Automatic Music Video Generation: Cross Matching of Music and Image

Xixuan Wu[1,2], Bing Xu[2]*, Yu Qiao[1], and Xiaoou Tang[1,2]
[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[2]Department of Information Engineering, The Chinese University of Hong Kong
{wxx010,xtang}@ie.cuhk.edu.hk,jethro.missing@gmail.com,yu.qiao@siat.ac.cn

ABSTRACT

Music and image are two most popular media on the Internet. Human perception of music and image are highly correlated. Music video is one of such products, in which music and image are complement to each other. In this paper, we present a system which can automatically generate music video for a given song. The challenge of such system comes from how to select relative images and align them with the song. This paper deals with this challenge by leveraging lyrics (if exists) and the semantic similarity between music and image. We retrieve related image in internet with lyrics keyword as query and use a learning based method to estimate a semantic score between an image and a music segment. Finally we construct a music video after quality filtering and refinement. Our system also allows users to upload their images and re-pick recommended images to personalize the music video.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Music,Image,Human Factors

Keywords

music video generation, music-image similarity estimation, multiple ranking Canonical Correlation Analysis

1. INTRODUCTION

Music and image are two popular forms of media. Although music and image are different types of human perception, image in vision and music in auditory, human perception of music and image show strong connections and correlations with each other. In our recent work [2], we tried to bridge music and image through computer processing. We developed an algorithm that can automatically estimate the matching degree between an image and a segment of music. In this work, we try to utilize the finding and results in [2] on a challenging application: automatically generating a music video for a given song.

*Xixuan Wu and Bing Xu contributed equally to this paper.

MM'12, October 29–November 2, 2012, Nara, Japan.
ACM 978-1-4503-1089-5/12/10.

To manually produce a music video for a song, a professional producer generates the visual content based on two sources of information, the lyric of the song and the music. The visual content has to capture either the semantic meaning of the lyric or the emotional inspiration of the music or both. In a way, we can consider lyric as a form of labeling of both music and visual information. Then, due to a common labeling, we can see that there is a semantic connection between music and image. This is demonstrated in our recent study of the relationship between music and image [2].

Apparently, it is extremely challenging for a computer to simulate a professional producer to generate a music video automatically based on an input song. Especially, if we directly deal with both music segments and video segments with different lengthes, then it is very difficult to simultaneously balance semantic matching and time warping. To simplify the problem, we use images instead of video segments to produce the visual content of the music video. Instead of pure static display of image sequences, we use some simple animation techniques to simulate a sense of motion in the final production. With a proper video database, it is also possible to replace some images with segments of videos.

To produce the visual content, we first have to decide where to find them. To produce interesting and dynamic visual content, we need to have a large pool of source images. In this work, we propose three sources of image data for the visual production. First, we know that the largest pool of images is the internet. We propose to use the lyrics of the song as search keywords to search the internet for relevant images and then use image quality filtering [1] to select high quality relevant images as one source of potential image candidates. This part of the data takes advantage of the relationship between images and lyrics. Secondly, we construct a large relatively stable image database with data collected from the internet. Then algorithms developed in [2] can be used to find the best few matching images from the database for each input music segment. This part of data utilizes the relationship between music and images. Finally, we allow users to use their own family photo album as another source of data, which helps users to personalize the final video production. The final video is generated by a weighted combination of the data selected from all three sources of data. We also allow users to adjust the final results through an interactive interface.

2. FRAMEWORK AND ALGORITHMS

Our system can generate music video automatically according to lyrics and musical content of an input song. The overall system is shown in Figure 1.

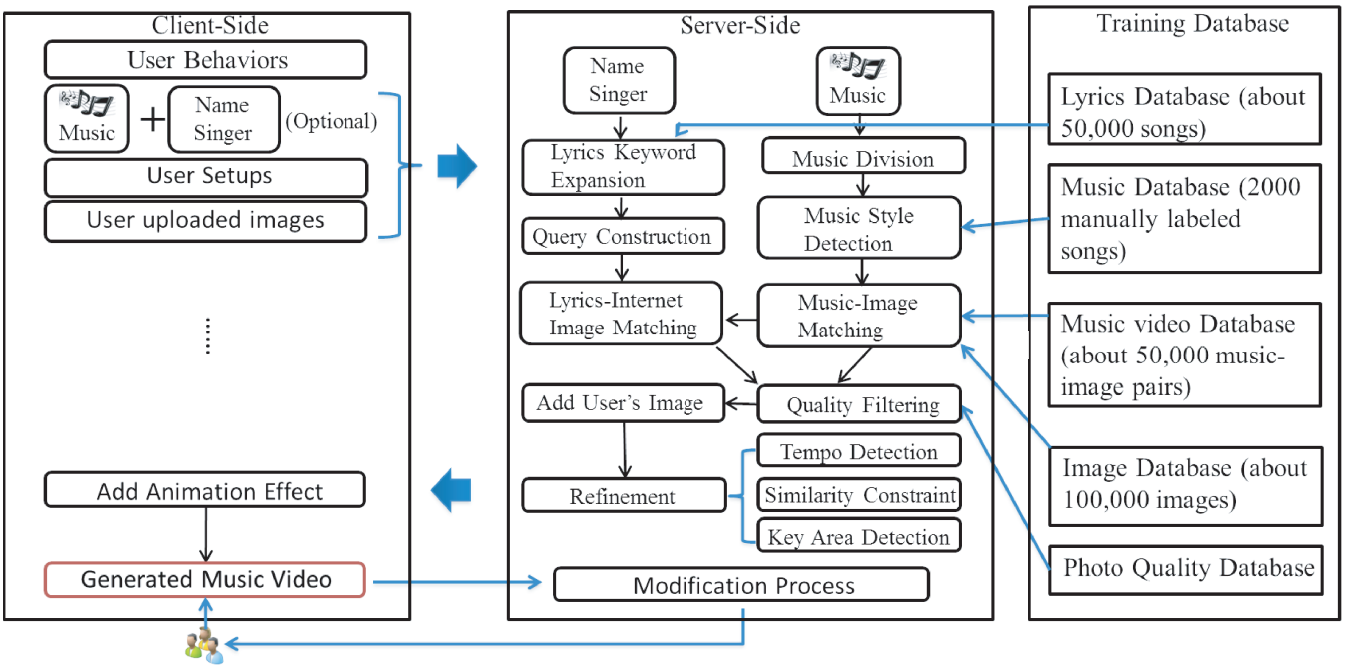

Figure 1: Architecture Overview of Our Music Video Generation System.

Users first upload a piece of music to the system. The music can be a song with lyrics or pure music. Users can also input the name and singer of the uploaded music. Users are encouraged to add their personal photos to personalize the produced video. After users finish making settings, all the information are sent to our server-side.

At the server-side, the processing is divided into two aspects: lyrics and musical content. For the lyrics part, we use natural language processing techniques to determine the key words of each lyric sentence. We construct search queries by key words expansion to look for relevant images from the internet. In order to cover different possible cases, we divide the set of images into four clusters based on histogram and spatiogram similarity. Given the song name and the singer name, our system automatically determines whether we have the lyrics available in our lyrics database (which contains lyrics of more than 50,000 songs).

For musical content, we first divide the music into several music segments with dynamic texture model or bar information. Given a sequence of music segments, we use the multiple ranking canonical correlation analysis (MR-CCA) [2] to learn a similarity function between music segment and image. In the generation phase, each segment is compared with all the images and get the 50 most similar images as good candidate images. We also use this approach to re-rank images obtained by using lyrics in order to get highly correlated image both with music content and lyrics. We construct an image dataset of 100,000 images collected from image search engine such as Google Image and Flickr with diverse styles for users to choose.

After the above steps, we have two kinds of candidate image resources: images from the Internet which are based on lyrics, and images from our image dataset which are based on musical content matching. We compute the image quality of candidate images for each track segment, and get the highest one as the final appearing image [1].

We also have a refinement process. Refinement contains three steps: tempo detection, similarity constraint, zoom detection. Tempo detection assigns images to music based on its tempo. If the music is soft, image are assigned according to its lyrics (if exists) or bar information. If the music is fast, image appears with the speed of its rhythm. For each image from users, the system replaces the most similar image in the generated image sequence with users' personal image. Similarity constraint ensures the similarity between successive images is in a reasonable range, which makes the video smoother. The system also adds several animation effects to the generated image sequence, including fade in, fade out, panning, and zooming. System utilizes face detection and subject area detection to determine the area of zooming. During the modification process, users are offered opportunities to modify the generated video. The system will choose 1 or 2 images from image clusters generated by lyrics, 2 images from musical content and another 1 personal image (if exist) for user to replace an image in the video that the user does not like.

The system demo and music videos generated by our system are shown at `http://mmlab.ie.cuhk.edu.hk`.

3. CONCLUSION

In this paper, we develop a system to automatically generate music video for a song. The key challenge in automatic generation of music video is how to bridge music and visual content. Our system deals with this challenge by combining several techniques, lyrics based image search, music-image matching, and photo quality estimation. Our system allows users to upload images to generate personalized music video. The proposed techniques also have applications in collaborative content analysis of music video.

4. REFERENCES

[1] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *ICCV*, 2011.

[2] X. Wu, Y. Qiao, X. Wang, and X. Tang. Cross matching of music and image. In *ACMMM*, 2012.