Synthesis of a Video of a Performer Appearing to Play User-specified Music

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

1 Introduction

We propose a method for using music to synthesize a video in which the performer appears to be playing user-specified music. Existing methods of synthesizing video from music employ a video summarizer [Foote et al. 2002] and the creation of a dance scene [Goto 2001]. While these methods synchronize video to music by analyzing the mood of the music based on the tempo or chord changes, our synthesis method uses a more detailed analysis of the music, i.e., video synchronization based on extracting the timing of musical notes. We create a feature vector so that its peaks represent the timing of musical notes (Fig. 1). We perform best match search using the feature vector, and copy the footage that has similar timing of peaks (Fig.1-a and b). In this poster, we demonstrate the application of our method to violin music and create a performance video that is fake but fun.

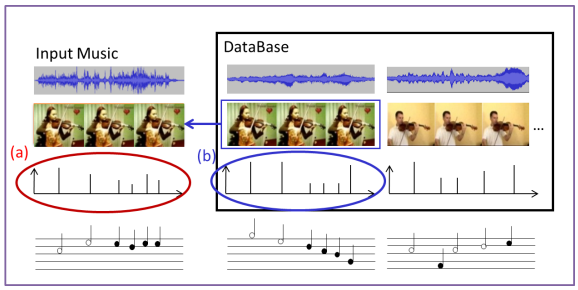

Figure 1: *Overview of the system.*

2 Method

We construct a video database in which each video contains footage and an audio track of a performance. Each audio track (Fig. 2-a) is analyzed in advance, creating a feature vector by extracting the timing of the musical notes (Fig. 2-e). The feature extraction starts by applying the short-time Fourier transform (STFT) to the signal of the audio track (Fig. 2-b). The spectrogram is usually noisy. To smooth it and to preserve a strong edge corresponding to the beginning and end of each musical note, we apply edge-preserving smoothing using a bilateral filter (Fig. 2-c) [Tomasi and Manduchi 1998]. We differentiate the spectrogram horizontally to extract the beginning and end of each musical note. Then, we integrate the spectrogram vertically to obtain one-dimensional (1D) signal, in which the peaks correspond to the timing of the musical notes (Fig. 2-d). Finally, we extract the feature vector by finding local maxima of the 1D signal (Fig. 2-e). We show the score that the performer actually plays (Fig. 2-f). The peaks match the timing of the score. We also include sound volume in the feature vector as it is important to capture the atmosphere of the music.

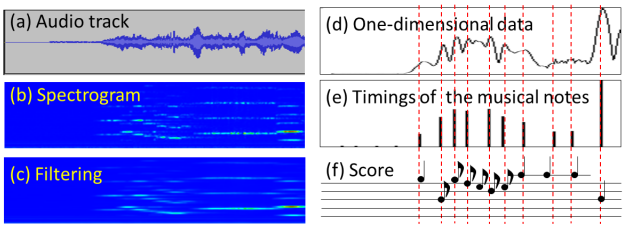

Figure 2: *Extracting the feature vector.*

Given input music, we divide it into fragments. We apply the analysis that was used to construct the database and create the feature vector of a fragment (Fig. 1-a). The feature vector is used to search for footage with a similar feature vector. Currently, the system searches for twenty candidate footages captured from various camera angles and allows the user to select the best one. We synchronize the footage to the fragment by modifying the speed of the footage locally so that the timing of the musical notes of both feature vectors aligns as much as possible. The footage is rendered at modified playing speeds by applying the motion interpolation technique.

3 Results and Discussion

Figure 3: *The frames of the rendered video.*

We applied our method to a violin solo from which it was possible to obtain relatively clean spectrograms by STFT. Our input music includes 1) “Salut d’amour” , 2) “Jounetsu Tairiku” at a faster tempo, and 3) “Etupirka” at a slower tempo. Watch the supplementary video. The movements of the right arm are synchronized to the music, or when the arm looks like it is not synchronized, the movements of the left hand are synchronized. As a result, the performer appears to be playing the music. Currently, with our non-optimized method, it takes about 3 hours to synthesize a 20-second video. It takes about 2 hours to search for appropriate footage for each fragment and about 1 hour for the final rendering. Note that our method is still faster than creating a synchronized video manually using video-editing software. We plan to speed the algorithm. To improve the quality of the resulting video, we plan to design a better feature vector and develop a user interface.

References

FOOTE, J., COOPER, M., AND GIRGENSOHN, A. 2002. Creating music videos using automatic media analysis. In *Proc. of ACM Multimedia*, 553–560.

GOTO, M. 2001. An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research 30*, 2, 159–171.

TOMASI, C., AND MANDUCHI, R. 1998. Bilateral filtering for gray and color images. In *Proc. of ICCV*, 839–846.

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