Lyrics-based Automatic Music Image Generation Using Scene Knowledge for Music Browsing

Go Kikuchi and Hiroyuki Kasai, *Member*, IEEE
The University of Electro-Communications, Japan

***Abstract* —Visual information of music makes users to understand it rapidly, and is very useful for a fast and efficient music browsing. In this paper, we propose an automatic music image generation that extracts features from lyrics such as an event and a place using the probabilistic parametric mixture model and the scene composition knowledge[1].**

I. Introduction

Recently, with the advance of online music distribution services and the high-capacity portable music players, the opportunities to browse and select favorite music from a large number of music has increased. Existing melodic recommendation mechanisms using not only meta-information such as title and genre but also acoustic features make music selection more efficient. Meanwhile, visual-information can represent music by pictures, animations or graphics, and make users to understand it rapidly, and it is very useful for a fast browsing. Therefore, some services may also include jacket pictures, and many researches have been done for music visualization. In [1] and [2], based on audio, they generate image pictures (like CG arts) that are consistent with human's auditory stimulus. In [3], based on lyrics, they get the image pictures from image search service, and select the most suitable image picture for the music. It is, however, in common in them to see that they use directly obtained image pictures and cannot provide an appropriate image to each number. In addition, lyrics have in general no complete sentences and have less number of words in it. Therefore, it may be hard to directly extract such information.

In this paper, we propose a technique that automatically generates a music image that matches the content of lyrics. We extract features such as event and place using a probabilistic approach of words in lyrics. Then, we generate a music image by synthesizing source images using our constructed knowledge of scene composition.

II. Lyrics-based Automatic Music Image Generation

We extract features place and event information from the lyrics of music, and generate the music image that matches the content of lyrics by synthesizing source images. Fig.1 shows the basic steps in the proposal. Place and event information are firstly extracted using keyword matching and a topic-based probabilistic approach. Then, some representative scene compositions for the extracted information are derived from the knowledge of scene composition. This knowledge has multiple representative scene compositions that are consist of a set of objects. These compositions have own scores that indicate appropriateness for the place and event information. In addition, the occurrence probability of each object in each composition is described.

Next, an appropriate scene composition is selected based on extracted objects in the lyric or randomly select if no object is detected. Finally, we generate a music image by selecting source images corresponding with the scene composition and synthesizing them.

[1] This work is supported in part by the National Institute of information,and Communications Technology (NICT), Japan, under,Grants for Research and Development on Network Virtualization,Infrastructure for New-generation network.

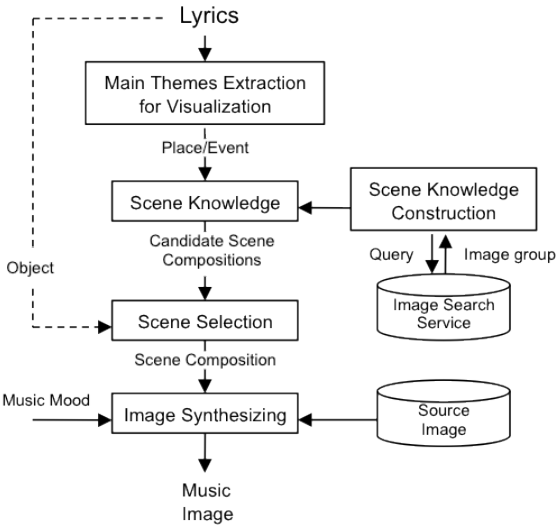

Fig.1. Music Image Generation using Scene Knowledge

A. Place and Event Information Extraction from Lyrics

Firstly, we have to extract some main themes for visualization because they can be helpful clues for visualization. To do this, we focus on extracting place and event information from lyrics as the main themes for visualization in this paper. However, because lyrics have in general no complete and full sentences and have less number of words in it compared with a general document, it may be hard to directly extract such information. That is why we adopt two tools; namely keyword matching and the Parametric Mixture Model (PMM) [4]. Keyword matching directly detects places and events shown in Table I from lyrics. Then, if no information is detected, the Parametric Mixture Model (PMM) is used. The PMM is one of text classification methods using Bag-of-Words (BOW). Words of multi-topic texts are a mixture of words of single-topic texts. Based on this idea, The PMM automatically constructs the probability model by learning from classified texts with multi-topic. The PMM was reported as that multi-topic classification performance is high than other representative classification methods such as Naive Bayes, Support Vector Machine, Nearest Neighbor and Neural Networks. Using the PMM, multiple events or places can be extracted even if lyrics have exact words for place and event. These multiple results give the system a wider variety of understandable music images.

TABLE I Target Place and Event Information Example.

Place	Church, Forest, Home, Mountain, School, Sea, Town
Event	Christmas, Drive, Graduation, Party, Sports, Valentine

B. Automatic Scene Knowledge Construction

We obtain a representative object composition according to the extracted places and events using our scene composition knowledge using object recognition on images [5]. Here, its outline is described briefly. First, we obtain image group with a targeted scene tag. To collect such images, we use an image search service by giving the extracted places and events as query words. Second, we recognize objects (shown in Table II) that exist in each image. In this process,

we use a recognition technique of [6] that recognizes a object at each pixel. First, a image got divided into many small regions(super-pixel), and score of each object are calculated using color and texture features in each region. Then, a object of image is recognized by applying these score to Conditional Random Field(CRF). Finally, we extract representative scene compositions based on inclusive relations and appearance frequencies of the extracted object compositions. Some of the representative object compositions are included in many images, and others are not. Then, a metric; named Scene Score to indicate how often each composition appears in all of images is defined. In addition, Object Score is defined to represent a frequency of each object. We calculate these scores of existence strength using its frequency and inverse frequency.

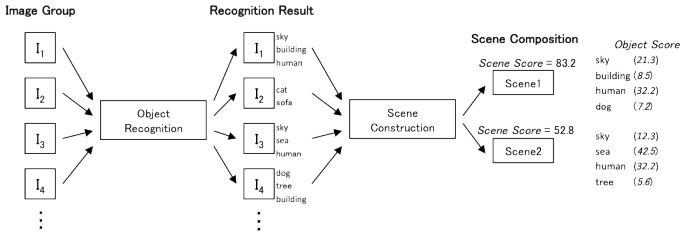

Fig.2. Scene Knowledge Construction

C. *Image Synthesizing based on Scene Composition*

The final step creates a music image based on the scene composition knowledge. This knowledge gives objects to be drawn, and multiple appropriate source images to the objects are combined creating the music image. The detailed steps are as follows;

1) Selection of Scene Composition (step1)

The most appropriate object composition will be selected from the candidate representative scene compositions. Specifically, the objects are extracted from the lyric based on the same keyword matching technique. Then, the scene composition with the highest total Object Score is selected. This total value is calculated by summing up all of the corresponding Object Scores of the detected objects. If no scene composition has the detected objects, the scene composition with the highest Scene Score is selected.

2) Image Synthesizing (step2)

In this last step, prepared source images depicted in Fig.3 are synthesized based on the selected scene composition. Each source image has its locatable region inside an image, and maximum and minimum size. The locatable region avoids unrealistic image where, for example, the sea locates above the sun. According to this, multiple object images are located by avoiding their overlaps to each other. The Object Score controls the size of each object. This means that the objects with higher score are drawn in a bigger size. Finally, we will also draw human characters whose face expression follows extracted emotion. Then, an entire color of the image is adjusted to mach the atmosphere of the image to the extracted emotion and music mood. These emotion and music mood can be extracted based on keyword matching against lyrics [7], or might be analyzed based on acoustic feature like Mel-Frequency Cepstrum Coefficients (MFCC).

Fig.3. Example of source image.

III. Case Study Example

We example our proposal using a case study example by assuming to process a lyric as follows; The place and event information in Table II represents the results from keyword matching and the PMM method for the main theme extraction for visualization process. Here, because the keyword matching successfully detected some words, they are used in the following processes. The Objects in the third line are the results from the keyword matching to the lyric. The scene composition in the fourth line is derived from the object-based selection process. Fig.4 shows the created music image. In this way, we can confirm that it is possible to create a music image from lyrics based on the multiple feature extraction and scene composition knowledge.

TABLE II Extracted Features from a Lyric.

Place/Event (Keyword Matching)	Drive, Sea
Place/Event (PMM)	Sea, Sports
Object	Bird, Road, Human, Tree
Scene Composition	Bird (3.2), Human (15.1), Sea (23.3), Sky (51.2), Tree (8.1)

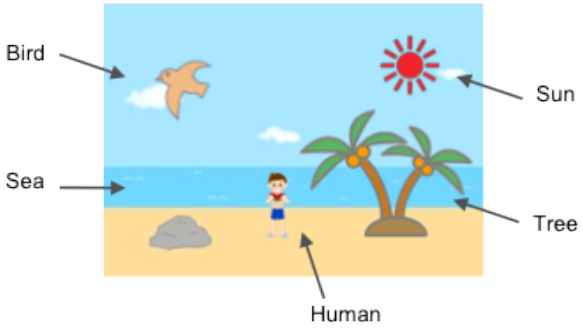

Fig.4. Example of Source Image

IV. Conclusion and Future Work

In this paper, we proposed a technique that automatically generates a music image that matches the content of lyrics. We extract event and place information using a probabilistic approach of words in lyrics as main themes for visualization. Then, we generate a music image by synthesizing source images using our constructed knowledge of scene composition. Future works include the extension of the PMM. MFCC-based music mood analysis scheme will be also integrated into this system.

References

[1] Reiko Miyazaki and Kouichi Matsuda, "DynamicIcon: A Visualizing Technique for Musical Pieces in Moving Icons Based on Acoustic Features," Journal of Information Processing, 51(5), 1283-1293, 2010-05-15, 2010.

[2] Philipp Kolhoff, Jacqueline Preuß and J¨orn Loviscach, "Music Icons: Procedural Glyphs for Audio Files," SIBGRAPI 2006, pp.289-296, 2006.

[3] Funasawa Shintaro, Ishizaki Hiromi, Hoashi Keiichiro, Takishima Yasuhiro, and Katto Jiro, "A Proposal for Synchronized Web Image and Music Playback System using Lyrics," Proceeding of Information Science Technology Forum 8 (2), 333-334, 2009-08-20, 2009.

[4] Naonori Ueda and Kazumi Saito, "Parametric Mixture Models for Multi-labeled text," Neural Information Processing Systems 17 (NIPS2002), pp.737-744, 2002.

[5] Go Kikuchi and Hiroyuki Kasai, "Automatic Knowledge Construction of Scene Composition using Object Categorization on Images," ICCE2012 (Submitted).

[6] Takeshi Okumura, Tetsuya Takiguchi and Yasuo Ariki, "Generic Object Recognition using CRF by Incorporating BoF as Global Features," *Meeting on Image Recognition and Understanding* 2009 (MIRU2009), 2009.

[7] Xiao Hu and J.Stephen Dowine, "Improving Mood Classification in Music Digital Libraries by Combining Lyrics and Audio," International Conference on Digital Libraries, pp.159-168, 2010.