AUTOMATED MUSIC VIDEO GENERATION USING WEB IMAGE RESOURCE

Rui Cai, Lei Zhang, Feng Jing, Wei Lai, and Wei-Ying Ma

Microsoft Research Asia, Beijing 100080, P.R. China

ABSTRACT

In this paper, we proposed a novel prototype of automated music video generation using web image resource. In this prototype, the salient words/phrases of a song's lyrics are first automatically extracted and then used as queries to retrieve related high-quality images from web search engines. To guarantee the coherence among the chosen images' visual representation and the music song, the returned images are further re-ranked and filtered based on their content characteristics such as color, face, landscape, as well as the song's mood type. Finally, those selected images are concatenated to generate a music video using the Photo2Video technique, based on the rhythm information of the music. Preliminary evaluations of the proposed prototype have shown promising results.

Index Terms — music video, image search, music analysis

1. INTRODUCTION

Music video (MV) presents rich visual representation to music songs, and became popular in music entertainment since 1980s. Through MVs, people receive information both from audio and visual channels, and have more fun by exploring the cross-modal relationships among music, lyrics, and visual information. Most commercial MVs are created by professional producers; however, with the increase of both individual and online media collections, it is now possible to automatically create MVs using both local and web media resource, to satisfy the requirement of creating personalized MVs, as well as to improve visual representation of current music players like Windows Media Player.

Some previous research works on automatic music video generation have been reported in literatures [1-3]. In [1], the content of home video and the repetitive structure of music are carefully analyzed and used to align video and music. Based on [1], a system named P-Karaoke was built in [2], in which personal image collections are also added to help create Karaoke MVs. Considering personal media collections are still limited in comparison with web media resource, and usually in low quality, some works have attempted to integrate on-line media collections in MV generation. For example, a MV creator called MusicStory was proposed in [3] to generate MVs using images retrieved from Google and Flickr. In [3], words of song's lyrics were used as queries and highly relevant images ranked by search engines were returned for video rendering. Web image resource is abundant and can provide a great diversity of candidates for MV generation. However, it should also be noticed that finding appropriate images from web resource which well match songs and lyrics is not a trivial task. Ranking strategies of general image search engines in general cannot meet such a requirement.

The challenges that should be addressed in automated MV generation with web image search are listed as follows:

- How to prepare queries for image search? As not all words in lyrics are informative to one song's concept, only those salient words or phrases reflecting the song's topic should be kept for query generation. Moreover, the distribution of queries along one song's timeline should also be considered to balance the image representation of various parts of the song.
- How to evaluate the qualities of the candidate images? Here, the qualities not only indicate physical factors like blurring or exposure, but also include content qualities, i.e., whether the image content is consistent with the semantics of lyrics. Both of these qualities should be taken into consideration in re-ranking those images returned by search engines.
- How to guarantee the selected images are in a similar style? And whether such a style is appropriate for the content of the song? Although some diversity can help increase the artistry of MVs [3], those finally selected images still should be in a common style, as they are used to describe a same intrinsic topic of the song.

In this paper, we systematically investigate the above three issues, and try to provide more reasonable solutions to select web images for MV generation. First, we propose a strategy to automatically select salient words/phrases from lyrics, and generate queries for image search. To evaluate image qualities, we also introduce some image content analysis techniques such as face detection and landscape classification to re-rank those candidate images. In addition, to keep all images in a similar style, we filter images according to their main colors via associating the hue dimension of color with the mood of a song. Finally, the song's rhythm information like beat and tempo is extracted and used to help align images with music, and the Photo2Video technique [4] is adopted to convert the image sequence to a MV.

The rest of this paper is organized as follows. The architecture of the prototype is introduced in Section 2. The

implementation details of the proposed approaches are presented from Section 3 to Section 5. Section 6 gives some conclusion and our future work.

2. PROTOTYPE ARCHITECTURE

The flowchart of the proposed prototype of automated MV generation is illustrated in Fig. 1. The framework mainly consists of three parts, including: (i) pre-analysis and query generation; (ii) image search and re-ranking; and (iii) post-filtering and MV creation.

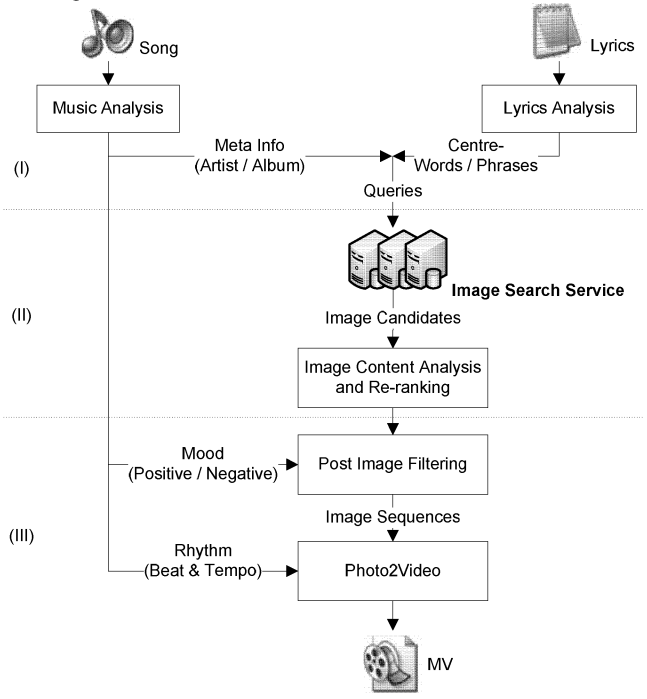

Fig. 1. The flowchart of the proposed MV generation framework: (i) pre-analysis and query generation; (ii) image search and re-ranking; and (iii) post-filtering and MV creation

Given an input song and its corresponding lyrics[1] as input, we first pre-analyze both music and lyrics to extract salient words and phrases from lyrics, and the music meta-information like names of artist and album. Such information will be utilized as queries to search relevant images. Then, through a general image search engine like MSN image search and a vertical image search engine for high-quality image search like EnjoyPhoto [5], a set of relevant image candidates are returned for each query. To satisfy the requirements of MV creation, these images are re-ranked according to their content characteristics. In the third part, based on the mood type analyzed from the song's acoustic signal, a global filtering is carried out to choose one most representative image for each query, as well as to ensure those selected images are in one color style properly matching the song's emotion. At last, the image sequence is aligned in timeline based on the beat and tempo information extracted from the song, and is converted to a MV through Photo2Video [4]. In the following sections, we will introduce the detailed implementation of these three parts.

[1] Given an input song, we can first identify it using audio fingerprinting technology in our database, and take back its lyrics. If the song is absent in our database, its meta information is extracted and used to search corresponding lyrics through some web lyrics services like "Leo's Lyrics Database" (http://www.leoslyrics.com/). In this paper, we assume the lyrics are available for simplicity.

3. PRE-ANALYSIS AND QUERY GENERATION

Query preparation is the basis of the whole framework. In the proposed prototype, the queries are generated through the pre-analysis of both the input music and its lyrics.

3.1. Music Analysis

The music analysis here includes meta-data extraction and acoustic content analysis, as shown in Fig. 1. Meta-data of a music file contains its basic properties like song title, artist, album, etc. Through investigating some commercial MVs, it is found shots of singers usually appear frequently. Thus, in automatically MV generation, it is also reasonable to integrate photos of singers and covers of albums. Here, we create two queries using the names of artist and album, respectively.

In this pre-analysis, we also extract some content-based features from music's acoustic signal, to facilitate the further post-processing. In our current implementation, these content features include mood and rhythm. For mood detection, our technique proposed in [6] is utilized, and the mood types are simplified to only two classes: *positive* and *negative*. The output of the mood detection is converted to a probabilistic score where one stands for positive and zero for negative. For rhythm analysis, we extract beat positions and tempo based on the methods introduced in [7].

3.2. Lyrics Analysis

Lyrics directly reflect the semantics of a song. Unfortunately, up to our knowledge, there is still no sophisticated method to automatically select salient words form lyrics. In [3], only stop words like "if" and "the" were removed from lyrics and remaining words are all used as queries. However, in practice, we found that besides stop words, many other words in lyrics are also useless in image search. For example, with most verbs in lyrics, it's usually hard to retrieve images being relevant with the song; while with most nouns the results are still acceptable. This indicates that a word's part of speech is important in lyrics analysis.

We conducted a user study by asking people manually labeling salient words in some lyrics. By investigating their behaviors, besides removing stop words, we summarized some heuristic rules based on which centre-words and phrases are selected:

1. Find out special phrases like location and people names in lyrics. Such special phrases usually have particular meanings for one song's conception. For example, the

word "Argentina" in the song "Don't Cry for Me Argentina" is important. In our implementation, the location and people names are detected based on a location database and an online name dictionary, respectively.

2. Find out nouns and noun phrases in lyrics. As mentioned above, nouns and noun phrases are meaningful and are ideal candidate queries for image search. Here, we first tagged the *part of speech* of each word in lyrics, and then merge two consecutive adjective and noun into one noun phrase.
3. Moreover, to balance the distribution of salient words along timeline, in situation where two candidate queries are too closed with each others, we just keep the query with more characters.

Figure 2 illustrates a part of the lyrics from the song "Yesterday Once More" by Carpenter, and the words in bold are salient-words and phrases selected according to our method.

When I was young	Those were such **happy times**	.....
I'd listen to the **radio**	And not so long ago	
Waitin' for my **favorite songs**	How I wondered where they'd gone	
When they played I'd sing along	But they're back again	.....
It made me **smile**	Just like a long **lost friend**	
	All the songs I loved so well.	

Fig. 2. Queries selected from the lyrics of "Yesterday Once More"

4. IMAGE SEARCH AND RE-RANKING

How to get relevant and high-quality images from online web resource is another key problem.

Most current online image search services can be classified into general web image search engines like Google image search and Windows Live image search, and vertical image search services like Flickr and EnjoyPhoto [5]. As introduced in Section 3.1, the queries generated include both meta-info and salient words from lyrics. Here, we adopt the strategy that sending different queries to different image search services. In more detail:

- Meta-info of the artist and album names are sent to Windows Live image search, as it can provide results of artist portrait and album covers.
- Centre-words and phrases selected from lyrics are sent to EnjoyPhoto.

This is because general search engines are with abundant image resource and are good at finding public information; while vertical image search services could focus on providing high-quality images. For example, the EnjoyPhoto service used in our framework was built by indexing about 3 million high-quality images from various photo forum web sites, and it leverages rich metadata like image title and comment from photo forums to provide more sufficient and accurate descriptions to images [5].

However, in practice we still found pictures selected by human testers are not always the top one photo returned by EnjoyPhoto. Via analyzing pictures suggested by people, we found that around 80% of those images are with human face, and around 65% are with natural sceneries. This phenomenon was also observed when examining some commercial MVs. Thus, it indicates that the rank list returned by EnjoyPhoto still cannot properly meet the requirements of MV generation; as such a rank list only considers query relevance and image qualities. To provide a more reasonable image rank for MV creation, in our framework, a re-ranking process is further conducted to integrate some image content analysis techniques for image evaluation.

Based on the above observations, for each image, we further perform face detection and indoor/outdoor classification using the techniques proposed in [8] and [9], respectively. Then the new rank value of an image r_{new} is computed as:

$$r_{new} = r_{old} \times e^{-|s_{face}/s_0 - 1|} \times P_{outdoor} \qquad (1)$$

where r_{old} is the original rank score returned by EnjoyPhoto, S_{face} is the ratio of face areas detected in the image and S_0 is the expected ratio of face areas, and $P_{outdoor}$ is the probability that the image is an outdoor scenery. With (1), the face detection and landscape classification results are combined to revise the original rank of each image. Here, we pre-set an expected ratio (S_0, which is empirically to 0.1) of face area in an image, and punish situations where faces in the image are too small or too large, using the second factor in (1).

Figure 3 gives an example of the top five images returned by Google image search, EnjoyPhoto, and the re-ranked results, of the query "happy time" from Fig. 2.

Fig. 3. Top five images for the query "happy time" returned by (a) Google image search, (b) EnjoyPhoto, and (c) the re-ranked results

From Fig. 3, it's clear that the pictures returned by EnjoyPhoto are with higher qualities than those from Google, by comparing (a) and (b). Pictures from general search engines are more diverse either on quality or on content, which make the image selection more difficult in MV generation. Moreover, the re-ranked images in (c) are more consistent in concept than those in (b), and seem to be more ideal as candidates for MVs. For example, the portrait picture (the 4th one in (b)) is removed from the top five of (c).

Consequently, after searching and re-ranking, we cache a set of image candidates (with the top 10 new rank values) for each query. The next step is to select one most representative image for each query, balancing both image qualities and harmoniousness among images of various queries.

5. POST-FILTERING AND PHOTO2VIDEO

With image material retrieved from web resource, in final MV editing, there still needs a global post-filtering to ensure that: 1) the finally adopted images are in one similar style; and 2) such a style is consistent with the song's concept and

emotion. In our framework, the post-filtering is performed by mapping dominant image colors with music emotions.

It is known that in painting, colors like blue and cyan are called "cool" colors and are usually used to exhibit emotions like sad and depressed; while colors like red and orange are called "warm" colors and are often associated with happy and pleasure. Like painting, music is also artworks reflecting some emotion styles. And we have been able to automatically estimate the probability of a song in positive mood, as described in Section 3.1. Thus, it's also reasonable and feasible to map the "cool-warm" color dimension to the "negative-positive" emotion dimension.

In implementation, for an image, its position on the "cool-warm" dimension is approximated as:

$$P_{color} = \frac{\sum_{n=1}^{N} |h_n - 0.5|}{N \times 0.5} \quad (2)$$

where h_n is the hue value (hue is related to color temperature in the HSI space) of the n^{th} pixel and there are totally N pixels in the image. With (2), the score of an image in cool colors is close to 0, and score of an image with warm colors is close to 1. Now, both the emotion and the color scores are in the range of [0, 1]. Given an emotion score predicted from music signal, for candidates of each query, we finally select the image whose P_{color} is the closest to the emotion score.

Figure 4 shows such an example. In Fig. 4, the top five image candidates of the queries from Fig. 2 are listed, and those images finally selected for each query are surrounded with red boxes. From Fig. 4, it can be seen that those selected images are mainly in cool colors such as blue and green, as the mood of "Yesterday Once More" is detected as somewhat negative—it is actually a sentimental song.

Fig. 4. The top five candidate images for the queries marked in Fig. 2, and the finally selected images are in red boxes

Finally, these images are aligned in timeline based on beat and tempo extracted the in the pre-analysis of music, as described in Section 3.1, and then converted to a video using the Photo2Video technique [4]. With Photo2Video, some camera motion patterns are simulated in concatenating images, and it makes the rendered video more vivid.

6. CONCLUSION AND FUTURE WORK

In this paper, we have presented a solution to automated MV generation through web image search. With this prototype, local music resource and global image resource are well combined to create personal MVs. In the implementation, some heuristic rules are first proposed to automatically select salient-words and phrases from lyrics; these words and phrases are then used as queries to retrieve relevant images from web image services. Furthermore, some sophisticate content analysis techniques are integrated to re-rank images, to better meet the requirements of MV creation. Finally, a global post-filtering is proposed to ensure all selected images are in a similar style which is consistent with the emotion of the song, by mapping the image color space with the music mood space.

As a prototype, it still needed more experiments to evaluate its performance on various music genres. Moreover, there is still room to improve the implementation of the proposed prototype. For example, context information in lyrics should be utilized to provide more accurate queries for image search; and to improve the quality of generated MVs, there is still more sophisticated editing rules to organize retrieved images into videos. In addition, user interaction could also be integrated into the framework, to help user create more personalized MVs. These are all directions of our future work.

7. REFERENCES

[1] X.-S. Hua, L. Lu, and H.-J. Zhang, "Automatic Music Video Generation Based on Temporal Pattern Analysis," in *Proc. 12th ACM International Conference on Multimedia*, ACM Press, New York, pp. 472-475, Oct. 2004.

[2] X.-S. Hua, L. Lu, and H.-J. Zhang, "P-Karaoke: Personalized Karaoke System," in *Proc. 12th ACM International Conference on Multimedia*, ACM Press, New York, NY, pp. 172-173, Oct. 2004.

[3] D.A. Shamma, B. Pardo, and K.J. Hammond, "MusicStory: a Personalized Music Video Creator, " in *Proc. 13th ACM International Conference on Multimedia*, ACM Press, Singapore, pp. 563-566, Nov. 2005.

[4] X.-S. Hua, L. Lu, and H.-J. Zhang, "Automatically Converting Photographic Series into Video," in *Proc. 12th ACM International Conference on Multimedia*, ACM Press, New York, pp. 708-715., New York, Oct. 2004.

[5] L. Zhang, L. Chen, F. Jing, and W.-Y. Ma, "EnjoyPhoto—A Vertical Image Search Engine for Enjoying High-Quality Photos," to appear in *Proc. 14th ACM International Conference on Multimedia*, ACM Press, Santa Barbara, Oct. 2006.

[6] L. Lu, D. Liu, and H.-J. Zhang, "Automatic Mood Detection and Tracking of Music Audio Signals," *IEEE Trans. Audio, Speech and Language Process.*, 14(1): 5-18, Jan. 2006.

[7] L. Lu, M. Wang, and H.-J. Zhang, "Repeating Pattern Discovery and Structure Analysis from Acoustic Music Data," in *Prof. 6th ACM International Workshop on Multimedia Information Retrieval*, ACM Press, NY, pp. 275-282, Oct. 2004.

[8] R. Xiao, M.-J. Li, and H.-J. Zhang, "Robust Multipose Face Detection in Images," *IEEE Trans. Circuits Syst. Video Techn.*, 14(1):31-41, Jan. 2004.

[9] L. Zhang, M.-J. Li, and H.-J. Zhang, "Boosting Image Orientation Detection with Indoor vs. Outdoor Classification," in *Proc. 6th IEEE Workshop on Applications of Computer Vision*, IEEE Press, Orlando, pp. 95-99, Dec. 2002.