A Musical-driven Video Summarization System Using Content-aware Mechanisms

Chen-Hsiu Huang
Dept. of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan (R.O.C)
chenhsiu@cmlab.csie.ntu.edu.tw

Chi-Hao Wu
Graduate Institute of Networking and Multimedia
National Taiwan University
Taipei, Taiwan (R.O.C)
nbadream@cmlab.csie.ntu.edu.tw

Jin-Hau Kuo
Dept. of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan (R.O.C)
david@cmlab.csie.ntu.edu.tw

Ja-Ling Wu
Graduate Institute of Networking and Multimedia
National Taiwan University
Taipei, Taiwan (R.O.C)
wjl@cmlab.csie.ntu.edu.tw

***Abstract*—In this paper, we propose a music-driven summarization system for home videos based on several content-aware mechanisms. Many audio and video features are employed to help analyzing and synchronizing input audios and videos. The synchronization is conducted by matching the rhythm of the video with that of the audio. Four profiles for synchronizing video with audio are proposed, which provide users more flexibilities in conducting the synchronization process. Experiments show that good subject test result for the summarized home video can be obtained.**

I. INTRODUCTION

With the rapid development of information technology, the digital capturing devices have been made affordable for end users. The result is that the cost per digital creation is quite low. There is a big challenging problem when people create huge amount of digital contents: how to efficiently manage those digital contents? For image data, the prototype of digital album system has been proposed to solve this problem[1,2]. However, the video data are hard to manage due to the huge data volume and high temporal complexity. In order to manage videos, efficient video editing tools are a must. Actually, there are indeed several powerful video editing softwares available in the market, such as Adobe Premiere[3], Ulead MediaStudio[4], and CyberLink PowerDirector[5]. However, they are too complex to be used for most end users. The phenomenon inspires us to develop a video summarization system which helps users manage their videos efficiently without the skills of using complicated tools mentioned above. The proposed system extracts the quintessence of home videos according to the context of a chosen music.

The remainder of the paper is organized as follows: In Section Ⅱ, we address the problem and present an overview of our system. It includes three stages: media analysis, media synchronization, and scripts generation. Sections Ⅲ and Ⅳ respectively describe the two kernels of the media analysis - audio analysis and video analysis. Section Ⅴ presents the synchronization mechanism between the audio and the video in detail. In Section Ⅵ, we describe how to evaluate the performance of the proposed system. Finally, conclusion is addressed in Section Ⅶ.

II. OVERVIEW OF THE PROPOSED SYSTEM

Usually, people are more impatient of watching videos without scenarios or voice-over, especially for those without dubbing (e.g. home videos). In this case, users always try to move the seek-bar to skip the unimportant and boring parts.

The proposed system is to summarize a lengthy home video into a short one and combine it with a piece of music in a fully-automatic way. In order to make the resulted music video professional-looking, the summarized video and the chosen audio are not combined together blindingly. The summarized videos are generated based on the principle that the rhythms of the video must fit that of the audio

A. System Overview

The block diagram of proposed system is shown in Figure 1. Three consecutive stages are included in the system:

1) Media analysis stage:

The stage covers audio and video analyses, which will respectively be described later in Sections Ⅲ and Ⅳ. The functionality of this stage is to divide audio and video contents into several clips and shots. They are the basic units for composing the resultant music video.

2) Media synchronization stage:

In the media analysis stage, the segmented audio clips and video shots are processed according to some selected features. By matching up with the input audio's rhythm and tempo, selection and abstraction of video shots are conducted. For different kinds of applications, four presentation profiles are proposed. Video shots are automatically edited according to the selected profile. The output video is the same length as that of the chosen audio. Detailed processing steps for media synchronization will be addressed in Section Ⅴ.

This work was partially supported by the CIET-NTU(MOE) and National Science Council of R.O.C. under the contract No. NSC93-2622-E-002-033, NSC93-2752-E-002-006-PAE, NSC93-2213-E-002-006.

3) Scripts generation/production stage:

Finally, the summarized video clips can be saved in some kind of script format. This work allows users to adjust the clips or save them as video files directly. In our work, we choose the AviSynth[6] standard as our script format.

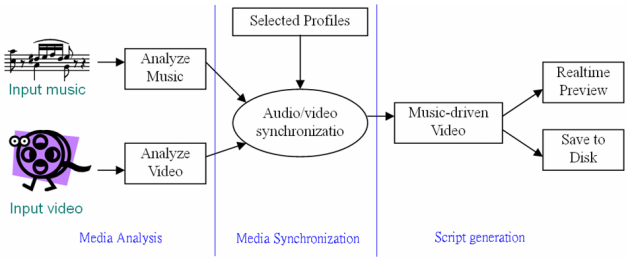

Figure 1. The diagram of proposed system

III. AUDIO ANALYSIS

A. Audio Feature Extraction

Audio features are extracted in two levels: short-term frame level and long-term clip level. Audio frame is defined as a group of neighboring samples which last about 10 - 40 ms. Signal in a frame is assumed stationary .Features like volume and Fourier transform coefficients are extracted in each frame. In our system, we use only time-domain frame-level features to analyze the whole audio as follows:

- Volume: The most widely used and easy-to-compute audio frame feature. Normally volume is approximated by the Root Mean Square (RMS) value of the signal magnitude within each frame.
- Zero Crossing Rate (ZCR): To compute the ZCR of a frame, we count the number of times that the audio waveform crosses the *x* axis.

B. Audio Clip Segmentation

The functionality of audio segmentation is to divide the whole audio stream into smaller units. We use spatial differences (i.e. volume changes) and the occurrence of attacks to conduct audio clip segmentation. First, the input audio is segmented into coarse clips according to the volume changes. The attacks within each clip then cut a clip into finer sub-clips. An attack is defined as the impulse or the release of audio signals. Through observations, we found that the peaks or valleys of the curve ZCR approximate attacks quiet well.

IV. VIDEO ANALYSIS

A. Shot Detection

We apply two methods together to detect shot change. The first one is the pixel domain approach that calculate the Minimal Absolute Difference (MAD) of two neighboring frames. The second one is a histogram domain approach which examines the color histogram difference between two neighboring frames. The shot changes can be detected as follows:

$$SC(m)=\begin{cases}1, & MAD(m,m+1)>Threshold_{MAD} \text{ and } HistD(m,m+1)>Threshold_{hisD}\\ 0, & \text{otherwise}\end{cases} \quad .1$$

where *MAD(m, m+1)* and *HistD(m, m+1)* respectively denote the MAD and histogram difference of two consecutive frames *m* and *m+1*. Th_{MAD} *and* Th_{hisD} are two pre-defined thresholds.

B. Video Features Extraction

Actually, video abstraction[7,8] lies in how to extract the representative image sequences from the original video such that the production result could be "important to the human perception enough". By observations, we used some video features that capture the amateur video-taker's intention for producing home video contents. The features are categorized into high-level, medium-level, and low-level respectively.

1) High-level features

Appearance of *human faces* in video frames certainly attracts users' attention. Thus, we regard the face feature as one of the high-level features. In our implementation, we use the Intel OpenCV library [9] to do the task of human face detection. Besides, we also use a skin-color detector to eliminate the false alarms and refine the face object detector[10].

Flashlight usually means that important event is happening and should be weighted with higher priority. The detection of flashlight is also performed when detecting shot changes. Actually, the flashlight will cause the situation of two consecutive shot changes due to its luminance variations. If two consecutive shot changes are detected and the luminance changes exceed a threshold, it will be denoted as a flashlight event.

2) Medium-level features

When watching video program, people tend to pay more attention to those frames with more motion. We conduct the motion features as below.

The motion information is computed base on motion vectors, which can be extracted from videos encoded in MPEG-like coding standards. The *motion strength* is defined as the sum of motion vector magnitude in each macroblock

Camera motions are widely used in general videos (i.e. zoom in, zoom out, pan, etc.). The camera motion type represents highly semantic information. We calculate the camera motion types from block-based motion vectors. We compare the magnitude of average of motion vectors and the average of magnitude of motion vectors, $\Sigma|v_i| / |\Sigma v_i|$ of each frame. If the ratio is between 1 and 3, we say that the frame has pan operation. If the ratio is greater than 3, it has zoom operation.

3) Low-level features

In order to filter out the under-exposure or over-exposure frames, we calculate the mean value of each frame in the luminance plane (i.e. video *frame's brightness*). A properly exposed video frame's brightness falls around 130.

Colorful video frames are more attractive. The feature *color variance* helps to select colorful frames. We use histogram distributions to model the color variance. The luminance values are quantized into 64 bins, and then we calculate the standard deviations of all bins. The frames with low standard deviation value represents the colorful frames.

C. Importance Functions

1) Frame-level importance function

We define the frame-level importance function as follows

$$\begin{aligned} Imp = & Coeff_H \times (R_{face} + E_{flash}) + \\ & Coeff_M \times (R_{motion} \times S_a + CameraType) + \\ & Coeff_L \times (\left| \frac{Mean - 130}{255} \right| + R_{ClrVar}) \end{aligned} \quad (1)$$

where

$$R_{face} = \frac{Area_{face}}{W \times H}, R_{motion} = \frac{Motion_i}{\max(Motion)}, R_{ClrVar} = \frac{ClrVar_i}{\max(ClrVar)}$$

denote the ratios of face, motion, and color variance respectively. $Coeff_H$, $Coeff_M$, and $Coeff_L$ are weighting coefficients which are set empirically to be 0.5, 0.3, and 0.2. E_{flash} is either 1 or 0. *CameraType* is set to 2 for zoom operation, 1 for pan operation, and 0 otherwise. S_a is a scaling factor which is defined as:

$$S_a = \frac{A_{Dynamic}(C_i)}{\max(A_{Dynamic}(C_i), \forall C_i)} \quad (2)$$

where $A_{Dynamic}(C_i)$ is the dynamic of the *ith* audio clip, and the dynamic is defined as the number of attacks in an audio clip.

2) Shot-level importance function

Besides the frame-level importance, we also define the shot-level importance function to identify which shots are more important than others. The function is defined as:

$$\begin{aligned} Imp = Len \times (\frac{Num_{face}}{Len} + \frac{\sum CameraType}{Len}) \times \\ (\frac{\sum Motion}{Len}) \times (\frac{\sum Heterogeneity}{Len}), \end{aligned} \quad (3)$$

Shot-level importance is proportional to the shot length (i.e. *Len*), the number of faces (i.e. Num_{face}), the presence of camera operations (i.e. $\Sigma CameraType$), the higher motions (i.e. $\Sigma Motion$), and the higher heterogeneity of a shot. (i.e. $\Sigma Heterogeneity$) Heterogeneity means whether a shot is dynamic. In the work, the total distance of color layout[11] between each frame is defined as heterogeneity

V. Audio and Video Synchronizations

Through observations, home videos can be divided into four categories according their properties: causal or non-causal, memorial or recreational. Causal means that all the shots are ordered in time and changes made in the sequence may cause confusion. Memorial means all shots are important, such as the video for marriages. Taking the properties into account, we propose 4 profiles to fit different applications: sequential rhythmic, sequential medium, non-sequential rhythmic, and non-sequential medium

In the rhythmic profile, the extracted video shots match the audio's tempo the most, but the resultant summary may discard some shots. On the other hand, the medium profile reserves most shots, so tempo matching is not the strongest criterion. The rhythmic and medium profiles are used to model recreational and memorial videos, respectively. The sequential parameter assures that the order of video shots will be preserved; On the contrary, the selection of non-sequential parameter will re-arrange the order of shots to make the best fit of video and audio. The two parameters model the causal and non-causal cases, respectively.

A. The Synchronization Mechanism

Figure 2 illustrates the adopted synchronization mechanism conceptually. At first, input video and audio are segmented by the techniques described in Sections Ⅲ and Ⅳ, respectively. Then for each audio segment, we select a video shot with the same length as that of audio segment. The selection criterion varies from profiles to profiles.

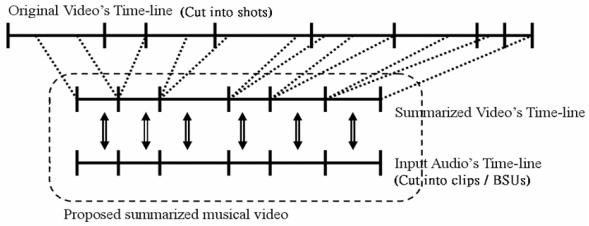

Figure 2. The synchronazation mechanism

The basic synchronization unit (BSU) is defined as the intervals between clip boundaries, between attacks, or between a clip boundary and an attack.

1) The Rhythmic profile

For each BSU, the starting and stopping points of BSU will be projected back to the video timeline (Figure 3). Base on the importance function, the system will search the projected range for finding a sequence with the same length as that of the BSU.

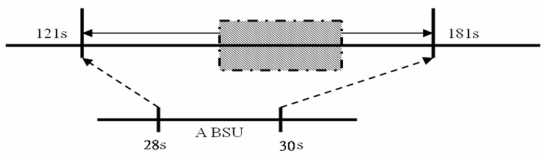

Figure 3. The synchronization mechanism of the rhythmic profile

2) The medium profile

The medium profile should make sure that every shot is not discarded. The synchronization mechanism is shown as in Figure 4. We introduce LBSU as a composition of 3 BSUs. The reason of using LBSU is to take multiple video shots into account. Then the LBSUs are also projected back to the video timeline. In this example, there are 4 shots in the projected range, and the LBSU is of 6 seconds. Our work is

to distribute the 6 seconds to the 4 shots according to the shot-level importance. Shots with higher importance value will be reserved longer.

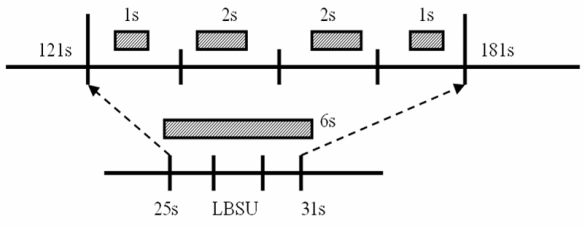

Figure 4. The synchronization mechanism of the medium profile

3) The sequential / non-sequential parameter

The sequential parameter can be applied to the rhythmic and medium profile directly. However, when it comes to the non-sequential case, shot permutation must be performed before applying to either profile. The goal of shot permutation is to align the distribution of shot length with that of audio clip length. In order to reach this goal, we employ a randomize algorithm to change the shot sequence and choose a sequence with a minimum average audio-to-video coverage (R_{avc}). Here, we define R_{avc} for each video shot as the number of projected audio clips covered by one shot (please refer to Figure 5).

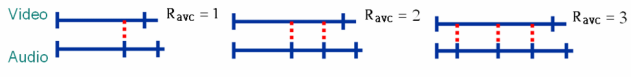

Figure 5. The audio-to-video coverage (R_{avc})

VI. EXPERIMENTAL RESULTS

In order to evaluate our system's performance, we invited 20 people to join the subjective test. Ten of them are with computer science background and others are not. Our subjective tests are divided into 3 parts. In part 1, testers are given 2 summarized videos, using rhythmic and medium profiles respectively. The question is about whether the difference between rhythmic and medium is obvious. In part 2, they are given another 2 summarized videos produced from the same original video but accompanied with 2 different kinds of music: slow tempo and fast tempo. The testers will be asked about their subjective feeling about the tempo matching.Table Ⅰ and Table Ⅱ show the testers' answer to part1 and part2, respectively. Finally, in part3, testers are asked about their degree of satisfication of the proposed system based on the generated video summarizations.

TABLE I. THE COMPARISON OF RHYTHMIC AND MEDIUM PROFILES

Questions	Answers of testers				
	5	*4*	*3*	*2*	*1*
The difference between Rhythmic and Medium profile is obvious.	2	10	8	0	0
It's necessary to adopt the Medium profile for some occasion	4	16	0	0	0

5: Very agree, 4:Agree, 3:no comment, 2:disagree, 1:very disagree

TABLE II. THE DEGREE OF AUDIO/VIDEO MATCHING

Questions	Answers of testers				
	5	*4*	*3*	*2*	*1*
The two edited video match the audio's tempo well.	2	10	8	0	0
The later video (the one with fast tempo music) has stronger motion intensity then the first one.	18	2	0	0	0

TABLE III. THE DEGREE OF SATISFICATION ABOUT THE MUSIC-DRIVEN SUMMARIZATION

Questions	Answers of testers				
	5	*4*	*3*	*2*	*1*
The summarized videos describe the original videos' sketch well. (In part 1)	11	9	0	0	0
The summarized videos describe the original videos' sketch well. (In part 2)	8	12	0	0	0
The system does help home users to manage their video works and is beneficial to video exchange.	10	10	0	0	0
The companion of music in videos does improve the production quality.	20	0	0	0	0

VII. CONCLUSIOIN

We have proposed and implemented a music-driven video summarization system that can help home users to post-process their creations in a fully automatic way. Many content-aware mechanisms are investigated to analyze the input media. Then the input video and audio are synchronized according to their content related features to form a musical videos. Four profiles are proposed to model various kinds of videos. The subject test shows high degree of satisfication of our system by most testees and most of them feel glad to have such a tool to help them editing their creations.

VIII. REFERENCES

[1] Tele Tan, Jiayi Chen, Philippe Mulhem, Mohan S. Kankanhalli, *"SmartAlbum: a multi-modal photo annotation system"*, ACM Multimedia 2002: 87-88

[2] J. C. Platt, *"AutoAlbum: Clustering digital photographs using probabilistic model merging"*, In Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries, pages 96--100. IEEE, 2000.

[3] http://www.adobe.com/products/premiere/main.html

[4] http://www.ulead.com/msp/runme.htm

[5] http://www.gocyberlink.com/english/products/product_main.jsp?ProdId=45

[6] AviSynth, http://www.avisynth.org/

[7] S. Pfeiffer, R. Lienhart, S. Fischer and W. Effelsberg, *"Abstracting digital movies automatically"*, Journal of Visual Communication and Image

[8] Ying Li, Tong Zhang, Daniel Tretter, *"An Overview of Video Abstraction Techniques,"* HP Laboratories Palo Alto

[9] Intel Open Source Computer Vision (OpenCV) library. Intel [Online]. Available: http://www.intel.com/research/mrl/research/opencv/

[10] Christophe Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," IEEE Trans. Multimedia, vol. 1, no. 3, pp. 246-277, Sept. 1999

[11] ISO/IEC 15938-3:2002(E), Information technology – multimedia content description interface – Part 3: Visual, May, 2002