A Real-Time User Interest Meter Based on Human Cognitive Model and its Applications

Chia-Han Chang[1], Wei-Ting Peng[1], Li-Wei Chan[1], Chien-Nan Chou[2], Wen-Yan Chang[2], Zita Chao-Ling Chen[1], Yi-Ping Hung[1,2]

[1]Graduate Institute of Networking and Multimedia
[2]Department of Computer Science & Information Engineering
National Taiwan University

hung@csie.ntu.edu.tw

***Abstract*—In this paper, we propose the Interest Meter, a system allowing computers understand users' reactions, based on multimodal interfaces for measuring users' interests in real time. The Interest Meter takes account of users' spontaneous reactions. In this work, we analyze the variations of users' eye movements, blinks, head motions, and facial expressions when they interact with computers. Furthermore, we propose the method of combining those signals into interest level and verify its reliability in our experiment. There are two integrated applications presented in this work. First, produces each user's personal memory according to their reactions relate to the Magic Crystal Ball. Second, from user's reaction, automatically edits the MV-style home video during watching. Experimental results have showed that the Interest Meter can measure user's interest and make a great improvement in the interaction.**

***Index Terms*—Human Computer Interaction, Affective Computing, Eye Detection, Human Facial Expression Recognition, Human Cognitive Model**

I. Introduction

The goal of Human Computer Interaction is to minimize the barrier between users and computers, that is, to make computers more usable and receptive to users' needs. Communication is an important social contact to understand each other in human society. The first step during communication is to understand reactions of each other. With these two concepts, we can make computers more receptive to users' needs by making computers understand users' reactions.

AIDA [2] is a marketing theory which describes a common list of events that happen very often when a person is selling a product or service. 「AIDA」 means Attention, Interest, Desire, and Action. When a person is selling a product or service, he should attract customers' attention before making them feel interesting. Once customers are interested in the product, they will generate the desire to buy the product. Interest has an influence on determining decision making. Therefore, understanding users' interest leads to know users' reactions. Moreover, interest can trigger emotion, at the same time, emotion can reveal interest. Finally, we propose the idea of measuring users' interest by constructing attention model and emotion model. Attention describes visual focus of the user and emotion de-scribes inner state of the user.

The details of bodily signals (proposed by Argyle, 1988 [6]) were references for the clues of reactions feeling interesting. There are six classes of clues. 1. Facial expression: laughing, more smiling (not false or miserable smile), eyebrows down, circular wrinkles and upward movements of mouth. 2. Gaze: smiling eyes, more gaze, glances, fixation, dilation, less blink, eyes track and look. 3. Gesture: lively movements of hands and shoulders, head-nods. 4. Posture: forward lean, draws back legs. 5. Bodily orientation: more direct, but side by side for some situations of group or pair work. 6. Non-verbal vocalization (tone of voice): higher pitch, upward pitch con-tours, orotund. We consider the first three classes of clues: facial expression, gaze and gesture in our work. Eye gaze plays an important role in attention because a speaker usually presents his focus of attention on a listener by looking. In emotion, the 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. In adiitoin, there is evidence that emotions play an essential role in rational decision making, perception, learning, and other cognitive functions [9]. Finally, the Interest Meter adopts blinking detection, saccade detection, head motion detection, and facial expression recognition for measuring users' interest.

In this paper, we have implemented Interest Meter in two applications: Magic Crystal Ball and MV-Style Home Video Automatic Editing System. In Magic Crystal Ball, users' reactions are continuously captured from a color camera during interaction process. According to these reactions, Interest Meter automatically keeps the clips with interesting reactions of the user and combines these clips into personal memory. In MV-style Home Video Automatic Editing System, users can conduct video editing by "watching videos". When users are watching videos, the Interest Meter indexes the important part of each shot in raw home video according to users' reactions.

In the experiments, we verify the Interest Meter can measure users' interest accurately, and dynamically adjust attention and emotion weight to obtain better performance.

The paper is organized as follows. In paragraph II, we show an overview of related works. The system framework is described in paragraph III and paragraph IV goes to the details of the Interest Meter implementation and asserts the evidence that the experiment results verify the system effect. Finally, we demonstrate two applications mentioned above in paragraph V, and verdict the conclusion and future work in paragraph VI.

II. Related Work

The related work can be divided into two parts. First, we describe the comparison of unimodal interface and multimodal interface. Second, we refer to some real-time affective multimodal interactive applications.

Multimodal interaction has become a key factor in developing novel, effective solutions of natural human machine interaction. In many situations, multimodal interfaces are preferred over unimodal interfaces. Oviatt [10] characterizes that multimodal interfaces satisfy higher user preference levels during interacting with these systems. More flexibility, expressiveness and control ability are available in such interfaces. Also, studies have reported enhanced performance when using multimodal instead of unimodal interfaces. Therefore, the Interest Meter adopts multimodal interfaces.

Gaze-X [5] is a context-aware affective multimodal interface that can adapt to the user's emotional state interface in an office scenario environment. It uses speech, eye gaze direction, facial expression, and keystroke and mouse movement as input factors. One of the major advantages of such a real-life working application is; it can be used as a research tool, to perform real-life experiments for affect measurement or usability aspects. A multimodal affective mirror [1] contain vocal and facial affect-sensing modules, and a component fuses the output of these two modules to achieve a user-state assessment, a user state transition model, and a component presenting audiovisual affective feedback, these should be keep or bring to the user in intended state. The mirror's interaction is to evoke positive emotions, to let people laughing and to increase laughter.

Attention Meter [4], most similar to ours, is a vision-based input toolkit which gives users an analysis of faces found in a given image stream, including face tracking, head motion detection, facial expression recognition. In face tracking, they use a face detection algorithm using the Intel Open Computer Vision library. This algorithm gives the locations and sizes of all faces in the image. In our case, we use the same face detection algorithm, but instead of tracking all faces in the image, we only detect the largest face in each frame taken from the video stream. In head motion detection part, Attention Meter detects large motion, nodding, and shaking by using a finite state machine to analyze sequences of small movements and smaller gestures of nodding and shaking. Different to Attention Meter, our approach only detects the head motion. In eye feature detection part, Attention Meter detects the blink by using basic knowledge of the structure of the face and looking for the distinctive brightness gradients of the eye. In our case, we detect not only blinking but also eyes movement by using more precise algorithm described in Section 3. In facial expression recognition part, Attention Meter only detects mouth shape such as open wide, close or smiling. In our case, we consider facial local regions and holistic face simultaneously.

The whole comparison between Attention Meter and Interest Meter shows in Figure 1. In attention part, the difference is that Interest Meter has saccade detection. In emotion part, Interest Meter uses facial expression recognition but Attention Meter use mouth shape recognition.

	Attention			Emotion	
	Head Motion	Saccade Detection	Blinking Detection	Facial Expression	Mouth Shape
Attention Meter	o	x	o	x	o
Interest Meter	o	o	o	o	x

Figure 1 Comparison of Attention Meter and Interest Meter

III. ATTNTION MODEL IN INTEREST METER

We propose the Interest Meter, a real-time system to measure a user's interest level. Figure 2 illustrates the system framework.

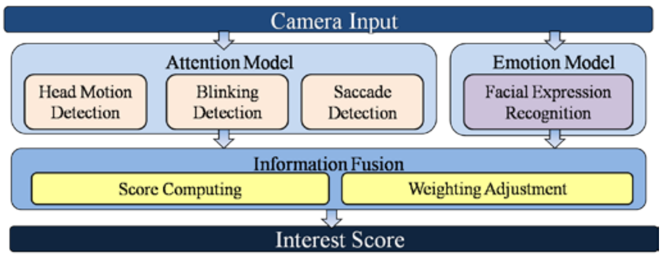

Figure 2 System framework of Interest Meter

As mentioned above, in this paragraph, we define the Interest Meter as two models: attention model and emotion model. The attention model contains head motion detection, blinking detection and saccade detection. The techniques we used in blinking and saccade detection can be found in [3]. In facial expression recognition, based on our previous work [13], we consider the local and holistic face components at the same time. Moreover, we focus on information fusion by combining those detection results into attention and emotion score, and dynamically adjusting attention and emotion weights of interest score. We also conducted the experiment results to verify the functionality of Interest Meter and the efficiency of two weighting adjustment rules.

The attention model is composed of three parts: head motion detection, blinking detection and saccade detection. In head motion detection, we adopt face movement as features. In blinking and saccade detection, we adopt three visual features: the center of the eyeball, two corners of the eye and the upper eye lid. To extract these eye features, face detection [8] is applied in advance for efficiently identifying possible eyes locations. Based on the facial geometry [7], we further simplify the procedure of eye detection only on the possible regions. As the face detection, the cascaded Adaboost is also used for eye detection. To find the center of the eyeball, we apply the Gaussian filter to the image in order to detect the dark circle of the iris. The location with the minimum value is regarded as the center of an eyeball. To detect the corners of the eye, we use the method proposed in [11], which utilizes Gabor wavelets to localize possible corners. After finishing these three detection methods, we can construct the attention score according to these three detection results.

A. Head Motion Detection

Interest Meter monitors the face found in the camera view. Each frame taken from the video stream would run through a face detection algorithm [8]. This algorithm gives us the locations and sizes of all faces in the image. The first process of head motion detection is to detect the face and calculate the face movement. In the second process, we acquire the mapping relationship between face movements and head motion score by adjusting the variance of the Gaussian kernel.

B. Blinking Detection

Figure 3 Definition of blinking

Figure 3 shows the blinking definition, the point Q points to the eyeball center. First, we find the point P by finding first intersection of upper eyelid and line PQ. When the eyeball center Q is covered by upper eyelid P, it means the user with blinking action. In the next step, we translate blinking signal into blinking score. We open a sliding window with one second and analyze this time duration to verify whether multiple blinking happened within one second. If there is more than one blinks found within one second, we label the one-second duration as abnormal blinking state.

C. Saccade Detection

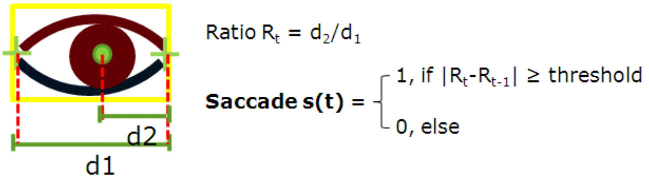

Figure 4 Definition of saccade

We define the ratio d2/d1 to represent the eye movement amount. If ratio difference $|R_t - R_{t-1}|$ between two adjacent frames is bigger than a preset threshold, it means a saccade happened (see Figure 4). Now, we have to convert saccade signal into saccade score. We open a one second sliding window first and analyze this time duration whether it has any saccade detected. If there is any saccade found in one second, we label the duration as abnormal saccade state.

D. Saccade Detection Attention Score Computing

Attention has two properties in our observation. First, people need a period of time to concentrate on paying attention, but they are distracted easily in seconds. Besides, attention is a continuous state, so the attention value of present frame should be determined according to the values of previous adjacent frames. Figure 5 shows the formula we set for the attention score.

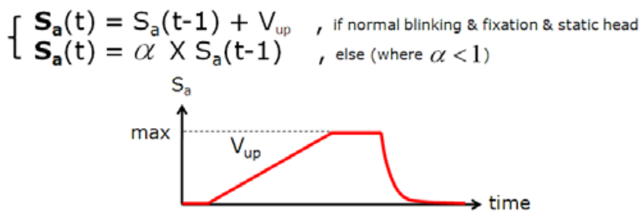

Figure 5 Definition of attention score

The initial score of $S_a(t)$ is set as zero. If there are any blinking, saccade, and head motion reactions found, it means the user is attentive to the object. The score of attention increases smoothly. On the other hand, the score of attention will suddenly degrade α (one-third in our implementation) time of original attention score when the user is inattentive.

III. Emotion Model in Interest Meter

We implement the emotion model by using the facial expression recognition proposed in [13].

A. Facial Expression Recognition

In our work, we only consider two types of emotion, positive and negative. Both local facial components and global face are adopted. We divide a face into seven components including left eye (LE), right eye (RE), middle of eyebrows (ME), nose (NS), mouth and chin (MC), left cheek (LC), and right cheek (RC). In addition, two components, upper face (UF) and holistic face (HF), are also considered. As the method in [13], we adopt manifold learning and fusion classifier to integrate the multi-component information for facial expression recognition. Given a face image I , a mapping M : $R^d \times c \longrightarrow R^t$ is constructed by

$$M(I) = [m_1(I_1), m_2(I_2), \cdots, m_c(I_c)], \qquad (1)$$

,where c is the number of components, $m_i(\cdot)$ is an embedding function learned from the manifold of component i, and I_i is a d-dimensional sub-image of the i-th component. Then, the multi-component information is encoded to a t-dimensional feature vector $M(I)$, where $t \geq c$. To characterize the significance of components from the embedded features, a fusion classifier F: $R^t \rightarrow$ {Positive, Negative} is used based on a binary classifier SVM. By applying this method, users' emotion can be recognized in our system.

After completing LDE models and SVM model construction, we can start to run the facial expression program. When we get each frame from camera, we do face registration and feature extraction of each component. Then, we project each component's feature to the corresponding manifold models and calculate the probability of belonging to each class. Finally, combine all probability as a new feature vector and use this new feature vector as the input of SVM classifier to produce the facial expression recognition result.

B. Emotion Score Computing

We use the positive probability of facial expression recognition result as the emotion score because we only interested in the positive emotion.

IV. INFRMATION FUSION

As above, we have calculated attention and emotion scores. Furthermore, we propose two rules to dynamically adjust weights of attention and emotion scores for better interest measurement. The ideas of these two rules come from our observations of 15 recorded videos during the users watching short videos selected from Youtube website, and Argyle's [6] psychology study. Based on psychology research [6], users will have the following reactions when they feel interested: laughing, more fixation, less blink, lively movements of shoulders and head-nods. On the other hand, users will have the following reactions when they feel bored: blank face, less fixation and more blink. Based on our observations we conducted two rules:

A. Interest = Attention + Emotion

B. Attention usually occurs before emotion

In first rule, we discuss the combination of attention and emotion for high and low situations. Figure 6 shows the four situations of people's emotional state.

When attention and emotion are both high, it means that people are interested in the object. When attention is high but emotion is low, it means that people pay attention to the object but with blank face. When attention is low and emotion is high, it means that people is excited. When attention is low and emotion is low, it means that people feel bored. In second observation, it is intuitive. Before having emotions triggered by an object, people must pay attention to the object first.

A. Interest Score Computing

The interest score formula is as below:

$$S_i = W_a \times S_a + W_e \times S_e, \quad (2)$$

The parameters Wa, Sa, We, Se, Si represent attention weight, attention score, emotion weight, emotion score, and interest score respectively.

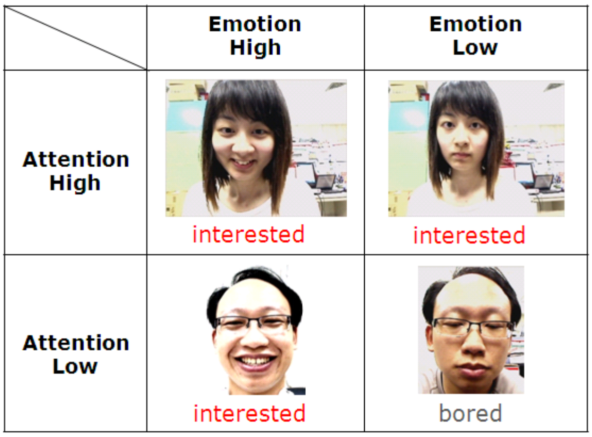

Figure 6 Four situations of people's emotional state

B. Weighting Adjustment

We propose two rules for weighting adjustment. Firstly, when the attention score is decreasing, we analyze the emotion score in weighting adjustment. Secondly, when the attention score is increasing, we consider that the user starts to concentrate gradually.

According to the second observation that attention usually occurs before emotion, we determine attention score first. Why do we adopt decrease and increase policies rather than low and high? Because attention is a continuous reaction, so we consider that adopting decrease and increase policies can measure user's interest better.

As mentioned before, Figure 6 shows four possible combinations of attention and emotion states. In the first column, with high emotion, shows that no matter what the attention value, people must be interested in. In this situation by first rule, we prefer using emotion factor to represent the interest level, so we increase the emotion weight and decrease the attention weight. In the second column, with low emotion, shows that the attention score is the major factor. When the attention score with high value, then people are interested in the object; otherwise, they feel bored with low attention score. In the same way, we prefer using attention factor to represent the interest score, so we increase the attention weight and decrease the emotion weight.

How to adjust the attention and emotion weights? We used the equation shown in below:

$$\begin{bmatrix} W_a \\ W_e \end{bmatrix} = (1-\beta)\times\begin{bmatrix} W_a \\ W_e \end{bmatrix} + \beta\times\begin{bmatrix} a \\ b \end{bmatrix}, \quad where(a,b)=(1,0)or(0,1) \quad (3)$$

$$\beta = W_1\times(1-S_b)+W_2\times(1-S_s)+W_3\times(1-S_m), \quad (4)$$

,where W_a is attention weight, W_e is emotion weight, W_1 is blinking weight, W_2 is saccade weight, W_3 is head motion weight and $W_1+W_2+W_3=1$, S_b is blinking score, S_s is saccade score and S_m is head motion score. We use β to control the variance of adjustment amount. We define the attention score according to head motion, blinking and saccade reactions. Therefore, when the inattentive reactions occur, the β value is increasing, and then the adjustment amount is also increasing. The weights of head motion, blinking and saccade are set with equal weighting.

In the second rule, we assign a higher score to a higher weight, that is, if the attention score is higher than the emotion score, we set attention factor to represent interest level. The formula shows below:

$$W_a = \frac{S_a}{S_a+S_e} \quad , \quad W_e = \frac{S_e}{S_a+S_e}, \quad (5)$$

Where W_a, W_e, S_a, S_e stand for attention weight, emotion weight, attention score and emotion score respectively.

IV. EXPERIMENT

In the experiment, test videos were shown on a monitor with a screen that is 40-cm wide. Participants were seated at a distance about 40-cm from the screen, and the viewing angle subtended by the screen is approximately 52 degrees. There are 6 participants (4 males and 2 females) volunteered in the experiment, from 20 to 35 years old, and they were not informed about the specific purpose in the experiment.

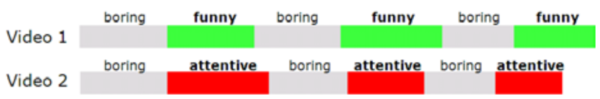

Figure 7 The structures of two test videos

Figure 7 shows the structures of the two test videos. The lengths of these two videos are all 3 minutes and are composed of interesting and boring video clips. All interesting video clips are collected from the Youtube website and the boring video clips are edited from our home videos. All participants are not familiar with the roles appearing in the videos. The difference of these two videos is that the interesting clips in video 1 may trigger participants with emotion reactions, while video 2 may not.

During participants watching two videos, the Interest Meter measures their interest level by analyzing their attention and emotion and calculates the interest score for each frame. There are two goals in our experiments. The first goal is to verify that Interest Meter can measure user's interest. Figure 8 shows the broken line graphs of two measuring results. In Figure 8, the interest scores are lower in the boring segments and higher in the interesting segments. Therefore, we can approximately divide interesting and boring segments in the broken line graph, that is, the Interest Meter can measure user's interest.

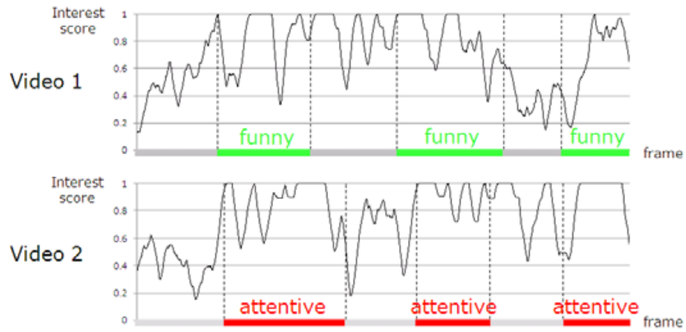

Figure 8 The broken line graphs of

two measurement results

The second goal is to verify that using weighting adjustment rules can measure user's interest better. Firstly, we produce personal boundary because the same source will trigger different levels and different kinds of emotion for each person. The personal boundary is used to eliminate people difference in about emotion. The method of producing personal boundary we used is maximum likelihood estimation. After having the personal boundary, we determine which frame is interesting if the interest score is higher than the boundary.

The motivation of this paper is, to collect the record clip when users are interested in the Magic

Crystal Ball application during the interaction. Therefore, we focus on users' interested responses rather than non-interested reactions. For the reason, we like to know the interesting detection rate of Interest Meter. We define the interesting detection rate as below:

$$\text{Interesting detecting rate} = \frac{\#\text{ of frames that user is detected interested in interesting segments}}{\#\text{ of frames in interesting segments}}$$

After calculating the intersection rate with the ground truth, we can acquire the interesting detection rate. For users' interested states lasting with a period of time, we open a window with three second width. If there is more than one second of interesting clips during this interval, it is assumed as interesting clips.

Table 1 shows the data that using weight adjustment rules can make better performance in interest measurement. There is a gap in the average improvement between video 1 and video 2. For that the interesting clips of video 2 may not trigger user's emotion reaction, the improvement of using weight adjustment rules can not be showed in these clips.

Table 1 The interesting detection rate of test videos

Video 1				Video 2			
User	Rule	Interesting Detection Rate (frame)	Improvement	User	Rule	Interesting Detection Rate (frame)	Improvement
User 1	none	64.8283%	10.718%	User 1	none	78.2561%	**8.4989%**
	all	**75.5463%**			all	**86.7550%**	
User 2	none	59.3132%	11.2383%	User 2	none	79.5806%	0.2207%
	all	**70.5515%**			all	**79.8013%**	
User 3	none	65.2445%	4.7867%	User 3	none	71.0817%	3.2009%
	all	**70.0312%**			all	**74.2826%**	
User 4	none	59.1051%	6.9719%	User 4	none	67.5497%	7.2847%
	all	**66.0770%**			all	**74.8344%**	
User 5	none	70.1353%	**12.2789%**	User 5	none	86.2031%	0.6622%
	all	**82.4142%**			all	**86.8653%**	
User 6	none	62.8512%	12.1748%	User 6	none	83.2230%	0.4415%
	all	**75.0260%**			all	**83.6645%**	
Average	none	**63.5796%**	**9.6948%**	Average	none	**77.6490%**	**3.3849%**
	all	**73.2744%**			all	**81.0339%**	

V. APPLICATIONS

There are two applications introduced in this work, the Magic Crystal Ball and the MV-style Home Video Automatic Editing System.

A. Magic Crystal Ball

Magic Crystal Ball (MaC Ball) is an interactive visual display system which allows the users to see a 3D virtual artifact appearing inside a transparent acrylic ball and to manipulate it with bare hands. The interactive components of MaC Ball contain hand motion and touch detections. In hand motion detection, they construct the MaC Ball coordinate by system calibration to obtain precise 3D directions and distances of hand motions during the interaction process. In touch detection, they integrate strain gauge sensors with a cantilever beam structure into the touch detection module to achieve the better stability of detection. Figure 9 shows the steps of collecting users' interested reactions and displaying their personal memory in the magic crystal ball.

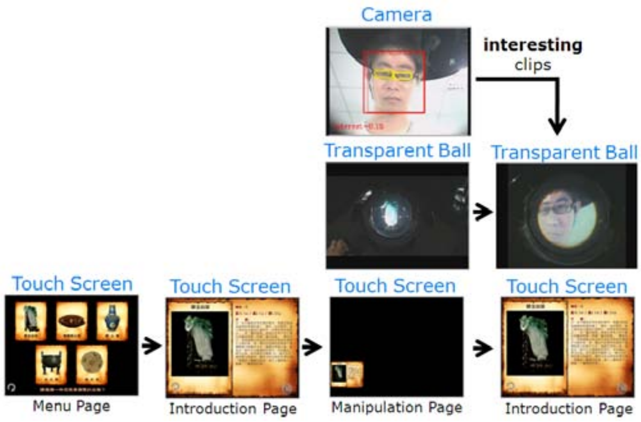

Figure 9 The flowchart of magic crystal ball

The equipment of this application contains a camera for capturing the users' reactions, a transparent ball for manipulating the relic and showing the personal memory, and a touch screen for controlling the interaction process. When a user manipulates the ball with his hands, the Interest Meter keeps the clips shown with high interest scores. Once the user returns to the homepage, his personal memory is presented in the transparent ball. Originally, the Magic Crystal Ball plays a passive role manipulating by the users. After combining with the Interest Meter, it changes to an active role that models users' interest level and interacts with the users by providing a personal memory.

B. MV-Style Home Video Automatic Editing System

MV-style home video automatic editing system, a system developed to automatically analyze video and a user-selected music clip. For video shots, the system eliminates shots with blurred content or drastic motion. For music, the system detects onset information and estimates tempo of the entire melody. With the aids of the editing theory and the concepts of media aesthetics, the system matches selected video shots with music tempo, and therefore facilitates users to make an MV-style video summary that conforms to editing aesthetics without difficulties.

The second application is combing Interest Meter into video editing system for automatically summarizing home video. The Interest Meter takes account of user's spontaneous behaviors when watching videos. Based on users' reactions when watching videos, we can construct a systematic framework to automate video summarization. With the aids of Interest Meter, the developed system can automatically generate a more receptive summarized video. The system architecture is illustrated in Figure 11 and the detail description can be found in [12].

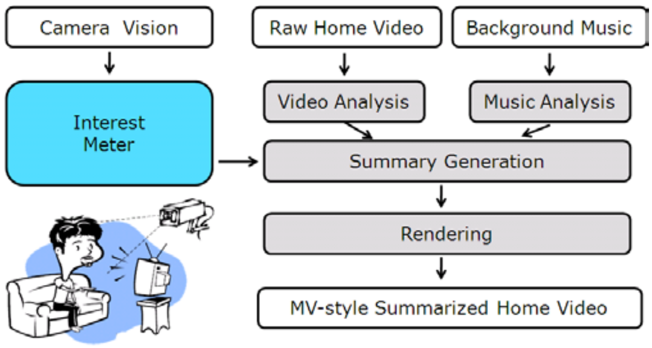

Figure 10 The system architecture of video editing application

VI. Conclusion and Future Work

We propose the idea that people's interests can be measured by the Interest Meter, a computer vision based approach to estimate people's interests. In this work, we analyze users' blinks, saccades, head motions and facial expressions when they interact with computers and provide interest scores for different applications. Therefore, applications can make a great improvement of the interaction by adjusting the interactive contents according to interest scores.

In future work, we will pay attention to incorporate with other human perceptions. For example, considering head orientation recognition or extending with modularized sensors. Moreover, the Interest Meter can be extended to measure multi-users at the same time in different applications, not only one user.

Acknowledgements

This work was supported in part by the Excellent Research Projects of National Taiwan University under grants 98R0062-04.

Reference

[1] A. Melder Willem et al., "Affective multi-modal mirror: sensing and eliciting laughter," ACM HCM, 2007.

[2] B. Alec, Glengarry Glen Ross, Panayiotis Papadakis, 1992.

[3] C. N. Chou, "Real-time three-stage eye feature extraction and its Applications," Master thesis in CSIE at National Taiwan University, 2009.

[4] L. Chia-Hsun Jackie, J. Wetzel, and T. Selker, "Enhancing interface design using attentive interaction design toolkit," ACM SIGGRAPH, Educators program, Boston, Massachusetts, July 30-August 03, 2006.

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

[6] M. Argyle, "Bodily communication," New York: Methuen & Co. Ltd, 1988.

[7] P. Ekman and W. V. Friesen, "Unmasking the face," Prentice-Hall, Englewood Cliffs, 1975.

[8] P. Viola and M. J. Jones, "Robust real-time face detection," International Journal of Computer Vision 57(2), 137–154, 2004.

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

[10] S. Oviatt, "User-centered modeling and evaluation of multimodal interfaces." in Proceedings of the IEEE, Vol.91 (9), 1457-1468, 2003.

[11] S. Sirohey and A. Rosenfeld, "Eye detection in a face image using linear and nonlinear filters," Pattern Recognition 34, 1367–1391, 2001.

[12] 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 Proceedings of International Multimedia Modeling Conference, 2009.

[13] W. Y. Chang, C. S. Chen and Y. P. Hung, "Analyzing Facial Expression by Fusing Mani-folds," in Proceedings of Asian Conference on Computer Vision Conference, 2007.