Automatic Mash up Music Video Generation System by Perceptual Synchronization of Music and Video Features

Tatsunori Hirai* Hayato Ohya Shigeo Morishima

Waseda University / CREST

1 Introduction

In this paper, we present an automatic music video generation system by segmenting and concatenating existing video clips. Recently, video sharing web services have been in fashion especially in a genre of music video. In such a situation, music video creation by inexperienced people is increasing rapidly. However, it is difficult to create music video with no experience because editing music video need a lot of skill and expensive software. With this system, all things to create music video are to select favorite song and choose music videos which users want to use for new music video. To create music video automatically synchronized with any input music, we performed experiment which subjectively evaluates optimum synchronization conditions between motions in a video and the music.

2 Synchronization of Music and Video

According to a psychophysical experiment, *Sugano* et al. reported that when a tempo of music is corresponded to an accent of a video such as movement or flicker, people feel the music video has been synchronized well. To synchronize video with music more suited to a human perception, we proposed synchronization method between the accent of the dynamic image and time change of the music. To understand detailed change of the music, tempo is not sufficient factor. In our method, we used RMS energy of sound which corresponds to loudness of a sound. By using RMS energy, it is enable to extract the time change of the sound. As the accent of the dynamic image, we used an acceleration of each object in video frame and time change of luminous in full screen which correspond to movement and flicker in the dynamic image.

We compared our synchronization method with Sugano's method by subjective evaluation experiment. In order to judge which method is more suited to human perception, we performed subjective assessment to 22 people asking which music video is more suited to their perceptual feeling using each synchronization method. As a result, people tend to feel it is better synchronized when an accent of a dynamic image corresponds to not only a tempo but also RMS energy; time change of a sound.

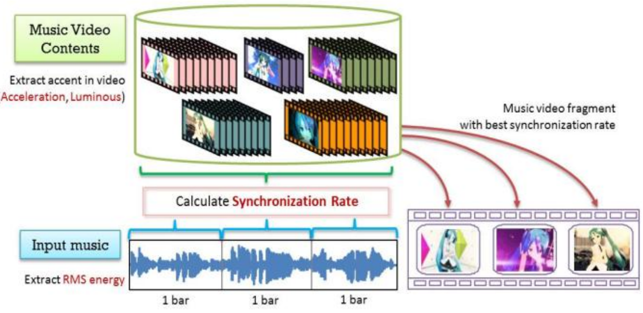
Music Video Contents
Extract accent in video (Acceleration, Luminous)
Music video fragment with best synchronization rate
Calculate Synchronization Rate
Input music
Extract RMS energy
1 bar
1 bar
1 bar

Figure 1: *Outline of the System.*

3 Database Construction

This system is composed of database construction part and music video creation part. In database construction part, object's accents in a music video are extracted. As information of movement, we used differential of optical flow in each moving area which corresponds to an acceleration of each moving object in a music video. At this point, we estimate the tempo of music which is added to the music video originally. A method to estimate tempo is to measure distance between peak and peak of a down-sampled waveform of the music.

4 Music Video Creation

In music video creation part, user gives a song and system calculates RMS energy of the sound in each bar. At the same time, tempo of the song is estimated with the method above.

For selection of the music video scene from database, we defined synchronization rate S as formula (1). When RMS energy of each bar in the song is $\boldsymbol{x}$, and acceleration or luminous in music video stored in database is $\boldsymbol{y}$, and a correlation coefficient of x and y is R, synchronization rate S is:

$$S = \frac{R + \left\{1 - \sqrt{(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}})^2}\right\}}{2} \qquad (1)$$

At this point, x and y are standardized with average value and variance. When value S is large and close to 1, transition and value between RMS energy and either acceleration or luminous in video frames from database music video are best synchronized. To make selection of music video fragment which most suit to each bar of the music, system calculates synchronization rate S in all $\boldsymbol{x}$ and $\boldsymbol{y}$ pair. To avoid often scene change in every 1 bar, we added weight to the synchronization rate of continuous video sequence. At this point, weight attenuates in time in order to prevent specific music video sequence to continue so long.

With this weight, scene change will not happen until a match with higher synchronization rate is found. When the video fragments are chosen, they get expanded or contracted in order to fit the length of 1 bar in the video to the length of 1 bar in the music. At last, system concatenates all chosen video fragments and adds music to the video sequence.

5 Results and Conclusion

With this system, we generated new music video with optional song by using fragments of existent music video clips. Generated music videos are based on a result of subjective evaluation experiment, that change of brightness and movement of an object is united with sound.

Reference

SUGANO, Y., IWAMIYA, S. 1999. The effect of structural relationships between motion of animation and rhythm of music on the audio-visual congruency and emotional impression. *Acoustical Society of Japan research paper*. 559-560

Copyright is held by the author / owner(s).
SIGGRAPH 2012, Los Angeles, California, August 5 – 9, 2012.
ISBN 978-1-4503-1435-0/12/0008

*e-mail: tatsunori_hirai@asagi.waseda.jp