A User Experience Model for Home Video Summarization

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

Abstract. In this paper, we propose a novel system for automatically summarizing home videos based on a user experience model. The user experience model takes account of user's spontaneous behaviors when viewing videos. Based on users' reaction when viewing videos, we can construct a systematic framework to automate video summarization. In this work, we analyze the variations of viewer's eye movement and facial expression when he or she watching the raw home video. We transform these behaviors into the clues of determining the important part of each video shot. With the aids of music analysis, the developed system automatically generates a music video (MV) style summarized home videos. Experiments show that this new type of editing mechanism can effectively generate home video summaries and can largely reduce the efforts of manual summarization.

Keywords: User experience model, video summarization, facial expression, eye movement.

1 Introduction

With the growing availability and portability of digital video cameras, making home videos has become much more popular. Although there is a number of commercial editing software that helps users to edit videos, not all of them can process a lengthy video easily; even friendly graphical interface and powerful editing functions are provided. Moreover, users need to have much domain knowledge of video editing and should be skilled in using the complicated tools.

Shooting video is fun but editing is proven frustrating. Hence, users incline to put the video footage on the shelf without further intention to elaborately editing. To ease video editing, video summarization has been studied for years. Ma et al. [2] proposed a framework of user attention models to extract essential video content automatically. Hanjalic [3] modeled the influence of three low-level features based on user excitement. Kleban et al. [4] describes contributions in the high level feature and search tasks. Mei et al. [5] further integrated the knowledge of psychology to classify the capture-intents into seven categories.

We also proposed an automatic home video skimming system [1]. In this work, a system was 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[6][7] and the concepts of media aesthetics[8][9][10], 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.

Although the systems described above can achieve satisfactory performance, we found that most of them are based on content-based audiovisual features. Video clips are often unreasonably selected because there is high motion or high color/intensity contrast in them. What human want to see or like to see is not properly considered. To this end, we propose a new approach to conduct video editing in this paper. A novel system based on the human viewing behaviors is developed to generate a home video summary. To our knowledge, the proposed approach is one of the first works to exploit human behavior and analyze users' intention for video editing.

Some studies [11][12]indicate that most people look at the same place all the time while watching movies, because movies consist of a series of shots and are well organized by editors to make a coherent story. Robert et al. [13] recorded the eye movements of twenty normally-sighted subjects as each watched six movie clips. More than half of the time the distribution of subjects' gaze fell within a region which area is less than 12% of the movie scene.

In our case, raw home videos are often not well organized or have clear targets. Therefore, in viewing a home video, humans are forced to move their eyes to search for targets of interest. On the contrary, if humans concentrate their gaze to a fixed region, it indicates that the corresponding video clips have clear targets/topics or have nice shooting conditions. This idea drives us to exploit the behaviors of eye movement in video editing.

In addition to eye movement, we also have to consider user's preference in selecting video clips of interest. Emotion analysis is a practical research issue in many fields. Much attention has been drawn to this topic in computer vision applications such as human-computer interaction, robot cognition, and behavior analysis. In this work, we perform facial expression analysis [14][15][16] to detect where the viewer likes or dislikes the displaying video clip, and use it as the foundation of video summarization.

By integrating eye movement and facial expression analysis, we introduce a user experience model and index the important part of each shot in raw home video automatically. Based on our approach, users can conduct video editing by "viewing videos". This approach makes home video editing more humanistic.

The remainder of this paper is organized as follows. The complete system framework is described in Section 2. Section 3 shows eye movement detection processes, and Section 4 describes the method of facial expression recognition. The development of the user experience model is presented Section 5. Experimental results are reported in Section 6, and conclusions are given in Section 7.

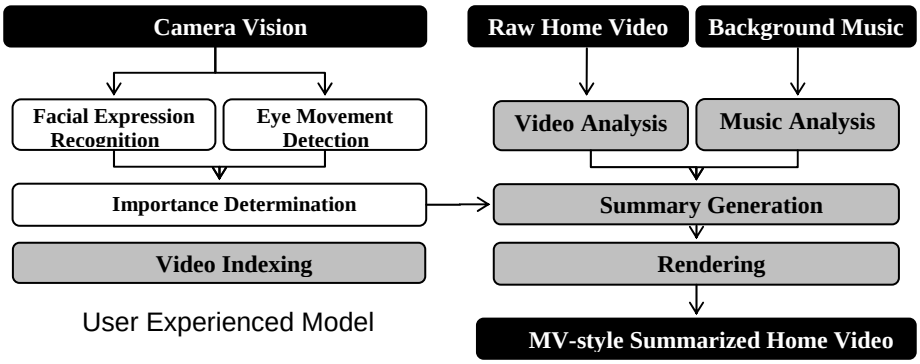

Fig. 1. Structure of the proposed system.

2 System Framework

In our previous work [1], to elaborately incorporate video clips with music, we respectively perform analysis from video and music perspectives. In video analysis, we first drop bad video frames that are ill-lit or blurred, then we segment the video into shots. For background music, we estimate the tempo information based on the occurrence frequency of onsets. We integrated them on the basis of the guidelines of media aesthetics.

In this work, we further integrate the content-based approach with the proposed user experience model. Figure 1 demonstrates the system framework. In addition to content-based importance measure, how humans behave in viewing the raw videos provides the clues about whether the user likes or dislikes the corresponding video clips. From the perspective of video editing, whether users like the video clips indicates the corresponding importance to form the video summary. Therefore, Figure 1 shows that the left part captures humans' behaviors and transforms them into importance measures to facilitate automatic video editing.

Note that the major difference between this work and conventional ones is that we incorporate psychometric model into video summarization, which was not well acquainted by computer science researchers before.

3 Eye Movement Detection

In eye movement detection, we adopt two visual features: the centers of the eyeballs and the corners of the eyes. To extract these features, face recognition is detected in advance and the position of eyes are then located based on the face region.

- **Face detection:** To extract eye movement features, we need to identify where the eyes are. Although we can directly apply an eye detection process to video frames, the search region is too large. Therefore, we perform face detection first

to locate the position of the face, and then we can detect eyes more efficiently. In this system, we exploit the Viola-Jones face detection algorithm [17].

- **Eye detection:** Based on the facial geometry [18], we simplify the procedure of eye detection only on the possible regions. As the face detection, the cascaded Adaboost is also used for eye detection.

- **Feature extraction:** Once the locations of eyes are obtained, we can extract the centers and corners of eyes. For finding the centers of the eyeballs, we apply the Gaussian filter to the image to detect the dark circles of the iris. The center of an eyeball is detected from the location with the minimum value. To detect the corners of the eyes, we use the method proposed in [19], which utilizes Gabor wavelets to localize possible corners. The positions of centers of eyeballs and corners of a video frame represent the characteristics of eyes. The variation of this characteristic along time is the eye movement information. Figure 2 shows the results of eye movement detection in a frame.

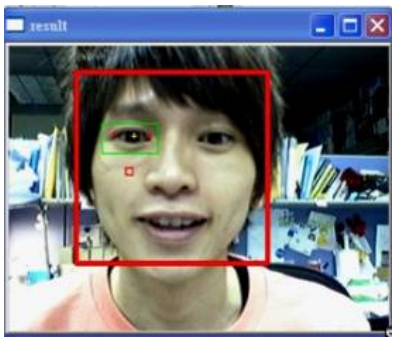

Fig. 2. Results of eye movement detection. The red rectangle represents the location of face. The center of an eyeball and the corners of an eye are presented inside the green rectangle.

4 Facial Expression Recognition

In addition to detecting eye movement, we also incorporate facial expression recognition in our system. Instead of analyzing the six-class expression [20], we only consider two types of emotion, positive and negative, in our work. By recognizing the positive and negative emotions, our system can understand users' intention and recognize video frames that users are interested in.

Recent advances in facial expression recognition have shown that a satisfied performance can be achieved by using hybrid representation [16][21] . Based on these studies, both local facial components and global face are adopted in our work. Besides the components of eyes, nose, and mouth, we also use the areas of middle of eyebrows and cheek to address the wrinkle variations. As the method in [16], we

adopt manifold learning and fusion classifier to integrate the multi-component information for facial expression recognition.

Given a face image I, a mapping $M : R^d \times c \to R^t$ is constructed by

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

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 $F : R^t \to \{Positive, Negative\}$ is used based on a binary classifier SVM. By applying this method, users' emotion can be recognized in our system. Figure 3 depicts the results of facial expression recognition.

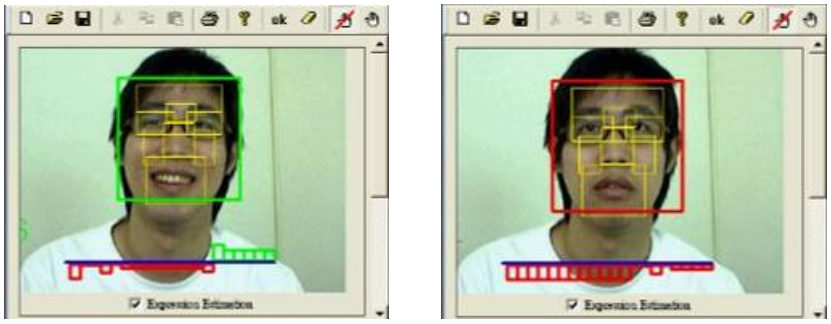

Fig. 3. Results of facial expression recognition. Left: positive expression. Right: Negative expression.

5 User Experience Model

After analyzing facial expression and eye movement, we can define a user experience model to determine important frames of each shot. The details are described in the following sections.

5.1 Importance Determination by Eye Movement

Goldstein et al. [22] classified eye movement into three categories, fixations, smooth pursuits or saccades. 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 the viewers. The more saccades occur in a shot, the lower interesting in this shot for the viewer. According to psychology researches [23][24][25], we can define the importance measure of each shot as follows.

Let $\delta(i)$ represents whether a saccade occurs at the ith video frame.

$$\delta(i) = \begin{cases} 1 & \text{if } v_e(i) > \epsilon, \\ 0 & \text{otherwise}, \end{cases} \quad (2)$$

where $v_e(i)$ is the estimated eye moving velocity, and ϵ is the threshold for saccade detection. Note that due to the limitation of the accuracy in eye tracking and the variant sampling rate of cameras, the threshold ϵ can be adjusted in different situations. The moving velocity is estimated by the difference between two neighboring detected eyeball locations divided by the time duration.

To measure the importance value of each frame, we apply a sliding window W with size $(2w+1)$ to the results of eye tracking. The importance value of the *i*th frame is

$$I_e(i) = (2w+1) - \sum_{k=i-\frac{w}{2}}^{i+\frac{w}{2}} \delta(k). \quad (3)$$

According to this measurement, the importance of the *i*th frame is reduced if more adjacent frames are detected with saccades. In other words, in video summarization, we prefer to skip video frames that the viewer doesn't fix his gaze.

5.2 Importance Determination by Facial Expression

As described in Section 4, we defined the results of facial expression as two types of emotion, positive or negative. Let $\phi(i)$ represent the recognition result of facial expression at the *i*th video frame. We set $\phi(i)$ is 1 if the result is positive. Otherwise, $\phi(i)$ is set to zero.

Because human facial expression doesn't change drastically in a short duration, we apply a sliding window with size $(2w+1)$ to the results of facial expression analysis. By using this strategy, we are able to filter out some noises caused by loss of face tracking and obtain more reasonable results.

The importance value of the *i*th frame is then calculated by

$$I_a(i) = \sum_{k=i-\frac{w}{2}}^{i+\frac{w}{2}} \phi(k). \quad (4)$$

By using this formulation, the frames that the viewer has high positive expression in viewing them will be selected in the video summary.

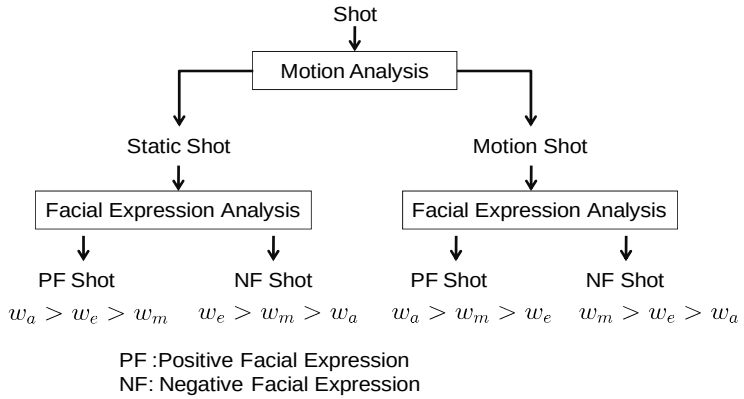

Fig. 4. Illustration of weighting coefficients conditions.

5.3 Importance Fusion

In the above, we have defined the importance for each frame based on human viewing behaviors. Then, we further integrate a content-based feature that substantially describes the importance of frames.

In our previous work [1], we consider camera motion as an important factor. If motion acceleration varies frequently and significantly, the video segment is usually annoying and is less likely to be selected in the video summary. Because the variation of viewer's eye movement and facial expression hardly represent this characteristic, we also take camera motion into account in this work.

Let f_{ij} denote the *j*th frame of the *i*th shot. We estimate the frame's importance values by combining the motion-based importance $I_m(f_{ij})$, the eye-based importance $I_e(f_{ij})$, and the expression-based importance $I_a(f_{ij})$:

$$I(f_{ij}) = w_m \times I_m(f_{ij}) + w_a \times I_a(f_{ij}) + w_e \times I_e(f_{ij}) \quad (5)$$

where w_m, w_a and w_e are weighting coefficients controlling the relative importance of camera motion, facial expression and eye movement. According to our studies, these weighting coefficients can be varied in different situations. For examples, after motion analysis we can label each shot as static or motion (including pan, tilt, and zoom). In motion shots, w_m can be larger than w_e, because eye movement research [12] states that eyes tend to concentrate on the center of screen when humans see videos with rapid moving content. In this situation, w_e doesn't provide useful information. In static shots, w_e can be larger than w_m conversely. Viewers try to search important objects in static shots. If the viewer can't find any attractive targets, then his eyes will move back and forth and produce events of saccades.

Furthermore, we can emphasize w_a when the corresponding shot is detected with positive facial expression. This indicates there is something important in that shot. The priority of these weighting coefficients can be illustrated in Fig. 4.

5.4 Summary Generation

We define the weighting coefficients and calculate the importance values of the frames in each shot after the processes described above. In this section, we will briefly describe the method of summarization generation. Basically, this method is similar to our previous work except we propose a new way to consider human's behaviors in video editing. Details of the summarization method please refer to [1].

According to the tempo of the user-selected music [1], the length of the targeted summarized shot has been determined. Assume that there would be *N* targeted shots in the summary videos, on the basis of music tempo information. In addition, we perform shot change detection for the raw home videos and accordingly obtain *M* video shots, $N < M$, which are called raw shots in the following. Now the problem is to select parts of these *M* raw shots to construct *N* target shots. Before the selection process, the shots with blur or over-exposure/under-exposure are first eliminated. Thus there would be fewer than *M* raw shots to be examined in the selection process.

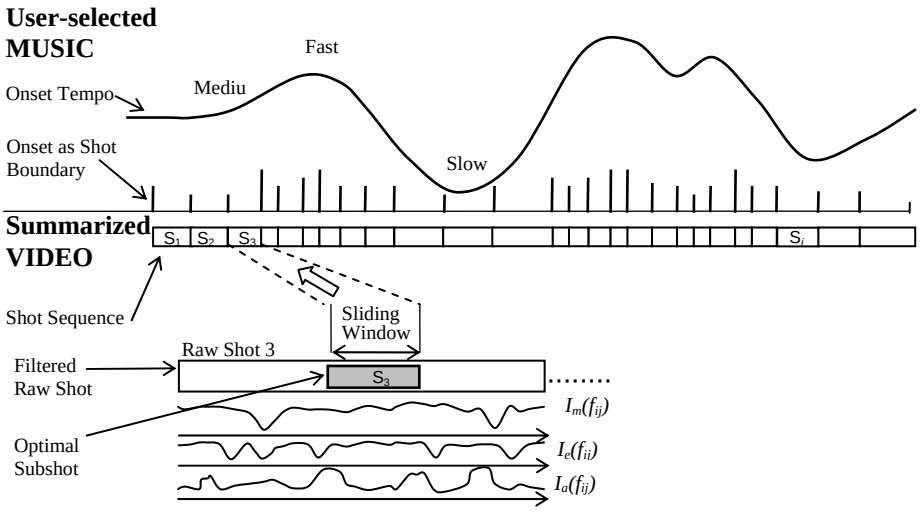

Fig. 5. Illustration of subshot selection.

For each raw shot, we define a sliding window which length is the same as the corresponding targeted shot. Based on this sliding window, the importance value $I(f_{ij})$ is calculated accordingly, and the optimal subshot with the maximum importance value is selected to be the representative part of this raw shot. The process is illustrated in Fig. 5.

Through the processes described above, the selected subshots are concatenated as the final video summary. In this work, human's behaviors play the role of determining the importance of each shot (or each frame further). Using human's behavior as the clues for selection rather than content-based characteristic is the most important contribution of this work.

6 Experimental Results

We describe the implementation framework and experiment settings as follows.

- Implementation Framework

In order to speed up the processes of facial expression analysis and eye tracking, we separate these tasks and respectively handle each of them on one computer. All these computers are connected with Network Time Protocol (NTP) to ensure synchronization. The signal captured from the user's face is forwarded to two computers, as shown in Figure 6.

Test clips 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.

- Participants

We invited 10 subjects (7 males and 3 females) who are students majoring in computer science and volunteered to be in the experiment. Participants are from 20 to 35-year old. All participants were unaware of and were uninformed about the specific purpose of the experiment.

- Evaluation Data

We evaluate the proposed method based on two video sequences, each with the length of about 5 minutes. These two sequences were captured by amateurs and are typical home videos that have worse video quality. The content in the first video is about traveling, and all subjects are not familiar with the people who appear in the video. On the contrary, all subjects are familiar with the people who appear in the second video. The specification of the test videos is listed in Table 1.

Table 1. Evaluation data.

Title	Category	Shot	Time
Video1	Travel	26	5min 41sec
Video2	Mountain Climbing	11	5min 39sec

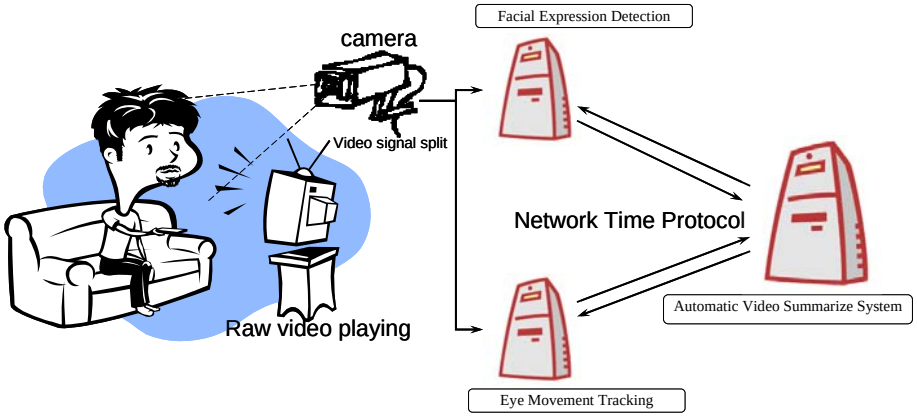

Fig. 6. The distributed scenario in the real-time implementation.

- Evaluation Method

All subjects viewed the videos twice. At the first time, the developed system records their eye movement and facial expression data. At the second time, subjects are requested to manually tag the important part of each shot that would be the ground truth and can be used to judge the goodness of the automatic summary.

In order to understand the influence of eye movement and facial expression, we use three different kinds of summarization methods. The first one is using only the data of eye movement; the second one is using only the data of facial expression; and

the last one exploits both eye movement and facial expression. The experiment last about an hour for one subject.

- Results and discussion

Since subjects had manually tagged the important parts of each shot, we can compare these clips with the generated summaries. Unlike other works that evaluate summary system by subjective scoring, we use a quantitative measurement called "*match rate*" to evaluate our system. Assume that the set of user-selected clips (ground truth) is *G*, and the set of automatically-selected clips is *A*, the match rate is defined as:

$$Match\ rate = \frac{|A \cap G|}{S}, \tag{6}$$

where the $|G|$ denotes the time duration (in seconds) of the set of ground truth and the length of summary time is *S*.

We set the length of summary as 20% of the original videos. Table 2 shows the mean performance of summarization judged by ten subjects, in terms of match rate. We can see that both using eye movement and facial expression are feasible methods to summarize videos. Fusing both factors according to the guidelines describe in Sec. 5.3 would introduce better performance.

To further demonstrate that a user experience model is practicable, we manually label the results of facial expression and combine them to produce summaries. It's not un-expectable that the match rates of the summaries based on true facial expression results are increasing. Although facial expression recognition is still a hard topic, we can see that the performances between UEM and FE in Table 2 are reasonably close. These results lead to the conclusion that the user experience model is quite useful in video summarization.

We also study the usability of video summaries in terms of match rate. According to our study, when match rate of the summarized videos is higher than 50%, the subjects usually feel that it's a good summary. Therefore, we can see that results in Table 2 are not far from appreciation.

Table 2. Experimental results.

UEM: User experience model FE: Facial expression

Title	Type	Eye Match (%)	Face Match (%)	Eye & Face Match (%)
Video1	UEM	33.7	29	34.3
	FE Ground Truth	---	29.9	40.8
Video2	UEM	39.2	44.4	49.9
	FE Ground Truth	---	47.2	61.8

7 Conclusions

A novel system based on a user experience model is proposed in this paper for automatic home video summarization. In this work, we address the variations of viewer's eye movement and facial expression when he or she watches the raw home videos. By analyzing user's intention, our system can automatically select the

important parts of video shots that they are interested. In our experiments, it shows that this new type of editing method can effectively generate home video summaries. A satisfied match rate of viewer's preference in shots also can be obtained. Currently, this work can be treated as the foundation of video summarization based on physiological studies. In the future, we will pay attention to this topic by incorporating other human perceptions.

Acknowledgments

This work was supported in part by the National Science Council, Taiwan, under grant NSC 95-2221-E-002-209-MY3, and by the Excellent Research Projects of National Taiwan University, under grant 95R0062-AE00-02.

References

[1] W.T. Peng, Y.H. Chiang, W.T. Chu, W.J. Huang, W.L. Chang, P.C. Huang, and Y.P. Hung, "Aesthetics-based Automatic Home Video Skimming System," *Proceedings of International Multimedia Modeling Conference*, 2008.

[2] 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 Trans. on Multimedia*, 7(5): 907– 919, 2005.

[3] A. Hanjalic, "Multimodal approach to measuring excitement in video," *Proceedings of IEEE International Conference Multimedia and Expo.*, 2003.

[4] J. Kleban, A. Sarkar, E. Moxley, S. Mangiat, S. Joshi, T. Kuo, and B. S. Manjunath, "Feature fusion and redundancy pruning for rush video summarization," *Proceedings of the international workshop on TRECVID video summarization*, 2007.

[5] T. Mei, X.S. Hua, H.Q. Zhou, and S. Li. Modeling and mining of users' capture intention for home videos. *IEEE TMM*, 9(1):66–77, Jan. 2007.

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

[7] G. Chandler. *CUT BY CUT : Editing Your Film or Video*, Michael Wiese, 2006

[8] H. Zettl. *Sight, sound, motion: Applied media aesthetics*, Wadsworth, 1998.

[9] C. Dorai and S. Venkatesh. *Computational media aesthetics: Finding meaning beautiful*, *IEEE Multimedia*, vol. 8, pp. 10–12, Oct./Dec.2001.

[10] P. Mulhem, M.S. Kankanhalli, H. Hassan, and J. Yi. *Pivot vector space approach for audio-video mixing*, *IEEE Multimedia*, Vol. IO, No. 2, pp. 28-40, Apr-JunZ003.

[11] L. Stelmach and W.J Tam, "Processing image sequences based on eye movements," *Proceedings of SPIE Human Vision, Visual Processing and Digital Display V*, vol. 2179, 1994.

[12] V. Tosi, L. Mecacci, and E. Pasquali, "Scanning eye movements made when viewing film: preliminary observations," *International Journal Neuroscience*, 1997.

[13] R. B. Goldstein, R. L. Woods, E. Peli, "Where people look when watching movies: Do all viewers look at the same place?" *Computers in Biology and Medicine*, 2006.

[14] I.A. Essa and A.P. Pentland, "Coding, Analysis, Interpretation, and Recognition of Facial Expressions," *IEEE Trans. on PAMI*, 19(7): 757-763, 1997.

[15] [13]M.S. Bartlett, G. Littlewort, M.Frank, and C. Lainscsek, "Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 568-573, 2005.
[16] W.Y. Chang, C.S. Chen, and Y.P. Hung, "Analyzing Facial Expression by Fusing Manifolds," *Proceedings of Asian Conference on Computer Vision Conference*, 2007.
[17] P. Viola and M.J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, 57(2): 137-154, 2004.
[18] A. Al-Oayedi and A.F. Clark, "An algorithm for face and facial-feature location based on gray-scale information and facial geometry," *Proceedings of International Conference on Image Processing and Its Applications*, vol. 2, pp. 625-629, 1999.
[19] S. Sirohey and A. Rosenfeld, "Eye detection in a face image using linear and nonlinear filters," *Pattern Recognition*, 34: 1367-1391, 2001.
[20] P. Ekman and W.V. Friesen, *Unmasking the face*, Prentice Hall, 1975.
[21] A. Schwaninger, C. Wallraven, D.W. Cunningham, and S.D. Chiller-Glaus, "Processing of identity and emotion in faces: a psychophysical, physiological and computational perspective," *Progress in Brain Research*, 156:321-343, 2006.
[22] R.B. Goldstein, E. Peli, S. Lerner, and G. Luo, "Eye Movements While Watching a Video: Comparisons Across Viewer Groups," Vision Science Society, 2004.
[23] R.M. Klein and A. Pontefract, "Does oculomotor readiness mediate cognitive control of visual attention: Revisited!" *Attention and performance*, vol. 15, pp. 333-350, 1994.
[24] F. Germeys and G. d'Ydewalle, "The psychology of film: perceiving beyond the cut," *Psychological Research*, 71: 458-466, 2007.
[25] S.P. Liversedge and J.M. Findlay, "Saccadic eye movements and cognition," *Trends in Cognition Sciences*, 4: 6–14, 2000.