AUTOMATIC MUSIC VIDEO CREATION SYSTEM
BY REUSING EXISTING CONTENTS IN VIDEO-SHARING SERVICE BASED ON HMM

Hayato Ohya, Shigeo Morishima
Waseda University / JST CREST

ABSTRACT
We propose automatic music video creating system by reusing music videos that are created by amateur users and existed on the video sharing service. These music videos are created by combining existing music with frames of video games, anime, and so on. They are called “MAD” video. In this paper, we improved learning method of DanceReProducer that is the system which our group proposed previously. We enabled the system to consider the time series information of videos by using Markov chain, and to generate movie by using Forward Viterbi algorithm. And we compared our system with DanceReProducer to carrying out subjective assessment experiment.

1. INTRODUCTION
User-generated music video clip called MAD movie, which is a derivative (mixture or combination) of some original video clips, are gaining popularity on the web and a lot of them have been uploaded and are available on video hosting web services. Such a MAD music video clip consists of audio signals and video frames taken from other original video clips. In a MAD video clip, good music-to-image synchronization with respect to rhythm, impression, and context is important. Although it is easy to enjoy watching MAD videos, it is not easy to generate them. It is because a creator needs high-performance video editing software and spends a lot of time for editing video. Additionally, a creator is required video editing skill.

DanceReProducer[1] is a dance video authoring system that can automatically generate dance video appropriate to music by reusing existing dance video sequences. It trains correspondence relationship between music and video. However, DanceReProducer cannot train video sequence information because it only trains one-bar correspondence relationship.

So we improved DanceReProducer to consider video sequence information by using Markov chain of latent variable and Forward Viterbi algorithm.

2. RELATED WORKS
DanceReProducer gets existed video contents and learns the synchronization relationship between music and video. Then, for input music, DanceReProducer selects video scenes of learned video contents based on the learned relationship between music and video, and automatically generates MAD video to combine video scenes with music. When learning the relationship between music and video, MAD videos which are created by amateurs are used as training data. Music and video features are extracted by a short time and gathered each music bar. In this case, DanceReProducer cannot learn video time sequence information because learning music and video features is one-on-one training.

Jiang et al. [4] automatically generate video summarization by learning time sequence information. They generate video summarization from original movie and its editing video which is summarized by professional. The shots of which editing video consists, the scene that has been continuous shooting without a break, are symbolized based on shot length, movement, luminance, etc. and its time sequence pattern is learned by Hidden Markov Model (HMM). But, this research doesn’t treat the sound of movie.

So in our work, we propose the method that uses Markov model for learning whose latent variable is video feature and observed variable is music feature, and automatically generates MAD video by Forward Viterbi algorithm.

3. SYSTEM OVERVIEW
Figure 1 shows our system’s flow chart. Our system consists of three parts. The first is database construction part. The second is model training part. The last one is video generation part.

In the database construction part, music feature and video feature are extracted from existing video contents. Then each feature is collected each music bar.

In the model training part, each feature is clustered for HMM and HMM parameters are trained by Markov chain.

In the video generation part, music feature and tempo are extracted from music we want to attach video on (input music) and feature is collected by bar-level feature. Then decide video sequence by Forward Viterbi
algorithm and using HMM parameter calculated from previous part. Considering state of previous bar video feature in addition to current bar music feature in estimating one-bar feature of video enables video generation training video sequence.

**Figure 1. System flow**

4. DATABASE CONSTRUCTION

First of all, we state database construction. Our system is given video contents of which a database is made. As preprocess, frame rate of each video is standardized 30 fps and sampling rate of music is standardized 44.1 kHz. Then, the system extracts music and video features each video frame. Also, the system collects obtained music and video features each music bar by estimating music bar of music of MAD videos.

4.1. Extraction of music features

We decided music features related to accent and impression by reference to Nishiyama et al [5] and the research of music genre classification [6]. Table 1 shows music features in our work. As the features related to accent, we use Spectral Flux (1 dimension) and Power by sub band (4 dimensions). In our work, as same as DanceReProducer, the number of filter bank is four. As the features related to the impression, we use Zero-crossing rate that is related to the tone of music (1 dimension), and the direct-current component and 12 lower term of MFCC (Mel-Frequency Cepstral Coefficients) (13 dimensions). The total of 19 dimensional features is extracted.

The shift width of analysis window is 1470 points (= 44100Hz / 30 fps) to take the corresponding to the video frame features, and window length is 2048 points.

4.2. Extraction of video features

We decided video features related to accent and impression by reference to Nishiyama et al [5]. Table 1 shows video features in our work. As the features related to accent, we use Optical flow and Differential of histogram of brightness value to indicate the movement of screen, its time change and the switching of screen (each 1 dimension). Optical flow is calculated by block-matching algorithm that parameters go as follow. The number of blocks is $64 \times 48$, the shift width is 1, and the max of shift width is 4. As the features related to the impression, we use the mean and the standard derivation of the hue, the saturation and the luminance of all pixels in the frame, to indicate the atmosphere of video (total of 6 dimensions). We also use the 12 dimensional coefficient of 2 dimensional DCT (Discrete Cosine Transform), to indicate the impression of the atmosphere of entire image in the frame.

As preprocess, we change frame size of video to $128 \times 96$, and the shift width of analysis window is 1 frame ($1 / 30$ s) that is the same as the frame rate of the video.

4.3. Music bar detection

To detect music bar, we calculate the method based on the correlation function of the power of acoustic signal. First, we calculate the peak time of the autocorrelation function of the acoustic power of input music. This shows the periodicity of the power, so we detect the tempo as the length of one beat. However, to avoid detecting double time tempo and half time tempo, we limit the tempo from 90 bpm to 180 bpm. Figure 2 shows the example of the tempo detection.

Then, the pulse string that has the peak each beat from detected tempo is generated and the peak time is detected by calculating the cross-correlation function that is the correlation between this pulse string and the power of input acoustic signal. We decide this peak time as the first beat of music. From this result, we consider first beat as the start point of music bar, and decide the position of music bar to assume 4/4 beat.

4.4. Gathering features each music bar

We use DCT as the method of gathering each feature into a unit of music bar, to consider time sequence information. We resample frame features at 16 points, perform DCT to this 16 points feature, and use the lower 4 dimensional feature that is transformed by inverse DCT. Figure 3 shows the result of inverse DCT in 4 dimensions. From this figure, we can confirm the increase of feature in the second half. We can consider that the loss of time sequence information caused by resampling is small because the loss is about 30 samples.
at 60 bpm. From this result, music bar feature is 76 dimensions and video bar feature is 80 dimensions.

Figure 2. the detection of the tempo

![Figure 2. the detection of the tempo](image)

Figure 3. Zero-crossing rate and the result of inverse DCT in 4 dimensions

![Figure 3. Zero-crossing rate and the result of inverse DCT in 4 dimensions](image)

5. LEARNING SYNCHRONIZATION RELATIONSHIP BETWEEN MUSIC AND VIDEO

The synchronization relationship is learned by HMM, to use calculated music bar feature and video bar feature. First, as preprocess, each feature is analyzed by principal component analysis and decreased dimension (cumulative contribution ratio is 95%). At the result of this process, music bar feature decrease from 76 dimensions to 62 dimensions. And video bar features also decrease from 80 dimensions to 64 dimensions.

5.1. Clustering of music and video feature

For automatic music video generation based on HMM, music and video features are clustered by k-means clustering algorithm. Music feature cluster is observed value and video feature cluster is state of latent value. Here, the number of clusters is ten.

We can consider that, as the number of clusters increases, the more proper frames are selected. However, underflow would cause too much increase the number of clusters, when the proper time sequence information of clusters is calculated by Forward Viterbi algorithm. In this work, the video time is from about 3 minutes to 5 minutes, and the number of cluster sequence (as same as the number of music bar) is from about 90 to 150.

After clustering, application of linear regression techniques to clusters allows for the calculation of the regression coefficient $A$; explanatory variable is music feature and objective value is video feature. The regression coefficient $A$ is used when the system automatically create music video.

5.2. Clustering of music and video feature

Figure 4 shows the training model of the synchronization relationship between music and video by Markov model. In the training model of Markov chain, we apply music feature cluster to observed data $O$ and video feature cluster to the state of latent value $x$. Then, we calculate following values in the Markov chain; State transition probability: $T$, Emission probability: $P(O|x)$, Initial state probability: $S_0$.

5.3. Weighted learning by views of video

The training data of our study is the videos created by amateur users in the video sharing website, so there is great variability among their qualities. This is why it is improper to equate them as the training sample. So we consider the views of video as statistical evaluation metrics and perform weighted learning.

Here, weight $\omega$ is denoted by formula (1).

$$\omega = \alpha \cdot \log(s) + \beta$$

(1)

The parameter $s$ is views. In formula (1), we configured $\alpha = 1.0$ and $\beta = -2.0$, as the weight factor is twice when views $s$ is 10000, and as the weight factor is triple when views $s$ is 100000. Videos whose views are less than 1000 are not used for training.
6. AUTOMATIC MUSIC VIDEO GENERATION

We state automatic music generation part of the system in this section. First, as preprocess, music feature is extracted from input music that is a music piece of new creation music video. Tempo, musical structure and music bar of input music are also detected as same as previous section.

Next, music feature is clustered each bar as well as previous section, too. Then, video feature cluster is detected from music feature cluster by Forward Viterbi algorithm. Figure 5 shows this conceptual map.

At last, output video is automatically generated from video feature cluster, music structure of input music and music bar feature of input music.

6.1. Detection of video feature cluster

We use Forward Viterbi algorithm to detect video feature cluster. In this detection, hidden state sequence \( x_0, x_1, \ldots, x_N \) is denoted by following recurrence formula (2), (3).

\[
V_{0,k} = P(O_0|k) \cdot S_p
\]

\[
V_{n,k} = P(O_n|k) \cdot \max_{\pi \in \mathcal{X}} (T_{\pi,k} \cdot V_{n-1,\pi})
\]

Here, each parameter is follow; \( V_{n,k} \): Maximum-likelihood state sequence, State transition probability; \( T \), Emission probability; \( P(O|x) \), Initial state probability; \( S_p \).

6.2. Video generation by video feature cluster

The system decides scene structure by \( V_{n,k} \) and musical structure information of the input music. Scene change is decided under the following conditions.

- Clusters of \( x_m \) and \( x_{m'} \) are different
- Musical structure changes between \( x_m \) and \( x_{m'} \)

After deciding scene structure, video feature of the input music is calculated from regression coefficient \( A \) and music feature. At last, the system combines the input music with the video frames of database videos whose video feature is nearest from the calculated video feature in Mahalanobis’ generalized distance.

When to generate video, to prevent the generated video from being unnaturally fast or slow, the video in the database whose tempo of music is different by more than 20 percent from tempo of input music is excepted. And the video frame that is once used is not used in the same generation.

5. CREATION RESULT AND DISCUSSION

At the result of automatic music video creation, the music video whose music and video are synchronized in terms of impression is created. In the case that input music is dance music, the system pasted the video frames whose scene has glittering neon and lights into input music. In the case that input music is popular music, the system pasted the video frames whose scene impression is different with each change in music structure. Especially, when input music gets moving, video frames that give the impression of dashy are pasted. In the case of jazz music that is instrumental, both light frames and dark frames are pasted. Light frames also fit the impression of jazz music because the saturation is not very high.

We also created music video by DanceReProducer using the same music above. In the case of dance music that doesn’t change melody frequently, the frames
whose saturation is not high. In the case of popular music, pasted video frames are comparatively light. However, created video is not so sharp by the musical structure. In the case of jazz music, pasted frames are dark frames. But, pasted frames are a piece of the same video. So, we consider that training the time sequence information enables paste video frames that matches the impression of music. We have uploaded created videos in this URL(http://youtu.be/Bg2TqVltpkk).

6. SUBJECTIVE ASSESSMENT EXPERIMENT

6.1. Contents of experimental

We automatically created 4 music videos using 4 input music by our system. And we also created 4 music videos using the same music by DanceReProducer. Moreover, we got 4 MAD videos whose music is the same above from video-sharing website. Then we compare three kinds of music videos.

Music videos are all dance videos because of comparing DanceReProducer. Training videos are 90 existing dance videos. The number of view of MAD videos is over 50,000. This indicates that MAD videos are high quality.

Examinees are 15 men. Examinees watched music videos and evaluated synchronization between music and video. Evaluation data is four level evaluation in each video frames; Very synchronize, Synchronize, A little synchronize, Not synchronize.

6.2. Experimental result

Figure 5 shows experimental result. “User-created” shows the evaluation of MAD videos got from video-sharing website, “DanceReProducer” shows the evaluation of the music videos that are automatically created by DanceReProducer, and “Our method” shows the evaluation of the music videos that are automatically created by our method. At the result of experiment, we can recognize that the percentage of “very synchronize” of “User-created” is very high, comparing with “DanceReProducer” and “Our method”. This is because detection miss of bar misaligns music and dance. When many focus on “DanceReProducer” and “Our method”, we can recognize that the percentage of sum of “very synchronize” and “synchronize” in “Our method” is higher than “DanceReProducer”. It is considered that training time sequence reflects that the impression of music matches the video’s one.

7. CONCLUSION

In this paper, we enabled automatic music video creation system to consider time sequence of video frames, by using Markov chain training model and forward Viterbi algorithm. In the result of subjective assessment experiment, our system can create music video that has more synchronization between music and video than DanceReProducer’s one. But, we have not been able to show quantitative data for considering time sequence. In the future, we need to examine new assessment method. In addition, we also need to consider the definition of “synchronization between music and video” and objective evaluation method.

In addition, GUI that is implemented in DanceReProducer is not implemented in our work. To make GUI that users can easily edit the generated video scene, the system could be improved to automatically generate high-quality music video at one time, to learn its editing data. We think that its obtained learning data is not “optimum solution” which music feature and video feature is physically conformed, but “satisfied solution” which human is qualitatively satisfied with the synchronization relationship between music and video. From the view point of automatic video generation, we also think that the optimal parameter should not be adjusted through the entire video, but the video which human feel good synchronization between music and video could be automatically generated to join the parts of the video which human feel good. So we will study GUI which can find “satisfied solution”, and the improvement of the automatic music video generation system.

11. REFERENCES


