AUTOMATED MUSIC SLIDESHOW GENERATION USING WEB IMAGES BASED ON LYRICS

Shintaro Funasawa† Hiromi Ishizaki‡ Keiichiro Hoashi‡
Yasuhiro Takishima‡ Jiro Katto†
†Waseda University ‡KDDI R&D Laboratories Inc.
shint@katto.comm.waseda.ac.jp, {ishizaki,hoashi,takisima}@kddilabs.jp
katto@waseda.jp

ABSTRACT

In this paper, we propose a system which automatically generates slideshows for music, by utilizing images retrieved from photo sharing web sites, based on query words extracted from song lyrics. The proposed system consists of two major steps: (1) query extraction from song lyrics, (2) image selection from web image search results. Moreover, in order to improve the display duration of each image in the slideshow, we adjust image transition timing by analyzing the duration of each lyric line in the input song. We have conducted subjective evaluation experiments, which prove that the proposal can generate impressive music slideshows for any input song.

1. INTRODUCTION

Music video, i.e., a series of visual content displayed with music, is a popular and effective way to increase the entertainability of the music listening experience. The synergetic effect generated by combining visual and audio signals is known as the sympathy phenomenon in the field of psychology [1]. While it is easy to enjoy music videos created by others (usually by experts), it is extremely difficult for common users to create music video by themselves. Namely, the cost to collect video and/or image material that is suitable for the selected music is expensive. Furthermore, the editing process to fuse the material with music also requires much intensive effort.

An important factor which reflects the image of a song is its lyrics. Many songs have lyrics which impressively represent its visual scenery, which are difficult to be extracted from their acoustic features. Numerous research efforts focusing on song lyric analysis have been presented recently. For example, extraction of song genre, topic and mood, have been investigated in recently presented work [2-5].

This paper proposes a system which generates a music slideshow automatically, by using images retrieved from the web based on query words that are derived from song lyrics. By utilizing images from the web, which provides an abundant and diverse resource of images, our proposal is able to generate slideshows of wide variety, without applying any burden to the user. In order to generate such a system, we focus on two major issues. One is the automatic extraction of words from the lyrics that are appropriate for web image search. The other is to select an optimal image to be displayed with each lyric line, from the set of candidate images obtained by web image search.

In this paper, we firstly propose a query extraction method from song lyrics based on the frequency of social tags attached to retrieved images. This method is effective to generate appropriate queries to avoid the retrieval of images that are unsuitable for slideshows. Secondly, we propose a method which selects images from the search results, based on entire impression of the song lyrics. This method is expected to increase the unity among the images within the slideshow. Moreover, we apply a method to adjust image transition time within the slideshow, by analysis of the duration time per lyric line. Subjective user evaluations will show that the proposal is capable of generating high-quality music slideshows automatically.

2. RELATED WORK

Mainly, two types of methods have been proposed for automatic generation of visual content from music. One is to generate visual contents using personal videos and/or photos [6-8], and the other is to utilize web images [9][10]. An advantage for using personal videos/photos is that the resulting slideshow will be more familiar to the user. However, in order to generate high-quality slideshows, a sufficient amount of personal material must be prepared, which is a heavy burden for casual users.

The web image-based approach has two major issues: query selection and image selection. Appropriate selection of query words is expected to be effective for the retrieval of images for slideshows. However, existing works [9][10] have utilized naive methods for query word selection, such as stop word rejection, and selection of specific parts of speech (e.g., nouns). Using values to measure the significance of words, e.g. TF*IDF, can be utilized to select query words which are significant within the lyrics. However, it is unclear whether or not such measures are appropriate to select query words for web image search to generate slideshows.

For the image selection problem, an idea has been proposed in [10] to select images containing human faces and

outdoor scenery. However, no evidence has been provided that such images are optimal for music slideshows. A naive approach is to use the top-ranked images in the search results for the image selection. In this case though, highly ranked images are expected to be selected repetitively for the same query, hence, the same image may be used for different songs with similar lyrics. Therefore, this approach is expected to generate slideshows with a lack of diversity, which may cause boredom for system users.

3. SYSTEM CONFIGURATION

The configuration of the proposed system is illustrated in Figure 1. The system selects one image for each lyric line of the input song. The selected image is displayed on the slideshow application (Figure 2) during music play. Images for the slideshow are collected from Flickr, a highly popular photograph sharing site [11], by using the Flickr API. As illustrated in Figure 1, we assume that a database which contains songs with their corresponding lyrics and timing information is prepared beforehand, as in the case of karaoke systems.

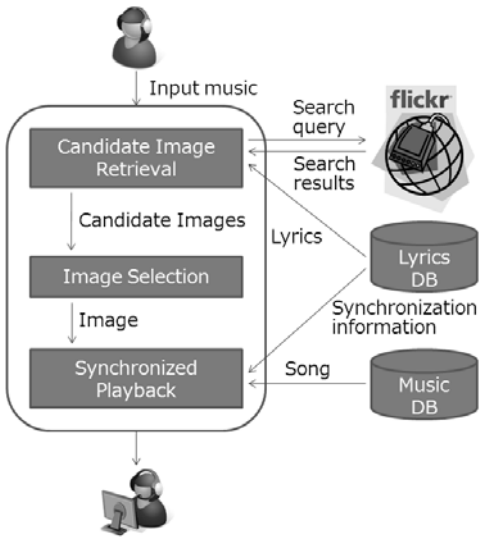

Figure 1. System configuration

The process flow of the system consists of the following three steps.

1. *Candidate Image Retrieval*
   This step extracts a candidate set of images per lyric line, by selecting appropriate query words from each line of the lyrics of the input song.
2. *Image Selection*
   This step selects an image from the previously extracted candidate image set for each line, to compose the slide show.
3. *Synchronized Playback*
   Selected images for each line are displayed with the song, according to the prepared timing information.

Figure 2. A screenshot of the proposed system

The following section explains the slideshow generation method, namely the candidate image retrieval and image selection steps, in detail.

4. SLIDESHOW GENERATION METHOD

4.1 Candidate Image Retrieval

In this step, the system generates a query (set of words) for each lyric line of the input song. The image search result from Flickr, obtained by the generated query is utilized as the candidate image set for the lyric line. The query is generated by analyzing the frequency of query words that are applied to the images in the search result, as social tags. This method is based on the hypothesis that, query words which are frequently used as social tags in Flickr have a significant meaning in the web image database, thus are expected to be effective to retrieve images which are expressive of the song lyrics. This method extracts the optimum combination of query words for each lyric line, based on the following three ideas:

- Words used in a lyric line should be prioritized, since such words accurately represent the content of the line.
- The query should be composed with as many words as possible, since such queries are more specific than single word queries, thus should result in more accurate image retrieval.
- Multiple words within a query tend to co-occur as image social tags.

4.1.1 Process Flow of Social Tag-Based Query Selection

Let $\mathbf{N}_{line}(l_i)$ represent the set of nouns used at the *i-th* line of the lyrics, $\mathbf{N}_{para}(l_i)$ represent the nouns used in the paragraph which contains the *i-th* line, and $\mathbf{N}_{all}(m)$ represent the word set which describes the general impression of song m (hereafter referred to as "general impression words", details explained in Section 4.1.2). Furthermore, when $\mathbf{W}$ expresses the set of words used as the query for

the Flickr API, let DF(**W**) (Document Frequency) represent the number of images in the search results, and UF(**W**) (User Frequency) represent the number of unique users (counted by the user ID information of the Flickr images) in the search results.

The proposed method extracts candidate query words for the *i-th* line in lyrics of music piece *m*, from $\mathbf{N}_{line}(l_i)$ and $\mathbf{N}_{para}(l_i)$. Words which have DF or UF value less than a pre-defined threshold are omitted. The thresholds for DF and UF are empirically set as 40 and 10, respectively.

Next, let $\mathbf{P}(\mathbf{N}_{line}(l_i))$ express the power set of $\mathbf{N}_{line}(l_i)$: $\mathbf{P}(\mathbf{N}_{line}(l_i))=\{\mathbf{W}_{line,1}, \mathbf{W}_{line,2}, \ldots \mathbf{W}_{line,x}\}$, where $\mathbf{W}_{line,x}$ expresses the *x-th* set of words in $\mathbf{P}(\mathbf{N}_{line}(l_i))$. From $\mathbf{P}(\mathbf{N}_{line}(l_i))$, $\mathbf{W}_{\max}$ is selected under the condition that DF($\mathbf{W}_{\max}$) is not zero and that $|\mathbf{W}_{\max}|$ is the highest in $\mathbf{P}(\mathbf{N}_{line}(l_i))$, where $|\mathbf{W}|$ expresses the number of words in **W**. If more than one $\mathbf{W}_{\max}$ can be selected, the set which has the highest UF($\mathbf{W}_{\max}$) is selected. In this way, $\mathbf{W}_{\max}$ is regarded as the set of queries for the *i*-th line, $\mathbf{Q}_{line}(l_i)$.

Then, in order to maximize the number of query words (which is assumed to reduce the number of candidate images, and improve search accuracy), we expand the query by using words in $\mathbf{N}_{para}(l_i)$. Namely, expanded sets of words, which are composed of the power set of $\mathbf{N}_{para}(l_i)$, plus the previously derived $\mathbf{Q}_{line}(l_i)$ are generated as $\mathbf{P'}(\mathbf{N}_{para}(l_i))=\{\mathbf{W}_{para,1}+\mathbf{Q}_{line}(l_i), \mathbf{W}_{para,2}+\mathbf{Q}_{line}(l_i), \ldots \mathbf{W}_{para,y}+\mathbf{Q}_{line}(l_i)\} = \{\mathbf{W'}_{para,1}, \mathbf{W'}_{para,2}, \ldots \mathbf{W'}_{para,y}\}$. Then, in the same way as explained above, $\mathbf{W'}_{\max}$ is selected from $\mathbf{P'}(\mathbf{N}_{para}(l_i))$.

Finally, by sending the all elements of $\mathbf{W'}_{\max}$ under the condition of 'AND' combination to Flickr, the system retrieves the candidate images for each line. If $\mathbf{W'}_{\max}$ has no elements, $\mathbf{N}_{all}(m)$ is used as the query.

4.1.2 Estimation of General Song Impression

As mentioned above, $\mathbf{N}_{all}(m)$ is the general impression word set, i.e., a set of words which expresses the collective impression of song *m*. This word set can be used for lyric lines from which no effective query words could be extracted. Furthermore, the general impression word set is also effective to generate slideshows with a sense of unity, as will be described in the next section.

The general impression of a song is estimated by text-based classification based on its entire lyrics. Namely, song classifiers are preliminarily constructed by SVM [12] for each of the categories showed in Table 1. The categories are divided into three concepts: Season, Weather, and Time. Each concept consists of several categories. The concepts/categories in Table 1 are selected because they all represent important aspects of song lyrics, and are expressed by discriminative words. For the classifier, we used the software *SVM*^{light} [13] with a linear kernel for learning. Here, lyrics have been vectorized by TF*IDF, and the training data for the classifiers learning have been obtained by a manually collected database of Japanese pop songs with human-applied labels. If the classifier determines that a song *m* is positive for its respective category, the name of the category is added to $\mathbf{N}_{all}(m)$. Note that multiple words may be included in $\mathbf{N}_{all}(m)$.

Concepts	Category labels
Season	Spring, Summer, Autumn, Winter
Weather	Sunny, Cloudy, Rain, Snow, Rainbow
Time	Morning, Daytime, Evening, Night

Table 1. Concepts and category labels for describing general impression of music.

4.2 Image Selection

The next step is to select an image to compose the slideshow from the candidate image set for each lyric line. We propose an image selection method based on an impression score, which represents strength of association between the image and the general impression words of the input song. Consideration of the impression score is expected to select images that are more fitting to the overall theme of the input song, thus increases the sense of unity among the images which compose the slideshow.

4.2.1 Relevant Tag Extraction Based on Co-occurrence Probability

Relevant tags for calculating the impression score are extracted based on co-occurrence probability of social tags on Flickr. In this paper, the co-occurrence probability is calculated based on UF instead of DF, since there are many tags with unusually high DF on Flickr, due to users who upload many images with the exact same tag set, while UF is more robust to the effect of such user behavior.

The relevance score between a general impression word $n_{all} \in \mathbf{N}_{all}(m)$, and a given tag *t,* is calculated by the co-occurrence probability of *t* and n_{all}, and also the impression words which belong to the same concept as n_{all}. For example, when the relevance score between "summer" and tag *t* is calculated, the same score for all other general impression words in the "Season" concept, i.e., "spring", "autumn", and "winter", are also calculated. In this way, it is possible to extract tags which have specifically high relevance to n_{all}, and decrease the score of generally popular tags, i.e., words which co-occur frequently with many other words.

The co-occurrence score between general impression word n_{all} and tag *t*, $CoScore(t,n_{all})$, is defined as:

$$CoScore(t, n_{all}) = \frac{UF(t \cap n_{all})}{UF(n_{all})} \quad (1)$$

Then, the relevance score R between n_{all} and t is defined as:

$$R(t, n_{all}) = CoScore(t, n_{all}) - \frac{\sum_{n \in C, n \neq n_{all}} P(t \mid n)}{|C| - 1} \times wgt \quad (2)$$

where $\boldsymbol{C}$ is the set of general impression words which belong to the same concept of n_{all}. For example, when n_{all} = "spring", $\boldsymbol{C}$ = {"spring", "summer", "autumn", "winter"}, since "spring" belongs to the "Season" concept. In the definition of the relevance score in Eq.(2), the first term increases the score of tags which have high co-occurrence probability with n_{all}. Subtraction of the second term decreases the score of tags with high co-occurrence probability of impression words which belong to the same concept as n_{all}. Note that *wgt* is a coefficient to adjust the impact of the second term. This coefficient is set to 3, empirically.

Based on Eq.(2), the relevance score between each general impression word, and all tags which co-occur with the general impression word, are calculated. Tags whose relevance scores are over 0.024, and UF value exceeds 5, are regarded as relevant tags of each impression word.

4.2.2 Definition of Impression Score

For image selection, we calculate the impression score for all images in the candidate image set, based on the tags applied to the image, and the above relevance score. The object of this method is to select images with tags which have high relevance to the general impression words of the input song. As a result of this process, the impression score of images with "noisy" tags, *i.e.*, tags with low relevance to the general impression of the input song, will be degraded.

The impression score of image i is determined by

$$score(i) = \sum_{n_{all} \in N_{all}(m)} \frac{\sum_{t \in T_i \cap T_{related}(n_{all})} R(t, n_{all})}{|T_i| - |T_i \cap T_{related}(n_{all})|} \quad (3)$$

where $\boldsymbol{T}_i$ is the set of tags applied to image i, $\boldsymbol{T}_{related}(n_{all})$ is the relevance tag set of general impression word n_{all}, and $R(t, n_{all})$ is the relevance score between n_{all} and tag t.

This impression score is computed for each candidate image, and the image with the highest score is selected to be displayed with its respective lyric line, during the slideshow.

4.3 Image Transition Timing Adjustment

In the proposed system, the images obtained per lyric line are displayed in synchronization with each line during the song playback. Adequate usage of the line information leads to natural image transition during the slideshow, since lines represent a semantic unit in the lyrics. However, display duration of each image may be too short/long when using the line information naively. For example, in a rap song with many lines, the image display time maybe too short, so that users may not be able to comprehend the images in the slideshow. On the other hand, in a slow ballad song, images may be displayed for a long time, which may cause boredom.

In order to improve the overall quality of the slideshow, we propose an image transition timing adjustment method, which adjusts the display time of images according to the duration of each lyric line. In this process, we first estimate the typical duration time of images in a song. Then, the line of lyrics is "combined" or "divided", based on the difference of the duration of the line and the typical duration time of the input song. In the "combining" process, lines with short duration time are combined with their adjacent lines, and a single image is displayed for the combined set of lines. In the "dividing" process, lines with long duration time are "divided" into plural sub-lines, and an image is to be displayed along with each sub-line.

The process flow for image transition timing adjustment consists of the following steps.

1. Typical duration time of song m is calculated from the lyrics data. Namely, the mode value of the line duration time is used as the typical image duration time I_m.
2. Lines which have less than 4 [sec] duration time are "combined" with the next line. If there is no next line, it is "combined" with the previous line. However, lines are "combined" only if they belong to the same paragraph. An image is retrieved for the newly combined line.
3. A line which has more than 12 [sec] duration time is "divided" equally. The number of divisions is controlled so that approximate duration time of the new "divided" line is equivalent to I_m. When the line is "divided" into n lines, n images are displayed from the candidate image set, which is retrieved based on the lyrics of the original line.
4. Interlude sections (which generally have no lyrics) are divided by the same process as step 3. The general impression words are used as query for image retrieval.

5. EXPERIMENTS

5.1 Outline

In order to evaluate the quality of the proposed method, we have conducted a subjective evaluation experiment. This experiment compares the proposed method with other conventional methods, by asking 42 subjects to rate the slideshows generated by all methods. The subjects are asked to view the music slideshows of the same song, which are generated by the proposed and comparative methods (details of the methods are explained in Section

5.2). Then, each subject is asked to apply a five-ranked rating for each slideshow, based on the following evaluation measures:

a) Accordance between lyrics and images [content]
b) Appropriateness of image display time [duration]
c) Unity of all images in slideshow [unity]
d) Overall quality [quality]

In this experiment, we use 10 Japanese pop songs and 28 ~ 29 subjects have provided evaluation results for each song. In order to evaluate the method described in Section 4.3, we have selected songs so that half of these songs include "combined" lyric lines (hereafter referred to as the "combined set") in the process of adjustment of image transition timing explained in Section 4.3, and the other half include "divided" lines (hereafter referred to as the "divided set").

5.2 Evaluated Methods

The next three methods were evaluated and compared.

A) *MusicStory* [9]

The first comparative method generates slideshows based on the method proposed for MusicStory [9]. Namely, all nouns are extracted from the entire lyrics of the input song, and are sent to the Flickr API under the 'OR' combination. The images in the search result are displayed according to the transition timing determined by the BPM (beats per minute) of the input song

B) *TF*IDF based method*

The second comparative method extracts query words from the lyrics based on TF*IDF. The process flow to obtain the image for the *i-th* line in the lyrics is described as follows. First, the nouns extracted from the *i-th* line in the lyrics, are sent to Flickr as query under the condition of 'AND' combination. If the result has no images, the noun with the smallest TF*IDF is removed, and the rest of the nouns are sent to Flickr again. This process is repeated until a set of images are obtained. If the system is unable to retrieve images by any of the nouns in the line, the images from the previous line are re-used. Finally, the highest-ranked image in the search result (according to the Flickr "interestingness" ranking) is selected. Images are obtained for each line and switched in synchronization with line appearance within input song. In this paper, the DF element of TF*IDF is calculated based on our database, which contains 3062 Japanese pop songs.

C) *Proposed method*

The third method is our proposal. Queries are generated from the lyrics by the social tag-based method, images are selected from the image search results based on the impression score, and the image transition timing is adjusted by the method described in Section 4.3.

5.3 Experimental Results

Figure 3 shows the average rating of all subjects, for each evaluation measure and method. The results in this Figure show that the proposed method has received the highest ratings, compared to the other methods for all evaluation measures. Most significantly, the proposed method has received the best rating for the overall quality, a difference which is statistically significant to the others based on t-test ($p<0.001$). These results prove that the proposed method is capable of generating high-quality slideshows.

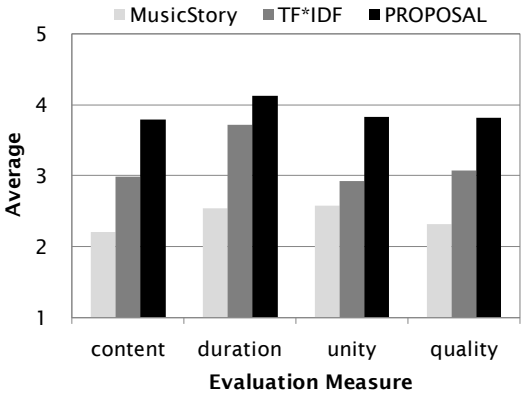

Figure 3. Average ratings of each evaluation measure.

Lyrics line	*"If this separation means departure, I will give my all smiles to you."*	
Query	TF*IDF based	"departure" (line word)
	Proposal	"smile" (line word)
Lyrics line	*"Will the memory of our encounter and the town we'd walked in be kept in our heart?"*	
Query	TF*IDF based	"heart" (line word)
	Proposal	"town" (line word)
Lyrics line	*"I wish I could stay with you for even a moment."*	
Query	TF*IDF based	"moment" (line word)
	Proposal	"car" and "night scene" (paragraph words)
Lyrics line	*"But the shining days will never return to me today or tomorrow."*	
Query	TF*IDF based	"today" (line word)
	Proposal	"evening" (general impression word)

Table 2. Examples of image search queries generated by TF*IDF based and proposed methods.

In order to analyze the query selection process of the proposed method, we compare the queries generated by the proposal to those of the TF*IDF based method. Examples are written in Table 2. This table shows examples of lyrics lines (English translations by the authors from the original Japanese lyrics) and the queries generated from the lines by the two methods. In the first two examples in this table, it is clear that the proposal has success-

fully selected words which represent visual concepts. Contrarily, the TF*IDF method has selected words which are important, but also are difficult to be represented in a visual manner. This is due to the characteristic of the proposed method, which considers the UF values of the words in Flickr. Furthermore, when there are no "visual" words in the lyrics, the proposal can appropriately generate queries, either from the lyric paragraph, or general impression words, as shown in the last two examples. These examples indicate that the proposed method is effective to generate good queries from any song lyric.

Moreover, even when the queries generated by the both methods are the same, the proposal is capable of selecting more suitable images for the song. For example, when both methods retrieve images by the query "town" for a winter song, the proposal appropriately selects an image of a town with falling snow, while the TF*IDF based method selects a general image of a town. Examples like this indicate that the proposed image selection method based on impression score can generate suitable slideshows which represent the overall theme of the song.

Additionally, in the "duration" measure, the proposal has achieved ratings superior to the TF*IDF based method for 9 songs, indicating that the proposed adjustment method has succeeded in improving slideshow quality. The difference of the average ratings between the proposal and the TF*IDF based method for "combined sets" is 0.21, while the difference for "divided sets" is 0.61. This result implies that the proposed method is more effective to improve slideshows for songs with lyrics that are slowly sung, as in slow ballads.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a system to generate slideshows for any given song, by using words in their lyrics to retrieve web images. We have proposed a query generation method for image search and an image selection method to compose slideshows from the image search results. Moreover, we proposed a method to adjust image transition timing based on the lines of lyrics. Results of subjective evaluations have shown that our system can generate highly satisfactory music slideshows.

In the future, we plan to expand our system to utilize not only the lyrics, but also the acoustic features of the input song. For example, displaying slideshows with various effects, such as zooming and panning, in accordance with the excitement of the song; as well as the use of beat information for image transition all are expected to improve the impression of the generated slideshows.

7. REFERENCES

[1] S. Iwamiya: "The interaction between auditory and visual processing when listening to music via audio-visual media," *The Journal of the Acoustical Society of Japan*, Vol.48, No.3, pp.146-153, 1992. [in Japanese]

[2] R. Mayer, R. Neumayer, and A. Rauber: "Rhyme and Style Features for Musical Genre Classification by Song Lyrics," *Proceedings of ISMIR 2008*, pp. 337-342, 2008.

[3] F. Kleedorfer, P. Knees, and T. Pohle: "Oh Oh Oh Whoah! Towards Automatic Topic Detection in Song Lyrics," *Proceedings of ISMIR 2008*, pp. 287-292, 2008.

[4] Y. Hu, X. Chen, and D. Yang: "Lyric-based Song Emotion Detection with Affective Lexicon and Fuzzy Clustering Method," *Proceedings of ISMIR 2009*, pp. 123-128, 2009.

[5] X. Hu, J. S. Downie, and A. F. Ehmann: "Lyric Text Mining in Music Mood Classification," *Proceedings of ISMIR 2009*, pp. 411-416, 2009.

[6] X. -S. Hua, L. Lu, and H. -J. Zhang: "P-Karaoke: Personalized Karaoke System," *Proceedings of the 12th Annual ACM International Conference on Multimedia*, pp.172-173, 2004.

[7] T. Terada, M. Tsukamoto, and S. Nishino: "A System for Presenting Background Scenes of Karaoke Using an Active Database System," *Proceedings of the ISCA 18th International Conference on Computers and Their Applications*, pp. 160-165, 2003.

[8] S. Xu, T. Jin, and F. C. M. Lau: "Automatic Generation of Music Slide Show using Personal Photos," *Proceedings of 10th IEEE International Symposium on Multimedia*, pp. 214-219, 2008.

[9] D. A. Shamma, B. Pardo, and K. J. Hammond: "MusicStory: a Personalized Music Video Creator," *Proceedings of the 13th Annual ACM International Conference on Multimedia*, pp.563-566, 2005.

[10] R. Cai, L. Zhang, F. Jing, W. Lai, and W. -Y. Ma,: "Automated Music Video Generation Using Web Image Resource," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007, Vol.2, pp. 737-740, 2007.

[11] Flickr: http://www.flickr.com/

[12] C. Cortes and V. Vapnik: "Support Vector Networks," *Machine Learning*, Vol. 20, pp.273-297, 1995.

[13] SVM-Light Support Vector Machine: http://svmlight.joachims.org/