Synthesis of a Video of Performers Appearing to Play User-Specified Band Music

Tomohiro Yamamoto* Makoto Okabe† Rikio Onai‡
The University of Electro-Communications *†‡ JST PRESTO†

1 Introduction

We propose a novel method to synthesize a video of multiple performers appearing to play the user-specified band music. The performance video often helps us to enjoy or understand the music. For example, watching the performance of a bass player makes it easier for us to focus on the bass sound, which is an experience different from listening to the music just using speakers or headphones. Our approach is based on the database of performance videos: given the band music, our system chooses appropriate footages from the database, slightly modifies their speeds and timings according to the music, and then concatenates them as the resulting video. In existing research, there are several methods proposed for synchronizing videos to a user-specified music, e.g., to synthesize a dance video [Nakano et al. 2011]. These methods synchronize a video to the music by analyzing its mood based on tempo or chord changes in the music. On the other hand, since we want to synthesize a performance video and require more precise synchronization, we analyze the music extracting the timings of musical notes from the audio signal (Fig.1). We perform best match search using the timings as feature vector, and copy the footage that has a similar set of timings (Fig.1-a and b). We demonstrate that our method enables to create a performance video of a band music, which is a fake but looks interesting.

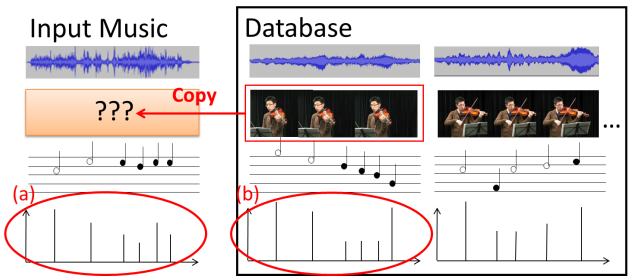

Figure 1: *The system overview*

2 Our Approach

We constructed a video database for each instrument independently. We asked each performer of viola, bass and drums to play the instrument for one hour. For each video, we analyze its audio track to estimate the timings of the musical notes (Fig.2-c). The process starts by applying the short-time Fourier transform (STFT) to the audio signal. The spectrogram is usually noisy (Fig.2-a). To smooth it preserving a strong edge corresponding to the beginning and end of each musical note, we apply a bilateral filter. We differentiate the smoothed spectrogram horizontally to extract the beginning and end of each musical note. Then, we integrate the spectrogram vertically to obtain one-dimensional (1D) signal (Fig.2-b). Finally, we extract the feature vector by finding local maxima of the 1D signal (Fig.2-c). We show the score that the performer actually plays (Fig.2-d). The peaks match the timing of the score.

The process described above is directly applied to the videos of viola and bass. However, in the video of drums, the audio track

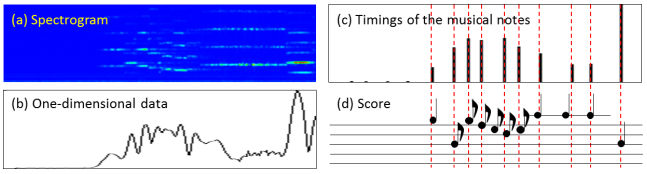

Figure 2: *Extracting the feature vector*

is mixture of four instruments, snare drum, bass drum, cymbals, and hi-hats. To analyze each sound, we apply audio source separation technique to the audio signal. We use probabilistic latent component analysis (pLCA) [Smaragdis et al. 2007]. Fig. 3 shows the result, where the original signal is separated into the four components. Actually, our drum kit has toms, but we distribute their sounds into the component of snare drum.

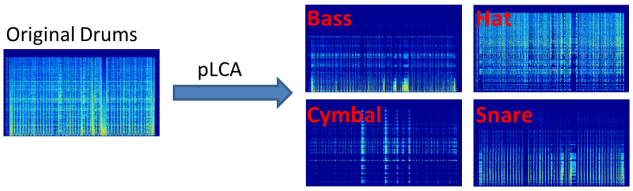

Figure 3: *Audio source separation*

The basic idea to synthesize the video is to compare the feature vectors between the input music and the video database and select a footage that has a feature vector similar to a part of the input music.We perform this process for each instrument. To increase the search rate, we allow slight differences between the feature vectors , which are solved in the rendering process by automatically modifying the speeds and timings of the footage using time remapping.

3 Results

We applied our method to two band music, 1) “Etupirka” of Taro Hakase, and 2) “Let It Be” of The Beatles. Given input music, we apply the same analysis as is used to construct the database and create the feature vector. Since Etupirka is originally a wave file, we applied pLCA to it to separate it into the violin part and the drum part. Let It Be is provided on the web with a violin solo part as a wave file, and the other parts as MIDI files. Given a 30 second audio track as input music, our method takes 30 to 40 seconds to search for footages for each instrument The selected footages are summarized with the information of time remapping as a script of Adobe After Effects. The user loads it into the software and renders the final video.

References

NAKANO, T., MUROFUSHI, S., GOTO, M., AND MORISHIMA, S. 2011. Dancereproducer: An automatic mashup music video generation system by reusing dance video clips on the web. In *Proc. of SMC*, 183–189.

SMARAGDIS, P., RAJ, B., AND SHASHANKA, M. 2007. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *Proc. of ICA*, 414–421.

*e-mail:yamamoto@onailab.com
†e-mail.m.o.acm.org
‡e-mail:onai@onailab.com

SIGGRAPH 2012, Los Angeles, California, August 5 – 9, 2012.
ISBN 978-1-4503-1435-0/12/0008