A REAL-TIME USER INTEREST METER AND ITS APPLICATIONS IN HOME VIDEO SUMMARIZING

Wei-Ting Peng[1], Chia-Han Chang[2], Wei-Ta Chu[3], Wei-Jia Huang[2],
Chien-Nan Chou[2], Wen-Yan Chang[2], Yi-Ping Hung[1,2]

[1] Graduate Institute of Networking and Multimedia,
National Taiwan University, Taipei, Taiwan
[2] Department of Computer Science & Information Engineering,
National Taiwan University, Taipei, Taiwan
[3] Department of Computer Science & Information Engineering,
National Chung Cheng University, Chiayi, Taiwan

Email: d93944004@ntu.edu.tw, wtchu@cs.ccu.edu.tw,
wychang@iis.sinica.edu.tw, hung@csie.ntu.edu.tw

ABSTRACT

In this paper, we propose the Interest Meter (IM), a system making computer conscious of user's reactions, to measure user's interest in real time. The Interest Meter takes account of users' spontaneous reactions when users interact with computers. In this work, we analyze variations of user's eye movement, blink, head motion, and facial expression. Furthermore, we propose an algorithm to combine those signals into interest score and determine important parts of video shots when people watch raw home videos. Experimental result shows that this new type of editing mechanism can effectively generate home video summaries.

***Keywords*—**Interest Meter, video summarization, facial expression, eye movement.

1. INTRODUCTION

Argyle [1] indicated that users will have the following reactions when they feel interested: laughing, more fixations, fewer blinks, and lively movements of shoulders and head-nods. Eye gaze plays an important role in attention because a speaker normally focuses attention on the listener by looking. In emotion, an intuitive and obvious clue of interest is from the facial expression. It has been demonstrated that emotions influence people's attitude towards their current and next action, and there is evidence that they play an essential role in rational decision making, perception, learning, and other cognitive functions [2]. Therefore, the Interest Meter adopts blink detection, saccade detection, head motion detection, and facial expression recognition to measure users' interest.

Despite shooting a home video is a lot of fun, editing videos can be tedious and troublesome. To do a good editing job, in addition to choose the right and convenient software, the user's basic domain knowledge and media aesthetics are also essential [3][4][5]. Commercial video editing software, such as Adobe Premier [6], Sony Vegas [7], or Apple iMovie [8], is equipped with a variety of editing tools. But for novice home users who don't have the above-mentioned prerequisites, the tools can be more confusing than handy.

Based on psychological analysis, Interest Meter is proposed to generate a more intuitive and personalized summarized video. In the experiments, user's reactions in watching a raw home video were analyzed before editing, such as his/her facial expressions, blinks, eye movements, and head motions, so as to know which parts of video clips s/he might be interested in. These parts of clips would then be chosen to generate video summaries. Experimental results show that our video summary system can make an appealing home video summary with ease.

The remainder of this paper is organized as follows. In Section 2, we show an overview of related works. Section 3 describes the details of the Interest Meter implementation. Finally, we demonstrate video summarized result in Section 4, and give the conclusion and future work in Section 5.

2. RELATED WORK

Gaze-X [9] is a context-aware affective multimodal interface that can adapt its interface to the user's emotional state and context in an office scenario environment. Attention Meter [10] is a vision-based input toolkit. It gives users analysis of facial expression, body motion, and attentive activities.

Comparing with the Attention Meter, we detect not only blink but also eyes movement information.

978-1-4244-7493-6/10/$26.00 ©2010 IEEE

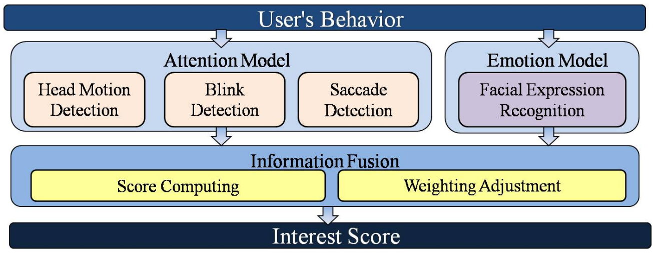

Fig. 1. The framework of Interest Meter.

For facial expression recognition, Attention Meter detects mouth shape such as open wide, close or smiling. In our work, we consider facial local regions and holistic face simultaneously.

To ease home video editing, video summarization has been studied for years. To categorize summary process, from a computer and human interaction perspective, it can be divided as follows: (1) manual, (2) fully automatic, and (3) semi-automatic. Most of the commercial editing software [6][7][8] are manual. Although they provide bountiful functions, even for experts, editing a video on a manual software can still be difficult, and so much more for a novice.

Fully automatic video editing system, such as systems [11][12][13] and automatic editing software Power Director [14], can summarize video through their inbuilt algorithms. Although this takes much less time, users are not able to make changes when they are not satisfied with the results.

Wang et al. [15] proposed a dynamic-programming based algorithm to perform fully or semi-automatic generation of personalized music videos. Shipmanet et al. [16] proposed Hyper-Hitchcock which includes a user interface and various techniques to semi-automatically generate hyper video summaries of one or more videos. MuVee[17], semi-automatic software, is installed with an automatic editing algorithm and a user interface, so that users can adjust the output results of a music video.

Although advanced computer technologies have been created to assist these tasks and developed for years, video summarization remains a hard research topic. Systems described above are all based on content-based summarization [18]. Video clips are often selected because there is high motion or high color/intensity contrast. However, what humans want to see or like to see are not considered. To this end, we propose a new approach to facilitate video summarization based on humans' behaviors when viewing videos. A psychometric model is incorporated into video summarization to accomplish a human-centric system, which was not well acquainted by computer scientists before.

3. INTEREST METER

We define the Interest Meter by two models: attention model and emotion model.

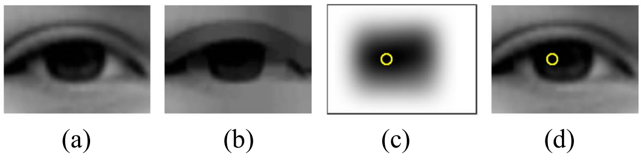

Fig. 2. Detection of eyeball center. (a) Original eye image; (b) Opening operator is applied on (a); (c) Gaussian filter is applied on (b); (d) Result of eyeball center.

Attention describes visual focus of the user, and emotion describes the inner state of the user. Figure 1 illustrates the framework of Interest Meter.

3.1. Attention model

3.1.1. Head motion detection and score calculation

To calculate head motion score, we calculate displacement of face positions between two consecutive frames. The score of head motion can be expressed as:

$$Score_{head\quad motion}(t) = e^{-\frac{(m(t))^2}{\sigma 1}}, \tag{1}$$

where $m(t)$ is the displacement of face position at time t from a previous face position at time t-1, and $\sigma 1$ is a control factor.

3.1.2. Blink and saccade detection

In blink and saccade detection, we adopt three visual features: center of an eyeball, two corners of the eye and the upper eye lid.

For finding center of the eyeball, in the initialization step, opening operator is first applied for eliminating the highlight which may be caused by the reflection on the cornea (Figure 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 (Figure 2(c)). Vezhnevets et al. [19] propose a similar function for the same purpose. We define the function as:

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

where the coefficient 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 the center of eyeball (Figure 2(d)).

To detect corners of eye, we modify the method proposed in [20], which utilizes Gabor wavelets to localize possible corners. They designed the wedge filters based on color information around the corners. The color distribution of the sclera region can be distinguished from the flesh tone in the face.

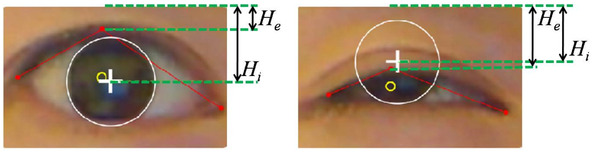

Fig. 3. Blink detection.

The largest wedge looks for the flesh tone and the smallest wedge looks for the sclera tone. A corner, either right or left, is detected if the average value of the pixels in each wedge satisfies its comparative color tone.

Based on the position of eyeball and two corners of the eye, we estimate eye movement situation by comparing the relative distances between eyeball center and eye corners chronologically.

If the velocity of eyeball movement between current frame and previous frame is larger than a threshold, it means a saccade.

To detect the user's eye blinks, we detect that the iris center is occluded by the upper eyelid. However, one blink means an action that a user opens his/her eyes after closing. Whether the iris center is occluded or not is just used to determine the status of the eye at each frame. Let *Blink*(*t*) represent the status of the eye at time *t*.

$$Blink(t)=\begin{cases} Open & if \quad H_i \geq H_e \\ Closed & otherwise, \end{cases} \tag{3}$$

where H_i and H_e are the values of the height from the upper boundary of the eye region to the iris center and the upper eyelid point respectively. Figure 3 shows the two status of the eye. As the eye changes from an open status to a closed status, we ensure that the blink occurs.

3.1.3. Blink score calculation

To calculate the blink score, we first define a blink detection function $b(t)$. If a blink is detected at time t, then $b(t)$=1, $b(t)$=0 otherwise. The score of blink can be expressed as:

$$Score_{blink}(t)=\left\{\begin{matrix} 1, & if \quad \sum_W b(t) \leq 1 \\ 0, & else \end{matrix}\right\}, \tag{4}$$

where W is a one-second sliding window. If there is more than one blink event in this window, it means abnormal blink in the one-second duration.

3.1.4. Saccade score calculation

Goldstein et al. [21] classified eye movement into three categories: fixations, smooth pursuits or saccades.

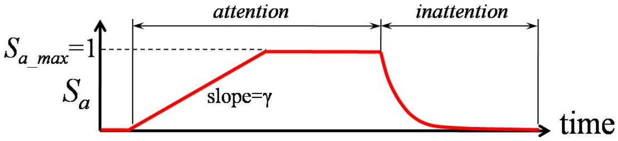

Fig. 4. The illustration of attention score.

They reported that if the moving velocity is larger than 200 degree/second, this period of eye movement is viewed as a saccade. In this work, we take saccades into account because they indicate attention shifting by viewers. The more saccades occur in a shot, the less interesting in this shot for viewers.

We also analyze saccade based on a one-second sliding window W. If a saccade is detected at time t, then the saccade detection function $s(t)$=1; otherwise $s(t)$=0. The score of saccade can be expressed as:

$$Score_{saccade}(t)=\left\{\begin{matrix} 1, & if \quad \sum_W s(t)=0 \\ 0, & else \end{matrix}\right\}, \tag{5}$$

where $s(t)$ is saccade detection function. If saccade is occurred at time t then $s(t)$ is one, $s(t)$ is zero otherwise.

3.1.5. Attention score calculation

In general, attention is a continuous reaction cumulated along time, but it can be lost immediately. Based on this observation, the value of attention in the present frame should change according to the value of the previous adjacent frame. Therefore, we can define attention score as follows:

$$Score(t)=Score_{head\quad motion}(t)+Score_{blink}(t)+Score_{saccade}(t)$$

$$S_a(t)=\left\{\begin{matrix} S_a(t-1)+\gamma, & if \quad Score(t)>\varepsilon \\ \alpha \times S_a(t-1), & else \\ S_a(0)=0 & \end{matrix}\right\}, \tag{6}$$

where *Score*(*t*) is the summation scores of head motion, blink and saccade. If *Score*(*t*) is higher than a threshold ε, it means the user is attentive to the object, and the score of attention increase stably with a slope γ. On the other hand, the score of attention would decrease α (α<1) times original attention score when the user is inattentive. $S_a(t)$ is the attention score at time t. Figure 4 shows the example of attention score.

3.2. Emotion model

When watching videos, users spontaneously express their feelings by facial expressions.

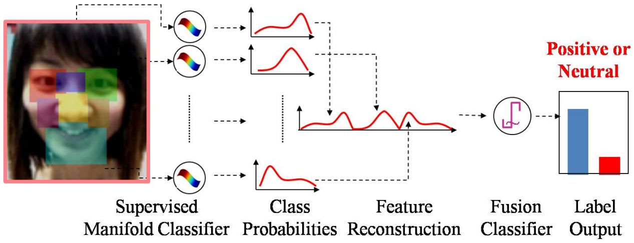

Fig. 5. The process of facial expression recognition.

For example, when something funny happens in videos, most users smile or laugh at what they see. To gain information from such user interests, we adopt facial expression analysis in our work.

For facial expression recognition, instead of analyzing six-class expression [22], we classify human expressions into two categories: positive expression and neutral expression.

A positive expression is defined as a positive human reaction, including smiling and laughing, which implies that the user is interested in this object. Expressions other than positive express are classified as neutral expression.

3.2.1. Facial expression recognition

We adopt a manifold learning and fusion classifier to integrate multi-component information for facial expression recognition. There are totally nine facial components in our work. Given a face image I, a representative feature is constructed by learning the mapping $M : R^d \times c \rightarrow R^t$ based on facial components. Essentially, the mapping M encodes the probability of each expression in facial components and can be defined as

$$M(I) = [m_1(I_1), m_2(I_2), \ldots, m_c(I_c)], \tag{7}$$

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 sub-image of the ith component. By learning the geometry of training data, an embedding function $m_i(I_i)$ can be obtained by projecting I_i onto the learnt manifold. In our framework, a probabilistic representation of $m_i(I_i)$ can be written as:

$$m_i(I_i) = \frac{1}{D^p + D^n}\{D^p, D^n\}, \tag{8}$$

where D^p is the shortest distance between I_i and positive training data, and D^n is the shortest distance between I_i and neutral training data. Based on these formulations, multi-component information is then encoded to a t-dimensional feature vector $M(I)$, in which t is 2×9 =18 in this case.

To characterize the significance of components from the embedded features, a fusion classifier $F : R^t \rightarrow$ {Positive, Neutral} is constructed based on a probabilistic SVM classifier.

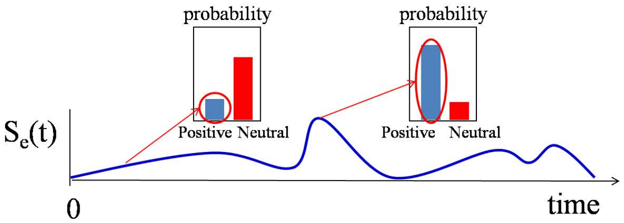

Fig. 6. The calculation of emotion score.

By using this method, users' emotion can be recognized in our system. Figure 5 shows the process of facial expression recognition. Details of facial expression recognition can be found in [23].

3.2.2. Emotion score calculation

Based on facial expression recognition results, we further determine the score of emotion. We use the result of positive probability from facial expression recognition as the emotion score $S_e(t)$, which is ranging from 0 to 1.

By using this formulation, we can determine that higher emotion score represents more important for viewer at time t. Figure 6 illustrates the calculation of emotion score.

3.3. *Interest score computing and weighting adjustment*

The interest score can be described as follow:

$$S_i = W_a \times S_a + W_e \times S_e, \tag{9}$$

where W_a and S_a are attention weight and attention score, W_e and S_e are emotion weight and emotion score, and S_i is interest score.

When positive probability is higher than neutral one in facial expression recognition result, we prefer emotion score to represent the interest score. In this case, we increase the emotion weight and decrease the attention one. When attention score increases, the user starts to concentrate. In this case, we let attention to represent interest more. The formula can be described as:

$$\begin{bmatrix} W_a \\ W_e \end{bmatrix} = (1-\beta) \times \begin{bmatrix} W_{a\_pre} \\ W_{e\_pre} \end{bmatrix} + \beta \times \begin{bmatrix} a \\ b \end{bmatrix}, where(a,b) = (1,0) or (0,1)$$

$$\beta = W_b \times (1 - S_b) + W_s \times (1 - S_s) + W_m \times (1 - S_m), \tag{10}$$

where W_a and W_e are attention and emotion weight. W_{a_pre} and W_{e_pre} are weights from previous frame. W_b is blink weight, W_s is saccade weight, W_m is head motion weight, and $W_b + W_s + W_m = 1$. S_b is blink score, S_s is saccade score and S_m is head motion score. When positive probability is higher than neutral one in facial expression recognition result, we set $(a,b)=(0,1)$, otherwise $(a,b)=(1,0)$.

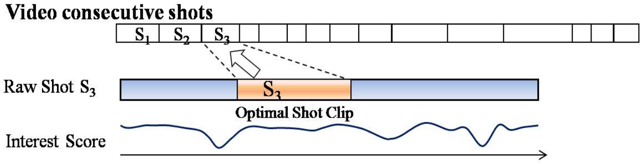

Fig. 7. The illustration of how to choose an optimal shot clip based on interest score automatically.

We use β to control the variance of adjustment amount. When there are more inattentive reactions, the value of β increases and the adjustment amount is enlarged.

3.4. Summary generation

In the process of summarizing home videos, for each raw shot, the interest score S_i is calculated accordingly, and the optimal shot clip with the maximum interest score is selected to be the representative part of this raw shot. Through the processes described above, the selected shots are concatenated as the final video summary. The procedure is illustrated in Figure 7.

4. EXPERIMENTAL RESULT

Test videos were shown on a monitor with a screen that is 40-cm wide. Participants were seated at a distance of about 40-cm from the screen, and the viewing angle subtended by the screen is approximately 52 degrees.

4.1. Experiment overview

All participants were invited to view our test videos and let the system record their eye information 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 higher satisfaction score represent more memorable for the viewer to share the story.

We invited 8 participants (6 males and 2 females) aged between 20 and 28 years old. The experiment lasts about one and a half hours for every participant. Participants of this experiment include video providers or those shown in the videos.

Table 1. The specification of the test videos

	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

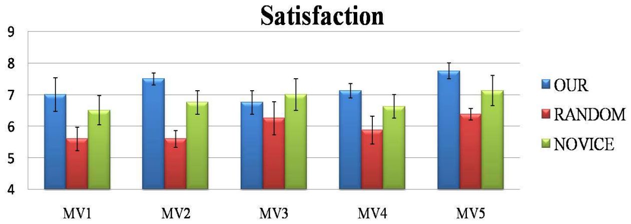

Fig. 8. The satisfaction scores of three methods.

4.1.1. Evaluation data

We evaluated the proposed method based on five video sequences, each of which lasts about 7 to 18 minutes. The specification of the test videos are listed in Table 1.

4.2.2.Procedure

Since there is no objective measure available today to evaluate the quality of summarized videos, we compare the automatically generated summaries with (1) the ones composed of randomly selected shots; and (2) the ones manually edited by a novice user who knows about the basic concepts of video editing. All participants were required to watch these three videos and give a satisfaction score from 1 to 10 to each edited video. Larger score means higher satisfaction. They did not know which summary was generated by which method. Detailed evaluation results are shown in Figure 8. We use satisfaction attributes to evaluate the summarized videos. Satisfaction value means that meaningful clips are reserved more from each shot in a summarized video.

4.3. Results and discussion

The satisfaction score results show that the work of our system strikes much higher scores than randomly selected ones and the edited results by the novice. The scores of NOVICE are higher than the randomly selected summaries, which is quite reasonable. The main reason is that random editing loses more important clips than OUR and NOVICE, and sometimes ill-quality frames are selected. We can also see that our summarization system receives higher scores than the NOVICE edited results. These results show that the Interest Meter can effectively select the shot clips from each raw shot that interest the subjects.

5. 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 the user's interest. In this work, we analyze user's blink, saccade, head motion and facial expression reactions when he or she interacts with the computers and provide interest scores for different applications. Therefore, this system makes a great improvement in video summarization according to interest scores. Satisfactory performance that matches user's interests can be obtained.

In future work, we will pay attention to incorporate with other human perceptions. For example, adding head orientation recognition or extending with modularized sensors. Besides, the Interest Meter can measure only one user at the same time. In the future, it can be extended to measure multiple users at the same time for different applications.

6. ACKNOWLEDGMENTS

This work was supported in part by the National Science Council, Taiwan, under grant NSC 98-2221-E-002-127-MY3, NSC 98-2221-E-002-128-MY3 and by the Construction of Vision-Based Intelligent Environment (Ⅱ), under grant 98-EC-17-A-02-S1-032.

7. REFERENCES

[1] M. Argyle, "Bodily communication," Methuen & Co. Ltd, 1988.

[2] Picard, R.W. "Affective computing," MIT Press, Cambridge, MA, 1997.

[3] H. Zettl. "Sight, sound, motion: applied media aesthetics," Wadsworth, 1998.

[4] R.M. Goodman and P. McGrath. "Editing digital video : the complete creative and technical guide," McGraw-Hill/TAB Electronics, 2002

[5] G. Chandler. "Cut by cut : editing your film or video," Michael Wiese, 2006

[6] Adobe Premiere Pro. http://www.adobe.com/products/premiere/.

[7] Sony Vegas Pro 9. http://www.adobe.com/products/premiere/.

[8] Apple iMovie'09. http://www.apple.com/ilife/imovie/.

[9] Maat, L. and Pantic, M. "Gaze-X: adaptive affective multimodal interface for single-user office scenarios," In Artificial Intelligence for Human Computing, vol.4451, pp. 251-271, 2007.

[10] Lee Chia-Hsun Jackie, Wetzel Jon, Selker Ted, "Enhancing interface design using attentive interaction design toolkit," ACM SIGGRAPH 2006 Educators program, Aug. 2006.

[11] J. Foote, M. Cooper, and A. Girgensohn, "Creating music videos using automatic media analysis," Proceedings of ACM Multimedia, pp. 553-560, 2002.

[12] X. Hua, L. Lu, and H. Zhang, "Automatic music video generation based on temporal pattern analysis," Proceedings of ACM Multimedia, pp. 472-475, 2004.

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

[14] CyberLink PowerDirector, CyberLink Corporation Inc., http://www.cyberlink.com/

[15] J. Wang , E. Chng , C.S. Xu , H.Q. Lu, and Q. Tian, "Generation of personalized music sports video using multimodal cues," IEEE Transactions on Multimedia, vol. 9, pp. 576-588, 2007.

[16] F. Shipman, A. Girgensohn, L. Wilcox, "Authoring, viewing, and generating hypervideo: an overview of Hyper-Hitchcock," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 5, pp. 1-19, 2008.

[17] MuVee AutoProducer, MuVee Technologies Pte. Ltd, http://www.muvee.com/en.

[18] Y.F. Ma, X.S. Hua, L. Lu, and H.J. Zhang, "A generic framework of user attention model and its application in video summarization," IEEE Transactions on Multimedia, pp. 907-919, 2005.

[19] Vezhnevets, V. and Degtiareva, "A Robust and accurate eye contour extraction." Proceedings of Graphicon, pp. 81-84, 2003

[20] S. Sirohey, A. Rosenfeld, "Eye detection in a face image using linear and nonlinear filters," Pattern Recognition, vol. 34, no.7, pp. 1367-1391, 2001.

[21] R.B. Goldstein, E. Peli, S. Lerner, and G. Luo, "Eye movements while watching a video: Comparisons across viewer groups," Vision Science Society, 2004.

[22] P. Ekman, W.V. Friesen, "Unmasking the face," Prentice-Hall, 1975.

[23] W.Y. Chang, C.S. Chen, and Y.P. Hung, "Analyzing facial expression by fusing manifolds," Proceedings of Asian Conference on Computer Vision Conference, pp. 621-630, 2007.