Aesthetics-Based Automatic Home Video Skimming System

Wei-Ting Peng[1], Yueh-Hsuan Chiang[2], Wei-Ta Chu[2], Wei-Jia Huang[2], Wei-Lun Chang[2], Po-Chung Huang[2], and 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

Abstract. In this paper, we propose an automatic home video skimming system based on media aesthetics. Unlike other similar works, the proposed system considers video editing theory and realizes the idea of computational media aesthetics. Given a home video and a incidental background music, this system generates a music video (MV) style skimming video automatically, with consideration of video quality, music tempo, and the editing theory. The background music is analyzed so that visual rhythm caused by shot changes in the skimming video are synchronous with the music tempo. Our work focuses on the rhythm over aesthetic features, which is more recognizable and more suitable to describe the relationship between video and audio. Experiments show that the generated skimming video is effective in representing the original input video, and the audio-video conformity is satisfactory.

Keywords: Video skimming, content analysis, media aesthetics.

1 Introduction

With growing availability and portability of digital video cameras, making home videos has become much more popular. But editing home videos remains difficult for most people because most home videos are captured without auxiliary tools, such as tripods, so that there is usually severe shaking and vibration. Insufficient photography knowhow of lighting techniques also results in bad visual quality. Moreover, editing is a skill as well as an art. Without solid editing knowledge and media aesthetics, it is not easy to generate an effective and lively summary video.

Therefore, we propose an automatic home video skimming system which conforms to the editing theory and enables amateurs to make an MV style video without difficulties. An MV style video means a music video accompanied by a piece of background music. Figure 1 illustrates the proposed system framework, which will be described in the following sections.

Video summarization systems have been developed for years, yet many problems still remain to be solved. Ma et al. [1] proposed a framework of user attention

S. Satoh, F. Nack, and M. Etoh (Eds.): MMM 2008, LNCS 4903, pp. 186–197, 2007.

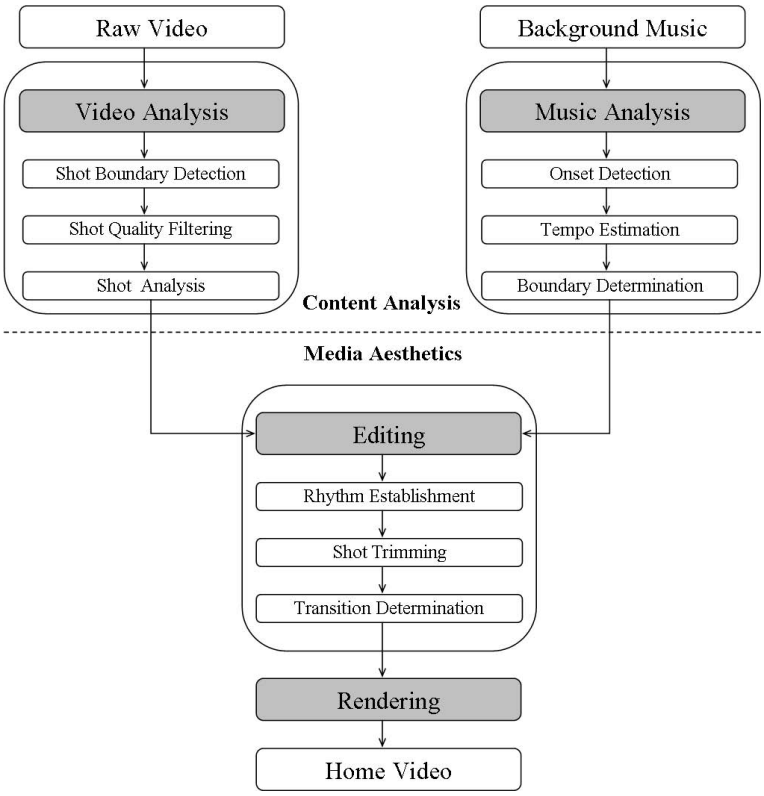


Fig. 1. Structure of the proposed system

models to extract essential video content automatically. Hanjalic [2] modeled the influence of three low level features on user excitement.

There also have been researches about automatic editing systems. Foote et al. [3] presented methods for automatic and semi-automatic creation of music videos. But they merely considered camera motion and image brightness and left out motion blur and ambiguity between underexposure images and night scenes. In [4], their work failed to elaborate on the application of transition effects, which are crucial for video composition and music tempo synchronization. Lee et al. [5] proposed a video pre-processing and authoring technique facilitating music video creation. However, the visual rhythm extraction is not accurate, and the editing theory is not considered in clip selection.

Computational media aesthetics [6] has aroused discussions in recent years. Mulhem et al. [7] used a pivot vector space method to match video shots with music segments based on aesthetic cinematographic heuristics. But their results failed to draw a link between video and audio for viewers. Therefore, our work focuses on the rhythm over aesthetic features, which is more recognizable and is more appropriate to describe the relationship between video and audio.

The remainder of this paper is organized as follows. Section 2 describes the editing theory. Section 3 shows content analysis processes in our system. The aesthetics-based editing method is proposed in Section 4. Section 5 shows experimental results, while conclusions are made in Section 6.

2 Editing Theory

In this section, we will briefly introduce the editing theory, and describe some crucial concepts for automatic editing. They are the foundations of our system. Details can be found in related literature [8,9,10,11,12].

2.1 Motion

The first essential editing concept is motion. Based on Zettl's theory [11], motion can be categorized into three classes: primary motion, secondary motion, and tertiary motion. Primary motion is event motion in front of the camera, including the movements of objects such as people and vehicles. Because extracting primary motion robustly is still an open issue in video analysis, we don't consider primary motion in this paper.

Secondary motion is camera motion. Camera motion analysis in our work will be described in the next section. More specifically, we will introduce pan, tilt, and zoom to clarify the priority rules of editing. Pan and tilt can be considered at the same time because they are similar and only differ in directions. The importance of pan is illustrated in [10]: *"It should be borne in mind that it is better to pan from a weak to a stronger dramatic situation than the reverse—in other words, to build toward strength."* In the case of zoom, zoom in stresses the last subject appearing in a shot, while zoom out emphasizes the scope. Therefore, in terms of automatic video editing, we should prioritize the tail over the rest of a video shot when there is pan, tilt or zoom camera motion.

Tertiary motion is editing motion. It is the visual rhythm caused by transition effects of shot changes. Tertiary motion is crucial for editing aesthetics. But most people just randomly apply transition effects without considering the editing theory. Three kinds of transition effects, i.e. cut, dissolve, and fade, are elaborately selected and incorporated into our system according to the characteristics of consecutive music tempo. These transition effects were depicted in the literature as follows: *"A cut usually generates a staccato rhythm; dissolves generate a legato rhythm"* [11]. *"Longer dissolves will slow the pace of the video, shorter ones will keep it moving quickly"* [12]. *"The function of the fade is to signal a definite beginning (fade-in) or end (fade-out) of a scene or sequence"* [11].

2.2 Rhythm

The length and playback rate of a video shot can be adjusted so that visual rhythm is in conformity to music tempo. We will introduce three methods to change video rhythm in the following.

Rhythmic Control. Lengths of successive shots constitute the rhythm of the output video. Different combinations of shots result in different rhythms. According to the temporal variations of music tempo, we can accordingly vary the visual rhythm by changing the lengths of successive shots.

Cut tight or Cut loose. Concatenating short shots would draw tight visual rhythm. On the contrary, consecutive long shots would moderate the visual presentation. In editing theory, they are respectively called as 'cut tight' and 'cut loose'. For example, in a movie trailer, the editor often applies cut tight editing to increase the visual rhythm and to attract viewers.

Slow and Accelerated Motion. Another method to change the visual rhythm is to adjust the playback rate. In reality, we seldom increase the playback rate, because it is visually unfavorable, and details could be neglected. However, slow motion is a common technique and is included in our system.

3 Content Analysis

Given the input video $\mathcal{V}$ and the background music $\mathcal{M}$, this section describes video and music analysis in our system. The results of these processes are the material for our automatic aesthetics-based editing method.

3.1 Video Analysis

In order to extract most favorable parts of the input video $\mathcal{V}$, we apply the following video analysis, including frame quality estimation, shot change detection, motion analysis, and face detection.

Quality Estimation. Ill-quality frames of the input video are detected and dropped at first. Blur, overexposure, and underexposure are detected in this work. When blur occurs, due to out-of-focus or object/camera motions, edges in the video frame will become indistinguishable [13]. In our system, we use a Laplacian filter to obtain edge intensities, and utilize them to achieve blur detection.

To detect overexposed/underexposed frames, we calculate the mean brightness M_{all}, M_L, and M_D of all pixel, top 10% lightest pixels, and top 10% darkest pixels, respectively. In overexposed frames, most pixels are over lit and have high brightness. So we consider a frame overexposed if both M_{all} and M_D exceed some predefined thresholds. On the contrary, most pixels are dark in underexposed frames, and M_{all} should be low in this case. However, night scene images share this characteristic with underexposed ones. But there are probably some bright pixels such as bulbs or street lamps in night scene images. So we further consider the difference between M_L and M_D to distinguish night scene images from underexposed ones:

$$\begin{array}{l}\text{If } M_{all} \text{ is low and } (M_L - M_D) \text{ is high} \Rightarrow \text{night scene image.} \\ \text{If both } M_{all} \text{ and } (M_L - M_D) \text{ are low} \Rightarrow \text{underexposed image.}\end{array} \quad (1)$$

Video Segmentation. To segment the home video $\mathcal{V}$ into clips, we just use the most well-known histogram-based shot change detection [14], since shot changes mostly occur with sudden cuts, instead of man-made transitions such as dissolve or fade. After dropping ill-quality frames and detecting shot change, we segment the input video $\mathcal{V}$ into N_{shot} filtered shots:

$$\mathcal{V}_{good} = \{\text{shot}_i : i = 1, \ldots, N_{shot}\}. \tag{2}$$

Motion Analysis. We also perform motion analysis to determine the camera motion types (pan, tilt, zoom, or still) with an optical flow based approach [15]. In addition to camera motion types, directions and magnitudes, we advocate considering camera motion acceleration. If motion acceleration varies frequently and significantly, the video segment is usually annoying and is less likely to be selected in the automatic editing phase.

Face Detection. Face information is an important clue to select attractive video segments at the automatic editing stage. So, we apply Viola-Jonse face detection algorithm [16] in our system.

3.2 Audio Analysis

To coordinate visual and aural presentation, shot changes need to be in conformity to the music tempo. So we estimate music tempo of the background music $\mathcal{M}$ at this stage.

We first detect onsets based on energy dynamics. Onsets generally occur when there is significant energy change. We apply the Fourier transform with a Hamming window $w(m)$ to $\mathcal{M}$. The kth frequency bin of the nth frame, $F(n,k)$, of the background music $\mathcal{M}$ can be described as:

$$F(n,k) = \sum_{m=\frac{-N}{2}}^{\frac{N}{2}-1} \mathcal{M}(hn+m)w(m)e^{\frac{2j\pi nk}{N}}, \tag{3}$$

where N is the windows size, and h is the hop size. If the sampling rate of the background music $\mathcal{M}$ is 44100Hz, N and h are set as 2048 and 441 in our system. Spectral flux [17] is one of the onset functions that can measure the changes of magnitudes between frequency bins:

$$\text{Flux}(n) = \sum_{k=\frac{-N}{2}}^{\frac{N}{2}-1} H(|F(n,k)| - |F(n-1,k)|), \tag{4}$$

where $H(x) = (x + |x|)/2$ is the half-wave rectifier function. Then a peak at the nth frame is selected as an onset if it fulfils the peak-peaking algorithm in [18]. Let peak(n) represent this onset detection function. If the nth frame conveys a peak, the output of peak(n) is one. Otherwise, the output is zero. Finally, we

formulate the tempo of the nth frame of the background music $\mathcal{M}$ as the sum of tempo(n) over a local window with size w:

$$\text{tempo}(n) = \sum_{k=n-\frac{w}{2}}^{n+\frac{w}{2}} \text{peak}(k). \tag{5}$$

4 Media Aesthetics-Based Editing

With the shots of the filtered video $\mathcal{V}_{good}$ and the tempo information of the background music $\mathcal{M}$, we are now ready to turn to our aesthetics-based editing method, which consists of three steps: rhythm establishment, shot trimming, and transition determination.

4.1 Rhythm Establishment

Since the lengths of the input video and background music are not necessary the same, the durations of video shots must be adjusted to match the length of the background music, and the visual rhythm caused by shot changes is desired to be synchronous with the music tempo. As we mentioned in Section 2.2, the easiest way to achieve this is exploiting 'cut tight' and 'cut loose'. Figure 2 illustrates how we use a transfer function for this purpose, and details are described as follows.

We first linearly map the shot durations to the length of the background music. The begin time of shot_i in $\mathcal{V}_{good}$ after this pre-mapping process can be written as:

$$t_i^{\text{pre}} = \frac{\sum_{k=1}^{i-1} \text{length}(\text{shot}_k)}{\text{length}(\mathcal{V}_{good})} \text{length}(\mathcal{M}). \tag{6}$$

To synchronize the visual rhythm with the music tempo, we try to alter the duration of each shot after pre-mapping. Motivated by the idea of histogram equalization, we try to design a transfer function, which is monotonically increasing and transforms the starting time of each shot according to the music tempo. The transfer function $TF(n)$ is defined as:

$$TF(n) = \sum_{k=1}^{n} (\text{tempo}_{max} - \text{tempo}(k)), \tag{7}$$

where tempo_{max} denotes the maximum value of all tempo(n). Then, the begin time of shot_i is further mapped according to this transfer function $TF(n)$ in this post-mapping process:

$$t_i^{\text{post}} = \frac{TF(t_i^{\text{pre}})}{TF(\text{length}(\mathcal{M}))} \text{length}(\mathcal{M}). \tag{8}$$

After post-mapping, the visual rhythm caused by shot changes is better synchronized with the music tempo of the background music $\mathcal{M}$.

In order to make shot changes occurred exactly at music onsets in the output music video, we further adjust t_i^{post} to align with its nearest onset peak t_i^{onset}, as shown in Fig. 2.

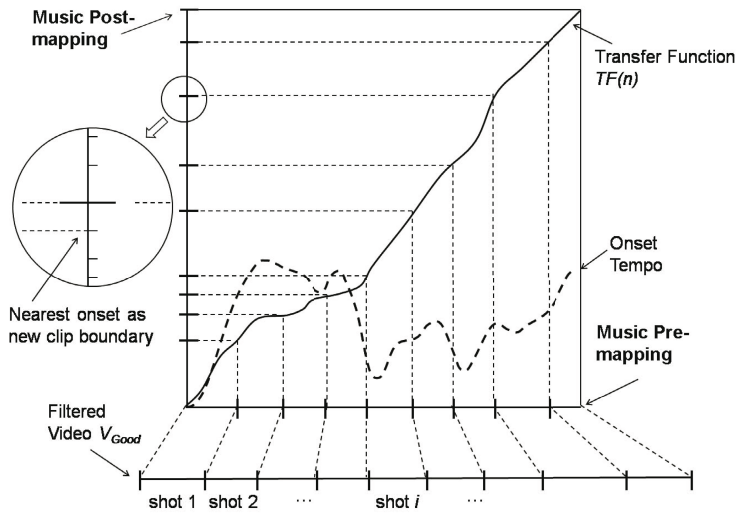


Fig. 2. Illustration of rhythm establishment

4.2 Shot Trimming

After the beginning time, so that the length of each shot in the output music video have been determined, there are still two choices to be decided: how many subshots should be extracted from each shot, and where the most favorable subshots are. We solve these two problems as follows.

We normally do not segment a shot into too many subshots. Because superfluous subshots result in information redundancy and decrease visual aesthetics. Generally, we choose only one subshot for static shots. For motion shots, the higher the music tempo is, the more subshots we extract. And the maximum number of subshots is decided as three in this work.

After the number of subshots for shot_i has been determined, we then select the most suitable subshots based on visual importance. Each subshot within the same shot has the same duration. We denote f_{ij} the jth frame of the shot_i, then we estimate its importance based on face region $\text{Face}(f_{ij})$, camera motion $\text{Motion}(f_{ij})$, and frame temporal position $\text{Pos}(f_{ij})$:

$$\text{Face}(f_{ij}) = \frac{\text{Region}(f_{ij})}{\max_j \text{Region}(f_{ij})}, \tag{9}$$

$$\text{Motion}(f_{ij}) = 1 - \frac{\text{Acc}(f_{ij})}{\max_j \text{Acc}(f_{ij})}, \tag{10}$$

$$\text{Pos}(f_{ij}) = \frac{j}{\text{length}(\text{shot}_i)}, \tag{11}$$

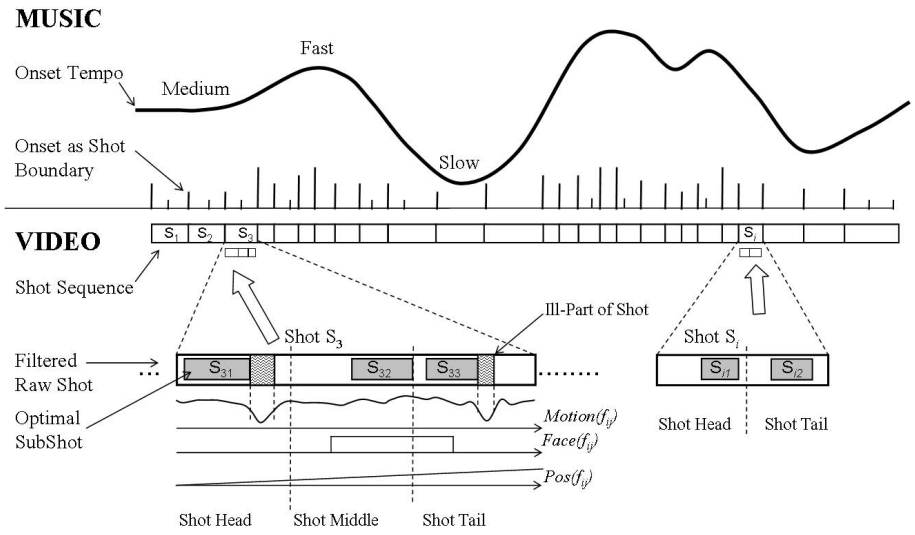


Fig. 3. Illustration of shot trimming

where $\text{Region}(f_{ij})$ and $\text{Acc}(f_{ij})$ denote the face region in frame f_{ij} and the magnitude of camera motion acceleration of f_{ij}. By these three elements, we then formulate the frame importance as:

$$\text{Imp}(f_{ij}) = w_f \text{Face}(f_{ij}) + w_m \text{Motion}(f_{ij}) + (1 - w_f - w_m)\text{Pos}(f_{ij}), \qquad (12)$$

where w_f and w_m are weighting coefficients controlling the importances of face, motion, and temporal position. Then, the optimal subshots with highest frame importance are extracted from different segments of each shot, as shown in Fig. 3.

4.3 Transition Determination

Now we consider transition effects, which occurs in three different situations: at the beginning and the end of the output video, between adjacent shots, and between adjacent subshots. According to the editing theory in Section 2.1, fade-in and fade-out are applied to the beginning and the end of the output video.

For transition effects between adjacent shots, we consider the average music tempo of shot_i:

$$\text{ShotTempo}(i) = \frac{\sum_{n=t_i^{\text{onset}}}^{t_{i+1}^{\text{onset}}-1} \text{tempo}(n)}{t_{i+1}^{\text{onset}} - t_i^{\text{onset}}}. \qquad (13)$$

Each shot is classified as fast, medium, or slow according to its average tempo. Then the transition applied to two adjacent shots is determined according to their shot tempos as Table 1.

When transitions between adjacent subshots are considered, unlike the case of shots, the tempo classes of subshots are derived from their containing shots

Table 1. Transition effects between adjacent shots

$shot_i$	$shot_{i+1}$	Transition
Fast	Fast	Cut
	Medium	Short Dissolve
	Slow	Long Dissolve
Medium	Fast	Short Dissolve
	Medium	Short Dissolve
	Slow	Long Dissolve
Slow	Fast	Short Dissolve
	Medium	Long Dissolve
	Slow	Long Dissolve

Table 2. Transition effects between adjacent subshots

$shot_i$	Transition
Fast	Cut
Medium	Short Dissolve
Slow	Long Dissolve

Table 3. Experiment 1 specification

	MV1	MV2
Video	16m02s	25m15s
Video description	Travel in Europe	Travel in Taiwan
Music	3m40s (popular)	3m08s (piano)

directly, instead of being further classified. Transition effects applied to adjacent subshots in fast, medium, and slow shots are cut, short dissolve, and long dissolve, respectively, as shown in Table 2.

5 Experiments

5.1 Experiment 1 and User Study

In the first experiment, we apply our method to two different kinds of input videos and background musics. The specification is shown in Table 3. We invited seventeen users (eleven males and six females) to compare the results of our system with that of two commercial software, PowerDirector [19] and MuVee [20].

In order to obtain convincing comparison, the evaluators do not know in advance which video was generated by which system. After watching the output music videos, the users are required to answer the following four questions and give scores ranging from one to ten for each question, where ten is the best:

Q1: In visual expression, please rate according to your perceptual satisfaction .
Q2: Do you think the transition effects are comparable with that of a designers touch?
Q3: Do you think the visual rhythm matches the music tempo?
Q4: In general, which result do you prefer?

Figure 4 shows the average scores of the three systems. The results indicate that both our system and MuVee excel PowerDirector in average. In terms of visual expression, our system and MuVee do not differ much, which is probably because MuVee provides fancy atmosphere for travel style and diverts evaluators' attention (**Q1**). Since PowerDirector fails to rule out shaky shots, its output

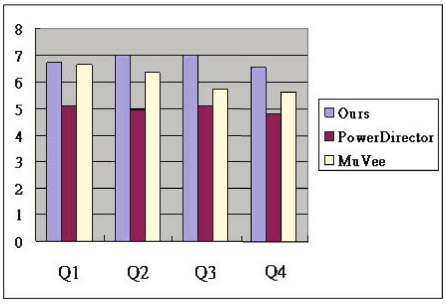
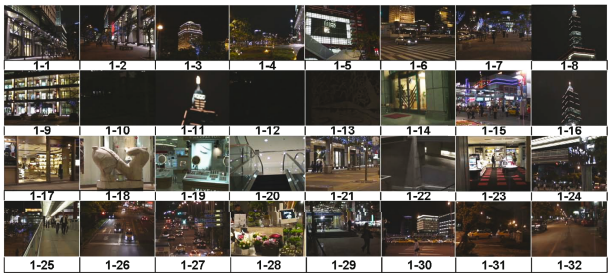


Fig. 4. Average scores of three systems in Experiment 1



(a)

(b)

Fig. 5. Keyframes of every shot in the two videos used in Experiment 2

videos induce visual discomfort. Our system outperforms the others with better conformity between visual rhythm and music tempo (**Q3**), which is one of the main objectives of this work.

5.2 Experiment 2 and Results

In Experiment 2, to compare these systems in terms of quality estimation, we use two videos that contain 68 shots totally, in which 53 shots are captured in night scenes. The night scene shots are likely to be underexposed (eight shots), and blur (five shots) in all clips. The keyframe of each shot is shown in Fig. 5.

Table 4. Detection results of underexposed and blur shots. (S1: Our System, S2: PowerDirector, S3: MuVee, F: fail, D: totally drop)

	Underexposure								Hit	Blur					Hit
	1-10	1-12	1-13	2-6	2-9	2-16	2-21	2-31	(%)	2-25	1-11	1-16	2-12	2-24	(%)
S1	D	D	D	D	D	D	D	D	100	D	F	D	D	D	80
S2	D	D	F	F	F	F	F	F	25	D	D	F	F	F	40
S3	D	D	D	D	D	D	D	D	100	D	D	D	D	D	100

Table 5. False alarm for ordinary night scenes

	Shots that caused false alarm	False Alarm (%)
S1	2-13, 2-22, 2-36	5.7 (=3/53)
S2	1-26, 1-27, 1-32, 2-22	7.5 (=4/53)
S3	1-3, 1-4, 1-7, 1-31, 1-32, 2-13, 2-17, 2-22, 2-23, 2-26, 2-27, 2-28, 2-29, 2-30, 2-34	28.3 (=15/53)

Tables 4 and 5 show the results of the Experiment 2. Both PowerDirector and MuVee can detect and remove most underexposed and blur shots. However, MuVee mis-detects fifteen ordinary night scene shots as underexposed ones, their false alarm rate is about 28.3%, while ours is 5.7%. PowerDirector can better preserve ordinary night shots, at the cost of selecting six underexposed shots. The hit rate of PowerDirector is about 25%, while ours is 100%. Overall, our system outperforms these two commercial software in bad shot removal and night scenes preservation.

6 Conclusions

We proposed an automatic home video skimming system based on media aesthetics. Editing theory is considered in algorithm design and is incorporated into our system. Experimental results demonstrated that our system is superior to two existing commercial software. Low quality shots can be detected and removed automatically and the conformity between video rhythm and music tempo can be achieved at the same time. More video processing techniques can be integrated into our system in the future, such as stabilizing or de-blurring, so that better video clip slection and more lively audiovisual presentation are possible. Moreover, it is also desired to improve possible inconsistent video rhythm when the music tempo changes significantly in a short period of time.

Acknoledgements

This work was partially supported by grants from NSC 96-2752-E-002-007-PAE.

References

1. Ma, Y.F., Hua, X.S., Lu, L., Zhang, H.J.: A generic framework of user attention model and its application in video summarization. IEEE Transactions on Multimedia 7(5), 907–919 (2005)
2. Hanjalic, A.: Multimodal approach to measuring excitement in video. In: ICME (2003)
3. Foote, J., Cooper, M., Girgensohn, A.: Creating music videos using automatic media analysis. In: ACM international conference on Multimedia (2002)
4. Hua, X.S., Lu, L., Zhang, H.J.: Optimization-based automated home video editing system. IEEE Transactions on CSVT 14(5), 572–583 (2004)
5. Lee, S.H., Yeh, C.H., Kuo, C.C.: Home video content analysis for MV-style video generation. In: International Symposium on Electronic Imaging (2005)
6. Nack, F., Dorai, C., Venkatesh, S.: Computational media aesthetics: finding meaning beautiful. Multimedia, IEEE 8(4), 10–12 (2001)
7. Mulhem, P., Kankanhalli, M., Yi, J., Hassan, H.: Pivot vector space approach for audio-video mixing. Multimedia, IEEE 10(2), 28–40 (2003)
8. Goodman, R., McGrath, P.: Editing Digital Video. McGraw-Hill/TAB Electronics (2002)
9. Chandler, G.: Cut by cut: editing your film or video. Michael Wiese (2004)
10. Communication Production Technology: The Pan Shot, `http://www.saskschools.ca/curr_content/cpt/projects/musicvideo/panshots.html`
11. Zettl, H.: Sight, sound, motion: applied media aesthetics. Wadsworth (2004)
12. Loehr, M.: Aesthetics of editing, `http://www.videomaker.com/article/2645/`
13. Tong, H., Li, M., Zhang, H., Zhang, C.: Blur detection for digital images using wavelet transform. In: ICME (2004)
14. Hanjalic, A.: Shot-boundary detection: unraveled and resolved? IEEE Transactions on CSVT 12(2), 90–105 (2002)
15. Dibos, F., Jonchery, C., Koeper, G.: Camera motion estimation through quadratic optical flow approximation. Technical report, Universite de PARIS V DAUPHINE (2005)
16. Viola, P., Jones, M.J.: Robust real-time face detection. IJCV 57(2), 137–154 (2004)
17. Masri, P.: Computer modeling of sound for transformation and synthesis of musical signal. PhD thesis, University of Bristol (1996)
18. Dixon, S.: Onset detection revisited. In: International Conference on Digital Audio Effects (2006)
19. CyberLink: PowerDirector, `http://www.cyberlink.com`
20. muvee Technologies: muvee autoProducer, `http://www.muvee.com`