A music video authoring system synchronizing climax of video clips and music via rearrangement of musical bars

Haruki Sato[1, 2]   Tatsunori Hirai[1]   Tomoyasu Nakano[3]   Masataka Goto[3]   Shigeo Morishima[2, 4]
Waseda University[1]   JST CREST[2]   AIST[3]   Waseda Research Institute for Science and Engineering[4]
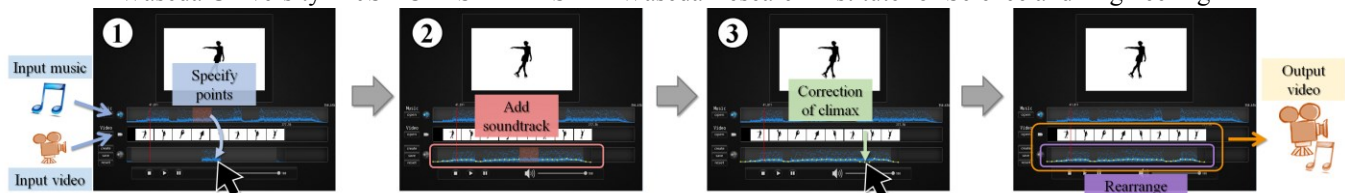
Figure 1: *Overview of the System*

1.   Introduction

This paper presents a system that can automatically add a soundtrack to a video clip by replacing and concatenating an existing song's musical bars considering a user's preference. Since a soundtrack makes a video clip attractive, adding a soundtrack to a clip is one of the most important processes in video editing. To make a video clip more attractive, an editor of the clip tends to add a soundtrack considering its timing and climax. For example, editors often add chorus sections to the climax of the clip by replacing and concatenating musical bars in an existing song. However, in the process, editors should take naturalness of rearranged soundtrack into account. Therefore, editors have to decide how to replace musical bars in a song considering its timing, climax, and naturalness of rearranged soundtrack simultaneously. In this case, editors are required to optimize the soundtrack by listening to the rearranged result as well as checking the naturalness and synchronization between the result and the video clip. However, this repetitious work is time-consuming. [Feng et al. 2010] proposed an automatic soundtrack addition method. However, since this method automatically adds soundtrack with data-driven approach, this method cannot consider timing and climax which a user prefers.

Our system takes all the patterns of rearranged musical bars into account and finds the most natural soundtrack considering a user's preference of intention for an audio-visual alignment and a climax of the resulting soundtrack. Specifically, musical sections between user's specified points and the beginning and the ending of the song are automatically interpolated by replacing and concatenating musical bars based on dynamic programming. To consider user's intention for a climax of the soundtrack, the system allows the user to specify the intended climax by an editing interface. The system immediately reflects the intention and the soundtrack will be interactively re-rearranged. These semi-automated processes of rearranging soundtrack for a video clip help the users adding songs without considering naturalness of the rearranged song by their own.

2.   Proposed System

Our proposed system consists of three parts. Figure 1 shows the overview of the system. The first part is specification of user's intention. User selects arbitrary musical sections of an input song and specified points in a video clip. Then, the first musical bar of an input song and the last musical bar are synchronized with the beginning of input video clip and the ending automatically. The second part is interpolation of a soundtrack. We generate a soundtrack by concatenating musical bars based on dynamic programming. To concatenate similar musical bars, we define a cost C between two musical bars as formula (1).

$$C = \alpha \sqrt{M + \Delta M + M_E + \bar{M} + M_\sigma} \qquad (1)$$

$M, \Delta M$ and M_E denote Euclidean distance of the MFCC, ΔMFCC and RMS audio feature between two musical bars. $\bar{M}, M_\sigma$ denote Euclidean distance of the mean and standard deviation value of the MFCC between two musical bars. α denotes a weighting factor for prompting to concatenate original musical bars. The third part is an interactive correction of climax of the resulting soundtrack. We rearrange musical bars of a soundtrack based on a new cost C' which is expressed as formula (2).

$$C' = C + |S - S_U| \qquad (2)$$

S denotes RMS mean of each musical bar, and S_U denotes climax that a user specified. If S_U becomes larger, musical bars that have large climax tend to be relocated to the point.

Table 1: *Median rank of the evaluation.*

Genre	Median	Genre	Median
World	2	Vocal	7
Jazz	3	Traditional Japanese	7
Latin	3	Dance	8
Popular	4	Classical	9
Rock	4	March	9

3.   Evaluation

We performed a subjective evaluation experiment to examine what kind of music genre does our method applicable to in terms of naturalness. 5 examinees are asked to evaluate 10 songs generated by our method based on naturalness. For this experiment, we used a song from each genre in RWC Music Database (Music Genre) [Goto 2004] and AIST Annotation [Goto 2006].

Table 1 shows the result of this experiment. Songs with small median rank tend to have a lot of similar melody in a song. Therefore, musical bar concatenations tend to be smooth. On the other hand, Songs with large median rank tend to have diversity of melody and tempo within a song. Therefore, musical bar concatenations do not tend to be smooth. We still have not considered the vocal section in current implementation. Therefore, if the song includes singing voices, they may appear or disappear suddenly. From these results, our method is effective when the input is an instrumental song which frequently has similar melody sections.

4.   Conclusion

This paper presented a music video authoring system. By this system, we can generate a new soundtrack considering a user's preference for an audio-visual alignment and a climax of the soundtrack. As the future works, we have to consider tempo of a song and existence of vocal section to apply our system to songs which have many tempo variations and vocal sections.

References

FENG, J., NI, B., AND YAN, S. 2010. Auto-generation of Professional background music for home-made videos. *ICIMCS*, ACM, 15–18.
GOTO, M., 2004. Development of the RWC Music Database. *ICA*. 553–556.
GOTO, M., 2006. AIST Annotation for the RWC Music Database. *ISMIR*. 359–360.