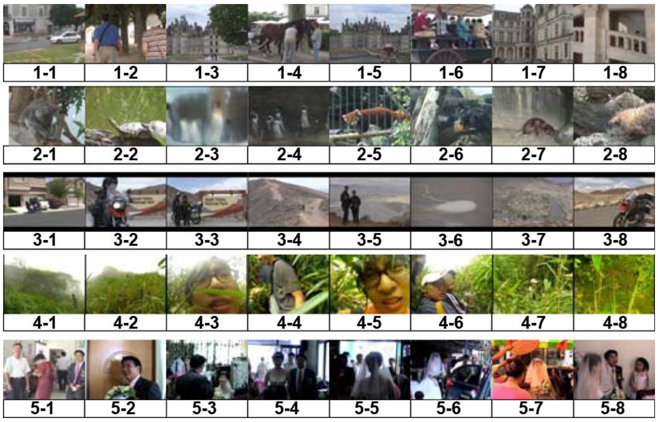

Fig. 14. Snapshots of the first eight shots in each test video.

D. Experiments on Summarization

To generate a personalized video summary, the developed system collects viewing behaviors of different participants and accordingly selects interesting clips to constitute a summary. We invited eight participants (six males and two females) aged between 20 and 28 years old. Participants of this experiment include video providers or those shown in the videos. They were invited to view the test videos and let the system record their eye blink/saccade and facial expressions. With the collected information, we were able to produce personalized video summaries by the proposed methods. The participants were then asked to give a satisfaction score for the generated summaries. The experiment lasts about one and a half hours for every participant.

We evaluated the proposed method based on five video sequences, each of which lasts about 7–18 min. The specification of the test videos are listed in Table IV. Fig. 14 shows snapshots

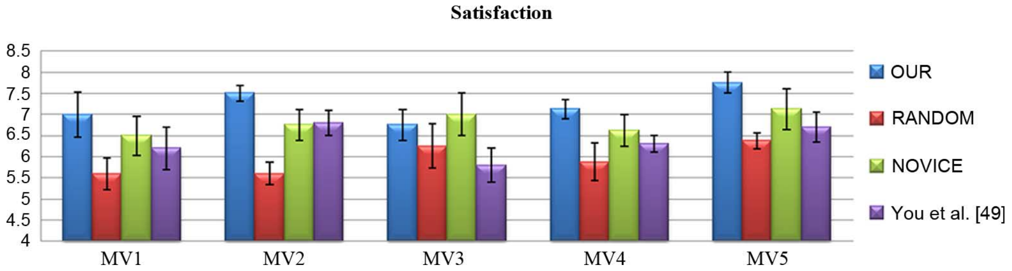

Fig. 15. Satisfaction scores of three methods.

of the first eight shots in the test videos. From the snapshots we see extreme variations in visual content, which covers ill-quality video frames and sometimes nonsense visual content.

Since there is no objective measure to evaluate quality of summarized videos, we compare the automatically generated summaries with: 1) the ones composed of randomly selected shots; 2) the ones manually edited by a novice user who knows about the basic concepts of video editing; and 3) the ones generated by a method based on perceptive analysis [46]. All participants were required to watch four kinds of video summaries and give each a satisfaction score from 1 to 10. Larger score means higher satisfaction. They did not know which summary was generated by which method. Detailed evaluation results are shown in Fig. 15. The satisfaction score results show that our system averagely obtains much higher scores than others. The scores of NOVICE are higher than the randomly selected summaries, which is quite reasonable. Random editing often loses important clips and sometimes ill-quality frames are selected. Our summarization system obtains higher scores than the NOVICE edited results, because the summary generated by the novice just reflects his preference, but not every subject agrees with this result. You *et al.* [46] utilize fusion of motion, color, face, and segment length to estimate human perception in viewing videos and accordingly define importance values of video segments. In contrast to [46], our system captures every subject's viewing behavior, directly analyzes user's interest and summarizes videos by their interests. The generated summaries more appropriately match with each subject's preference. Overall, the automatic method in [46] still works worse than manual editing by the novice. Our system eases editing efforts by allowing users to conduct summarization by viewing the video rather than using a complex interface to do laborious editing works. At the same time, although manual instructions are reduced, the summarization results are better than the fully automatic method in [46].

We also investigate how video content affects satisfaction judgement. In these five videos, the first three videos are irrelevant to participants at all, and Video 3 is commonly viewed as the most boring one. Video 4 was captured by one of the participants and contain activities of some participants. No participant was shown in Video 5, but this video contains activities from friends of some participants. Fig. 15 shows that our approach performs worse for Video 3 because the video content attracts little attention from viewers. On the other hand, for unfamiliar video content that contains interesting objects or events (Videos 1 and 2), our approach works better because more attention and emotion clues can be captured. Comparing performance for videos consisting of familiar people or not, our approach obtains on average a satisfaction score of 7.08 for Videos 1, 2, and 3 and 7.44 for Videos 4 and 5. Participants who were also shown in the test videos or know someone in the test videos tend to give higher satisfaction score. This trend is also true for the NOVICE results. However, the average NOVICE results for the first three videos and the last two videos are 6.75 and 6.88, respectively. We see larger performance improvement in our approach (0.36) than the NOVICE (0.13). This means that our system well captures viewer's intention when someone/something of familiar was shown in videos.

E. Discussion

The main idea of this work is "video editing by viewing." We start the work of video summarization by exploiting viewing behaviors in [41]. However, in this previous work, we simply detect eye movement to evaluate human's interest. In this work, head motion and blink characteristics are further added to more appropriately capture viewing behaviors. This work also proposes that user's interest can be estimated by fusing information from attention and emotion. By incorporating more advanced attention models and emotion recognition approaches, e.g., laughter or surprising sigh from sound information, the proposed approach would be more practical.

Any video domain that easily or apparently draws human's attention or emotion can be analyzed by the proposed system. One of the most appealing domains would be movie videos. Conventional approaches perform movie summarization by analyzing visual variations, object motion or shot change patterns to estimate video tempo. Video shots with higher tempo are then concatenated as a video summary for movies [39]. Although this kind of content-based approach has been adopted for years, the semantic gap problem still impedes the construction of a human-centric video summary. The proposed system directly takes human's viewing behavior into account and generates video summaries by considering human factors.

Another characteristic of the proposed system is personalization. Many personalized summarization methods have been proposed for sports videos or videos with specific objects/events, because in such videos user's preference can be clearly defined [40]. However, conducting personalized summarization for general domain is still not a well-discovered field. We believe that the proposed system provides a new way to approach this task. Parts of interest or disinterest are directly determined from user's reaction, which definitely depends on user's preference and has no standard benchmark.

On the contrary, we have to describe limitation of the proposed method. First, users have to view the whole video at least once to generate a summary. Though users do nothing but view the video, the time spent to generate video summaries is linearly dependent on length of the video. Second, variations of reactions across different users may be large. Subtle reactions or special expression cannot be well detected by imperfect detection modules. Third, not all videos can easily draw user's attention and emotion. For example, users may stay neutral when viewing a documentary about historical events.

VI. Conclusion and Future Work

We propose the idea that user's interest can be measured by the Interest Meter, a computer vision based approach to measure user's interest. In this work, we analyze user's blink, saccade, head motion, and facial expression reactions when he or she views a video. An attention model and an emotion model are then constructed to estimate user's interests based on viewing behaviors. To enrich browsing experience of video summary, the proposed system constitutes a video summary with accompanying music, in which video shots change as significant music beats strike. Satisfactory performance is reported to show that the generated summaries well match with user's interests, as compared to manual editing results and a content-based summarization method.

In the future, we will pay attention to incorporate more human perception factors into this system. For example, results of head orientation recognition or audio cues may be used to capture more human behaviors. In addition, viewing behaviors of multiple users may be analyzed at the same time for different applications.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

[1] M. Argyle, *Bodily Communication*. London, U.K.: Routledge, 1988.

[2] R. W. Picard, *Affective Computing*. Cambridge, MA: MIT, 1997.

[3] H. Zettl, *Sight, Sound, Motion: Applied Media Aesthetics*. Belmont, CA: Wadsworth, 1998.

[4] R. M. Goodman and P. McGrath, *Editing Digital Video : The Complete Creative and Technical Guide*. New York: McGraw-Hill/TAB Electronics, 2002.

[5] G. Chandler, *Cut by Cut : Editing Your Film or Video*. Studio City, CA: Michael Wiese, 2006.

[6] A. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *J. Visual Commun. Image Representation*, vol. 19, pp. 121–143, 2008.

[7] P. Mulhem, M. S. Kankanhalli, H. Hassan, and J. Yi, "Pivot vector space approach for audio-video mixing," *IEEE Multimedia*, vol. 10, no. 1, pp. 28–40, Jan. 2003.

[8] J. Foote, M. Cooper, and A. Girgensohn, "Creating music videos using automatic media analysis," in *Proc. ACM Multimedia*, 2002, pp. 553–560.

[9] X. Hua, L. Lu, and H. Zhang, "Automatic music video generation based on temporal pattern analysis," in *Proc. ACM MultiMedia*, 2004, pp. 472–475.

[10] J. C. Yoon, I. K. Lee, and S. Byun, "Automated music video generation using multi-level feature-based segmentation," *Multimedia Tools Applic.*, vol. 41, pp. 197–214, 2009.

[11] J. Wang, E. Chng, C. S. Xu, H. Q. Lu, and Q. Tian, "Generation of personalized music sports video using multimodal cues," *IEEE Trans. Multimedia*, vol. 9, no. 3, pp. 576–588, Apr. 2007.

[12] A. Money and H. Agius, "Analysing user physiological responses for affective video summarization," *Displays*, vol. 30, pp. 59–70, 2009.

[13] H. Joho, J. M. Jose, R. Valenti, and N. Sebe, "Exploiting facial expressions for affective video summarization," in *Proc. Int. Conf. Image Video Retrieval*, 2009.

[14] W. T. Peng, C. H. Chang, W. T. Chu, W. J. Huang, C. N. Chou, W. Y. Chang, and Y. P. Hung, "A real-time user interest meter and its applications in home video summarization," in *Proc. IEEE Int. Conf. Multimedia & Expo*, 2010, pp. 849–854.

[15] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[16] A. Al-Oayedi and A. F. Clark, "An algorithm for face and facial-feature location based on gray-scale information and facial geometry," in *Proc. Int. Conf. Image Process. Its Applic.*, 1999, vol. 2, pp. 625–629.

[17] H. Gu, G. Su, and C. Du, "Feature points extraction from face," in *Proc. Conf. Image Vis. Computing*, 2003, pp. 154–158.

[18] M. Castrillón, O. Déniz, C. Guerra, and M. Hernández, "Encara2: Real-time detection of multiple faces at different resolutions in video streams," *J. Visual Commun. Image Representation*, vol. 18, no. 2, pp. 130–140, 2007.

[19] V. Vezhnevets and Degtiareva, "A robust and accurate eye contour extraction," *Proc. Graphicon*, pp. 81–84, 2003.

[20] A. Yuille, P. Hallinan, and D. Cohen, "Feature extraction from faces using deformable templates," *Int. J. Comput. Vis.*, vol. 8, pp. 99–111, 1992.

[21] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, pp. 381–395, 1981.

[22] S. Sirohey and A. Rosenfeld, "Eye detection in a face image using linear and nonlinear filters," *Pattern Recognit.*, vol. 34, pp. 1367–1391, 2001.

[23] R. B. Goldstein, E. Peli, S. Lerner, and G. Luo, "Eye movements while watching a video: Comparisons across viewer groups," in *Proc. Conf. Vis. Sci. Soci.*, 2004, Art. ID 643.

[24] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Trans. Syst., Man Cybern.*, vol. SMC-15, pp. 116–132, 1985.

[25] P. Ekman and W. V. Friesen, *Unmasking the Face*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[26] W. Y. Chang, C. S. Chen, and Y. P. Hung, "Analyzing facial expression by fusing manifolds," in *Proc. Asian Conf. Comput. Vis. Conf.*, 2007, pp. 621–630.

[27] *"The BioID Face Database,"* BioID Technol. Research, 2001 [Online]. Available: http://www.bioid.com

[28] R. Valenti and T. Gevers, "Accurate eye center location and tracking using isophote curvature," *Proc. IEEE Comput. Vis. Pattern Recognit.*, pp. 1–8, 2008.

[29] M. Turkan, M. Pardás, and A. E. Cetin, "Human eye localization using edge projection," *Comput. Vis. Theory Applic.*, pp. 410–415, 2007.

[30] S. Asteriadis, N. Nikolaidis, A. Hajdu, and I. Pitas, "An eye detection algorithm using pixel to edge information," in *Proc. Int. Symp. Control, Commun. Signal Process.*, 2006 [Online]. Available: http://www.eurasip.org/Proceedings/Ext/ISCCSP2006/defevent/papers/cr1124.pdf

[31] L. Bai, L. Shen, and Y. Wang, "A novel eye location algorithm based on radial symmetry transform," in *Proc. Int. Conf. Pattern Recognit.*, 2006, pp. 511–514.

[32] P. Campadelli, R. Lanzarotti, and G. Lipori, "Precise eye localization through a general-to-specific model definition," in *Proc. BMVC*, 2006, pp. 187–196.

[33] M. Hamouz, J. Kittlerand, J. K. Kamarainen, P. Paalanen, H. Kalviainen, and J. Matas, "Feature-based affine-invariant localization of faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 9, pp. 1490–1495, Sep. 2005.

[34] D. Cristinacce, T. Cootes, and I. Scott, "A multi-stage approach to facial feature detection," in *Proc. BMVC*, 2004, pp. 277–286.

[35] O. Jesorsky, K. J. Kirchbergand, and R. Frischholz, "Robust face detection using the Hausdorff distance," in *Audio and Video Biom. Pers. Auth*, 1992, pp. 90–95.

[36] *"CyberLink PowerDirector,"* CyberLink Corp. Inc. [Online]. Available: http://www.cyberlink.com/

[37] F. Shipman, A. Girgensohn, and L. Wilcox, "Authoring, viewing and generating hypervideo: An overview of Hyper-Hitchcock," *ACM Trans. Multimedia Computing, Commun. Applic.*, vol. 5, no. 2, pp. 1–19, 2008.

[38] *"MuVee AutoProducer,"* MuVee Technologies Pte. Ltd [Online]. Available: http://www.muvee.com

[39] H.-W. Chen, J.-H. Kuo, W.-T. Chu, and J.-L. Wu, "Action movies segmentation and summarization based on tempo analysis," in *Proc. ACM SIGMM Int. Workshop Multimedia Inf. Retrieval*, 2004, pp. 251–258.

[40] C. Xu, J. Wang, H. Lu, and Y. Zhang, "A novel framework for semantic annotation and personalized retrieval of sports video," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 421–436, Mar. 2008.

[41] W.-T. Peng, W.-J. Huang, W.-T. Chu, C.-N. Chou, W.-Y. Chang, C.-H. Chang, and Y.-P. Hung, "A user experience model for home video summarization," in *Proc. Int. Multimedia Modeling Conf.*, 2009, pp. 484–495.

[42] M. Chau and M. Betke, "Real time eye tracking and blink detection with USB cameras," Boston Univ. Comput. Sci., Tech. Rep. 2005-12, 2005.

[43] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, May 2005.

[44] P. Masri, "Computer modeling of sound for transformation and synthesis of musical signal," Ph.D. dissertation, Dept. Electr. Electron. Eng., Univ. of Bristol, Bristol, U.K., 1996.

[45] S. Dixon, "Onset detection revisited," in *Proc. Int. Conf. Digital Audio Effects*, 2006, pp. 133–137.

[46] J. You, G. Liu, L. Sun, and H. Li, "A multiple visual models based perceptive analysis framework for multilevel video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 273–285, Mar. 2007.

Wei-Ting Peng received the B.S. degree in mechanical engineering, M.S. degree in drama and theatre, and Ph.D. degree from National Taiwan University, Taipei, Taiwan, in 1995, 2004, and 2010, respectively.

His research interests include human-computer interaction, digital content analysis, and image processing.

Wei-Ta Chu received the B.S. and M.S. degrees from National Chi Nan University, Taiwan, in 2000 and 2002, and the Ph.D. degree from National Taiwan University, Taipei, Taiwan, in 2006.

Since 2007, he has been an Assistant Professor with the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan. He was a Visiting Scholar with Digital Video & Multimedia Laboratory, Columbia University, New York, during July-August 2008. His research interests include digital content analysis, multimedia indexing, digital signal process and pattern recognition.

Dr. Chu was the recipient the Best Full Technical Paper Award at ACM Multimedia 2006.

Chia-Han Chang received the B.S. degree in computer science and information engineering from National Central University, Taipei, Taiwan, in 2007, and the M.S. degree from National Taiwan University, Taipei, Taiwan, in 2009.

He is currently a UI engineer with the Digital Entertainment Department, Company, Taipei, Taiwan. His interests include computer vision, image processing, and human–computer interaction.

Chien-Nan Chou received the M.S. degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2009.

He is currently a Research & Development Engineer with MediaTek Inc., Taipei, Taiwan. His research interests include mobile technology, persuasive computing, gaze tracking, and image processing.

Wei-Jia Huang received the B.S. degree from National Chiao Tung University, Hsinchu, Taiwan, in 2006, and the M.S. degree from National Taiwan University, Taipei, Taiwan in 2008, both in computer science and information engineering.

He is currently an Engineer with Industrial Technology Research Institute, Hsinchu. His research interests include computer vision and computer graphics.

Wen-Yan Chang received the B.S. degree from Tunghai University, Taichung, Taiwan, in 1998, the M.S. degree from National Cheng Kung University, Tainan, Taiwan, in 2000, and the Ph.D. degree in from National Taiwan University, Taipei, Taiwan, in 2008, all in computer science and information engineering.

His research interests include computer vision, pattern recognition, image processing and computer graphics.

Yi-Ping Hung (S'84–M'89) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1982, and the M.S. degrees in engineering and applied mathematics and the Ph.D. degree in engineering from Brown University, Providence, RI, in 1987, 1988, and 1990, respectively.

He is currently a Professor with the Graduate Institute of Networking and Multimedia and with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan. From 1990 to 2002, he was with the Institute of Information Science, Academia Sinica, Taipei, where he became a tenured Research Fellow in 1997 and is currently a Joint Research Fellow. He has served as the Deputy Director of the Institute of Information Science from 1996 to 1997 and the Director of the Graduate Institute of Networking and Multimedia, National Taiwan University, since 2007. He was the Program Cochair of ACCV'00 and ICAT'00 and the Workshop Cochair of ICCV'03. He has been an editorial board member of the *International Journal of Computer Vision* since 2004. His current research interests include computer vision, pattern recognition, image processing, virtual reality, multimedia and human-computer interaction.