Tenth IEEE International Symposium on Multimedia

Automatic Generation of Music Slide Show using Personal Photos

Songhua Xu[♮,♯,‡,*] Tao Jin[‡] Francis C.M. Lau[‡]

♮: College of Computer Science and Technology, Zhejiang University
Hangzhou, Zhejiang, P.R. China, 310027
♯: Department of Computer Science, Yale University
New Haven, Connecticut, USA, 06520-8285
‡: Department of Computer Science, The University of Hong Kong
Pokfulam Road, Hong Kong, P.R. China
*: Correspondence author. Contact him by songhua DOT xu AT gmail DOT com.

Abstract

We present an algorithmic system capable of automatically generating a music slide show given a piece of music with lyrics. Different from previous approaches, our method generates slide shows using personal photos which are without annotation. We introduce a novel algorithm to infer the relevance of personal photos to the lyrics, based on which personal photos are optimally selected to match the music. The proposed system first detects the keyframes of the input music. For each music keyframe, it optimally selects an image from the personal photo collection via our image content analysis procedure. Once the keyframe images have been selected, the in-between frames are then generated via an image morphing process. Experiment results have shown that our method can successfully generate music slide shows which follow the rhythms of the music and at the same time match the lyrics.

1 INTRODUCTION

Music video is a popular type of media contents on the Internet. However, due to their somewhat higher production costs, the amount of online music video is significantly smaller than that of online music. Sight and sound together is certainly more entertaining than pure sound; so many modern music players would add visualization to the music being played. In most cases, however, the visualization is pre-programmed to fit any kind of music, and can not be easily changed to reflect the character (e.g., the mood, rhythm, etc.) of the music. Ideally, every piece of music should have its own custom visualization—that is, visualization that "dances to the tune". To choreograph a visualization to fit a piece of music is however non-trivial, and time-consuming if done manually, even for just a single piece. This motivates our research on automatic generation of visuals to accompany music.

In this paper, we introduce an algorithmic system to automatically generate a music slide show for a given input piece of music by taking into account the rhythms and the lyrics of the input. The key difference from previous music video or music slide show generation approaches is that our method takes personal photos as the source to generate the slide show video. We pick personal photos because they are part of everybody's life, and showing them along with the music would add a sense of intimacy. And the result can also be seen as a way to browse and enjoy personal photos with nice music in the background. However, there are challenges with using personal photos. A big one is that images in a personal photo collection are rarely annotated which makes automatic image understanding by computers difficult or impossible. And even if some may be annotated, the texts would likely not be reliable. To solve the problem, we introduce a novel algorithm to infer the relevance of personal photos with respect to the given music lyrics by referring to similar online images with text annotations from the Internet. Benefitted from this algorithmic process, we can optimally select personal photos to match the keyframes in the input music.

The rest of the paper is organized as follows. We survey the most related work in Sec. 2. Sec. 3 presents our problem statement and overviews the main ideas behind our algorithm design. Sec. 4 discusses how to determine the image transition point in the music slide show according to the rhythms in the music. Sec. 5 introduces how to optimally select images from a personal photo collection to match the music lyrics. Sec. 6 discusses how to generate images for

978-0-7695-3454-1/08 $25.00 © 2008 IEEE
DOI 10.1109/ISM.2008.39

the gap between adjacent selected images through an image morphing process. Sec. 7 presents some experiment results and Sec. 8 concludes the paper.

2 RELATED WORK

We survey the most related work on automatic generation of visuals to accompany the music below.

There have been some attempts to build automatic music video generation systems, e.g., [2], [4]. The system described in [2] takes an audio clip and video as the input. The video is segmented and analyzed. Shots with fierce camera motion or poor contrast are removed. From the remaining video shots, suitable video shots are selected to tag the different parts of the audio. The criterion they use in video shot to audio track alignment is to match the video shot boundaries with the major peaks in the audio. Their system also provides interactive control for the user to edit the video. The music video system proposed in [4] refines the above method by incorporating some more elaborate content analysis algorithms. The visual elements used for music video generation in their system are scenes from home videos, which are automatically extracted from the source videos through shot segmentation. Given a clip of home video, their algorithm first applies content analysis to the input video to derive the motion intensities in its shots. And then for a piece of user selected music, the music's rhythms and repetitive patterns are analyzed. Their algorithm then assigns input video shots to various parts of the music so that the tempos of the music (in terms of repetitive patterns in the music) are reflected by the motion intensities in the video shots. The main difference between their algorithm and the algorithm we propose in this paper is that they use video shots and we use still photos from a personal photo collection. Both types of music visual accompaniment add entertainment values to the original music. We argue however that using photos for the visualization would make the resulting system more accessible to more users simply because photos or still pictures are more abundantly and conveniently available to most users than videos, and photos are much easier to manipulate by anyone. In fact, most users today keep countless personal photos on their computer, but a small amount of personal videos, or none at all. Also, videos might contain too much motion and thus cause distraction to music listening. For these reasons, we favor music slide shows and present a system in this paper for automatically generating a slide show from personal photos.

There also exist some work on automatic music accompanying visual generation based on images. The P-karaoke system by Hua et al. [6] can generate karaoke videos from personal photos. Their system is capable of converting single photos into a motion photograph clip using simulated camera motions. The resultant motion photograph clip is then used to garnish the music, namely to align video shot boundaries with the music beats. Unlike their approach which emphasizes on the animation, our system focuses (through intense content analysis) on matching personal photos to the lyrics. Later, Hua et al. extended their work by including style and template support [7]. Shamma et al. [10] proposed a personalized music video creation system which can generate a music video for a piece of music using images found through public image search engines. But they did not elaborate on how to find images that are most related to the lyrics. Most recently, Cai et al. [1] proposed an automatic music video generation approach using images from the Internet. Their work is the most related study to our work here. In choosing images to use for composing the visual contents of music videos, they use a hand coded heuristic to re-rank images obtained from online image search. The heuristic considers the likelihood the photo is taken outdoor and also the percentage in the image that is occupied by a human face. Thus their image ranking method is biased towards photos taken with larger human faces and outdoor. Unlike their method, we introduce an image content analysis method to infer the relevance of personal photos to text queries arising from the lyrics. Thus our method represents an improvement over their image reranking method for producing more semantically meaningful music slide shows using personal photos.

Also related is the automatic home video editing system introduced in [5], where home videos are segmented and matched with a music piece according to the beats and tempos of the music as well as the video content. Their system can optimally select a music piece to match a given video, which is in fact the reverse problem of the music slide show generation task we study here.

3 OVERVIEW

This paper studies the problem of generating a music slide show for a given input piece of music. The generation is considered successful if the resultant video appears to synchronize with the rhythms of the music as well as match the lyrics. Although there are other attributes of music (such as tempo, timbre, tone color, mode, dynamics, etc.) which could be used for the synchronization task, we find rhythm to be most natural for guiding the behavior of the accompanying visualization. Here we assume the input music piece contains lyrics that are correctly synchronized. We stress the point again that the visual contents used for generating a music slide show in our method come from photos in a personal photo collection.

The key steps of our algorithm to generate a music slide show are as follows. Given a piece of input music, we first detect its keyframes according to the rhythms. For each of the keyframes, the algorithm optimally selects an image from the personal photo collection to match the corresponding lyrics. After all the music keyframes have been assigned an optimal image, for every pair of two adjacent key frame images, we generate the in-between frames through an image morphing procedure. In the above steps, one key algorithmic decision is in the selection of images from the personal photo collection to accompany the music so that the image contents would tally with the lyrics. Because in a typical personal photo collection, images are mostly not annotated, we introduce a method to infer the contents of these unannotated photos based on images retrieved from the Internet.

4 DETERMINING PHOTO TRANSITION POINTS VIA MUSIC KEYFRAME DETECTION

4.1 Music features and pairwise music distance

Many of the operations we need deal with music features and pairwise music distance. Therefore, we look at these two issues first before discussing the algorithm.

4.1.1 Music features

The speech recognition community has proposed a number of audio features for tackling various audio signal processing problems such as to characterize and describe sounds, to classify musical instruments by their sounds, and to perform psycho-acoustical studies. In developing the music sound classification system [3], Herrera summarized and classified the most popular audio features into six groups [9]. In this paper, we adopt the following audio features which have been proposed in existent literature: a) Spectral Centroid (SC), which is the balancing point of the spectrum. b) Spectral Flux (SF), which is the L$_2$-norm of the difference between the magnitudes of the Short Time Fourier Transform spectrum (STFT) evaluated in two successive sound frames with energy normalization. c) Mel-Frequency Cepstral Coefficients (MFCC), which are commonly used in speech recognition. They are a perceptually motivated compact representation of the spectrum. d) Linear Prediction Reflection Coefficients (LPC), which are used in speech research as an estimate of the speech vocal tract filter. e) Root Mean Square (RMS) of the audio signal, which is defined as the square root of the sum of the squared values of the signal. f) Spectral Rolloff (SR)—for example, if the percentage is 0.90 then Rolloff would be the frequency where 90% of the energy in the spectrum is below this point. These audio features are abbreviated as SC, SF, $MFCC$, LPC, RMS, and SR respectively. The code for calculating them is available from an open source software framework, Marsyas (Music Analysis, Retrieval and Synthesis for Audio Signals). For each of these features X and for each music frame M_i, we calculate the feature's mean value $\overline{X}$ and variance $\delta(X)$ over the music duration spanned by the preceding five frames of M_i as well as its following five frames. We use $\mathbf{F}(M_i)$ to represent them, i.e., $\mathbf{F}(M_i) \triangleq \big(\overline{SC}, \delta(SC), \overline{SF}, \delta(SF), \overline{MFCC}, \delta(MFCC), \overline{LPC}, \delta(LPC), \ \overline{RMS}, \delta(RMS), \overline{SR}, \delta(SR)\big)$. The vector $\mathbf{F}(M_i)$ becomes our audio feature vector for M_i. Given the above definition over the music features for a single music frame M_i, we can easily define the music features for an entire music piece. The basic equation is the same as the one above for a single frame; the only difference is the music duration being spanned. For simplicity, we do not differentiate these two types of music features and use the notation $\mathbf{F}(M_i)$ to refer to either a single frame or an entire piece of music.

4.1.2 Pairwise music distance

Given a pair of music pieces M_x and M_y, once their music feature vectors are derived as $\mathbf{F}(M_x)$ and $\mathbf{F}(M_y)$, we can define the pairwise music distance between them as the square root sum of the differences between their corresponding feature vector components, i.e.: $\theta_{music}(M_x, M_y) \triangleq \sqrt{\sum_{k=1}^{12} \big(\mathbf{F}_k(M_x) - \mathbf{F}_k(M_y)\big)^2}$, where $\mathbf{F}_k(M_x)$ and $\mathbf{F}_k(M_y)$ are the k-th components of the feature vectors $\mathbf{F}(M_x)$ and $\mathbf{F}(M_y)$ respectively.

4.2 Music keyframe detection

We need to detect the keyframes of the input music to serve as targets for our algorithm to try to optimally select images to match. Keyframes are those musical moments at which the music exhibits some substantial changes. For those in-between frames, we generate their images using an image morphing procedure, to be discussed in Sec. 6.

For each frame M_i in a piece of music, we compare its audio features $\mathbf{F}(M_i)$ with the audio features of its ten neighboring music frames. When the average difference exceeds a certain threshold, we regard the frame as a music keyframe. Mathematically, M_i will be identified as a music keyframe if

$$\frac{1}{10} \sum_{j=i-5, j\neq i}^{j=i+5} \|\mathbf{F}(M_j) - \mathbf{F}(M_i)\| > \varepsilon, \qquad (1)$$

where $\|\mathbf{x}\|$ computes vector $\mathbf{x}$'s magnitude. We empirically set ε as one tenth of the average magnitude of a music frame's audio feature vector throughout the music piece, i.e., $\varepsilon = \frac{1}{10} \frac{\sum_i \|\mathbf{F}(M_i)\|}{\sum_i}$. We also constrain two adjacent music keyframes to be at least 1 second and no more than 5 seconds apart. A minimum of 1 second because the user needs time to see a photo; maximum 5 seconds so that the user will not be bored looking at the same picture for too long.

5 OPTIMALLY ASSOCIATING PERSONAL PHOTOS WITH INPUT MUSIC

5.1 Main idea

In our system, we would like to optimally associate personal photos with the input music so that the content in the photo would be a best possible match with the meaning of the corresponding lyrics. Unfortunately, personal photos are seldom annotated, which makes our problem difficult. Even when some of them are annotated, the annotations usually are not reliable or comprehensive. To solve this problem, we introduce a novel image content analysis method which can infer image-to-text-query relevance for personal photos, i.e., images with no text annotation. The idea is to find images with annotations in the Internet and then compute the image content similarity between these reference images and images in the personal photo collection. Through this reference, our algorithm may be able to estimate the relevance of a personal photo with respect to a text query.

5.2 Our method

We have in Sec. 4.2 introduced the method to detect music keyframes. Since in our experiment, we assume the lyrics of the song are available and are in-sync with the music, we thus can cluster all the lyric words according to the breakpoints detected (i.e., the music keyframes). Assuming the music keyframes are at time moments $t_1, t_2, \cdots, t_n$ respectively, all the lyric words that fall into the time period $[0, t_1]$ form a word sequence $\mathbf{ws}_1 \triangleq \{w_{1,1}, w_{1,2}, \cdots, w_{1,l_1}\}$ (assuming there are l_1 such words). Similarly, we denote the lyric words that fall into the time period $[t_i, t_{i+1}]$ as $\mathbf{ws}_i \triangleq \{w_{i,1}, w_{i,2}, \cdots, w_{i,l_i}\}$ (assuming there are l_i such words). We call each $\mathbf{ws}_i$ a lyric segment. In the clustering process, we remove stop words in the lyrics. We then sequentially select images to match each lyric segment. That is, we first optimally select an image from the personal photo collection Λ to match the first lyric segment $\mathbf{ws}_1$, and then optimally select another image to match the second lyric segment $\mathbf{ws}_2$, and so on.

For a given lyric segment $\mathbf{ws}_i$, we submit it as a search phrase to a commercial image search engine to obtain some reference images. In our current experiment setting, we use Google Image Search.Assuming there are $z(\mathbf{ws}_i)$ images returned as search result, forming a reference image set as $\mathbf{I}(\mathbf{ws}_i) \triangleq \{I_1(\mathbf{ws}_i), I_2(\mathbf{ws}_i), \cdots, I_{z(\mathbf{ws}_i)}(\mathbf{ws}_i)\}$.

The optimal image-to-query relevance score between the lyric segment $\mathbf{ws}_i$ and an image I_x in the personal photo collection Λ is estimated as:

$$\eta(I_x, \mathbf{ws}_i) \triangleq \frac{\sum_{j=1}^{m} \theta_{img}\big(I_x, I_j^{ref}(I_x, \mathbf{ws}_i)\big)}{m}, \quad (2)$$

where $\theta_{img}(I_x, I_y)$ computes the image content similarity between the pair of images I_x and I_y based on the Scale Invariant Feature Transform (SIFT) image features [8]; the images $I_1^{ref}(I_x, \mathbf{ws}_i)$, $I_2^{ref}(I_x, \mathbf{ws}_i)$, $\cdots$, $I_m^{ref}(I_x, \mathbf{ws}_i)$ are the m distinct reference images from the image set $\mathbf{I}(\mathbf{ws}_i)$ which yield the highest image content similarity scores with I_x. In our current experiment, we empirically tune $m = 3$.

Thus the image $I_{\text{opt}}(\mathbf{ws}_i)$ which we optimally select from the personal photo collection Λ to match the lyric segment $\mathbf{ws}_i$ is determined as:

$$I_{\text{opt}}(\mathbf{ws}_i) \triangleq \arg\max_{I_x \in \Lambda} \eta(I_x, \mathbf{ws}_i). \quad (3)$$

Since showing the same photo multiple times tends to bore the viewer, in the above optimal image determination process, we only search among images in the personal photo collection Λ which have not been previously selected to accompany a lyric segment.

6 PRODUCING IN-BETWEEN FRAMES

For each music keyframe, we have optimally selected an image to match with it (Sec. 5.2). To generate the in-between frames, we adopt a morphing based approach. Assume we have two music keyframe images M_1 and M_2, whose in-between frames need to be generated. We use the multi-resolution image morphing algorithm [26]. The morphing algorithm operates entirely automatically. However, there is a free parameter $\alpha \in [0, 1]$ involved in the morphing process which specifies how closely the generated image resembles the two source images respectively: If $\alpha = 0$, then the morphing result image is the same as M_1; if $\alpha = 1$, then the morphing result image is the same as M_2; in general, the bigger the α value, the more similar to M_2 the morphing result image would look. In the original morphing algorithm, the α values are uniformly sampled from the interval [0, 1] for generating the intermediate morphing images. This is the standard way adopted by morphing algorithms designed for computer graphics applications. In our system, to make the intermediate images appear more reflective of the flow of the underlying music, instead of using the standard uniform distribution to assign values for α, we make these values dependent on the rhythms of the underlying music. More concretely, for an intermediate music frame M_x to be processed, let its music content distances to

M_1 and M_2 be $\theta_{music}(M_x, M_1)$ and $\theta_{music}(M_x, M_2)$ respectively, which are computed via the method introduced at Sec. 4.1.2. Then the α value we choose for generating the morphing result image for the music frame M_x is determined as: $\alpha(M_x) \triangleq \frac{\theta_{music}(M_x,M_1)}{\theta_{music}(M_x,M_1)+\theta_{music}(M_x,M2)}$. To verify whether this simple equation satisfies our need, if M_x sounds similarly to M_1, then $\theta_{music}(M_x, M_1) \longrightarrow 0$ so that $\alpha(M_x) \longrightarrow 0$. In the same way, if M_x sounds similarly to M_2, then $\alpha(M_x) \longrightarrow 1$. Therefore, through the above equation we can effectively make the images generated through the above morphing process for the in-between music frames closely reflect the rhythm of the music piece. Once all the in-between frames between every pair of adjacent music keyframes are generated, the music slide show for the given music piece is eventually produced.

7 EXPERIMENT RESULTS

Figures 1–2 present two generation results from our experiments. Both the personal photos identified and the corresponding lyric segments are shown. Due to space limit, only the top two personal photos identified as most relevant to the respective lyrics are included. These personal photos are obtained through mining personal blogs of graduate students in the authors' institutions as well as the personal blogs of these students' immediate friends. Examining these photos along with their accompanying lyrics, most of the pictures seem to be able to express the semantics of the lyric segment or the keyword in the lyric segment vividly. These results confirm that the personal photos identified by our algorithm are satisfactorily matched to the song lyrics. Hence this proves the algorithm introduced at this paper can effectively augment the music listening experience through automatically generating music slide show using personal photos.

In terms of the time performance of our algorithm, downloading the top 200 images ranked by Google Image Search took around 10 minutes using a local ISP for home users; and the re-ranking process took less than 10 seconds. Thus, the overall time spent is dominated by the online image downloading step. Such a feature is advantageous for the wide deployment of our algorithm since in many home and office situations the web access bandwidth is not fully utilized most of the time.

8 CONCLUSION

We have presented an algorithmic system capable of automatically generating a music slide show given a piece of music with lyrics. The key contribution of our work is that our system can generate music slide shows using personal photos and the visual contents in the generated video correspond to both the lyrics as well as the rhythms of the music. This is realized through an image content understanding process which infers the relevance of personal photos without text annotations with respect to a text query based on reference images obtained from an online image search process. Besides using lyrics to optimally select the photos, we also employed a music rhythm analysis process which detects music keyframes to be used as photo transition points in the music slide show. Such an analysis is based on music features that collectively represent the rhythms of the music. This is different from prior work, e.g., [1], which simply uses beat positions, tempo or mode for determining the photo transition points.

References

[1] R. Cai, L. Zhang, F. Jing, W. Lai, and W.-Y. Ma. Automated music video generation using web image resource. *ICASSP '07: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2:737–740, April 2007.

[2] J. Foote, M. Cooper, and A. Girgensohn. Creating music videos using automatic media analysis. In *MULTIMEDIA '02: Proceedings of ACM International Conference on Multimedia*, pages 553–560, New York, NY, USA, 2002. ACM.

[3] P. Herrera, G. Peeters, and S. Dubnov. Automatic classification of musical sounds. *Journal of New Musical Research*, 32(1):3–21, 2003.

[4] X.-S. Hua, L. Lu, and H.-J. Zhang. Automatic music video generation based on temporal pattern analysis. In *MULTIMEDIA '04: Proceedings of ACM International Conference on Multimedia*, pages 472–475, New York, NY, USA, 2004. ACM.

[5] X.-S. Hua, L. Lu, and H.-J. Zhang. Optimization-based automated home video editing system. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5):572–583, 2004.

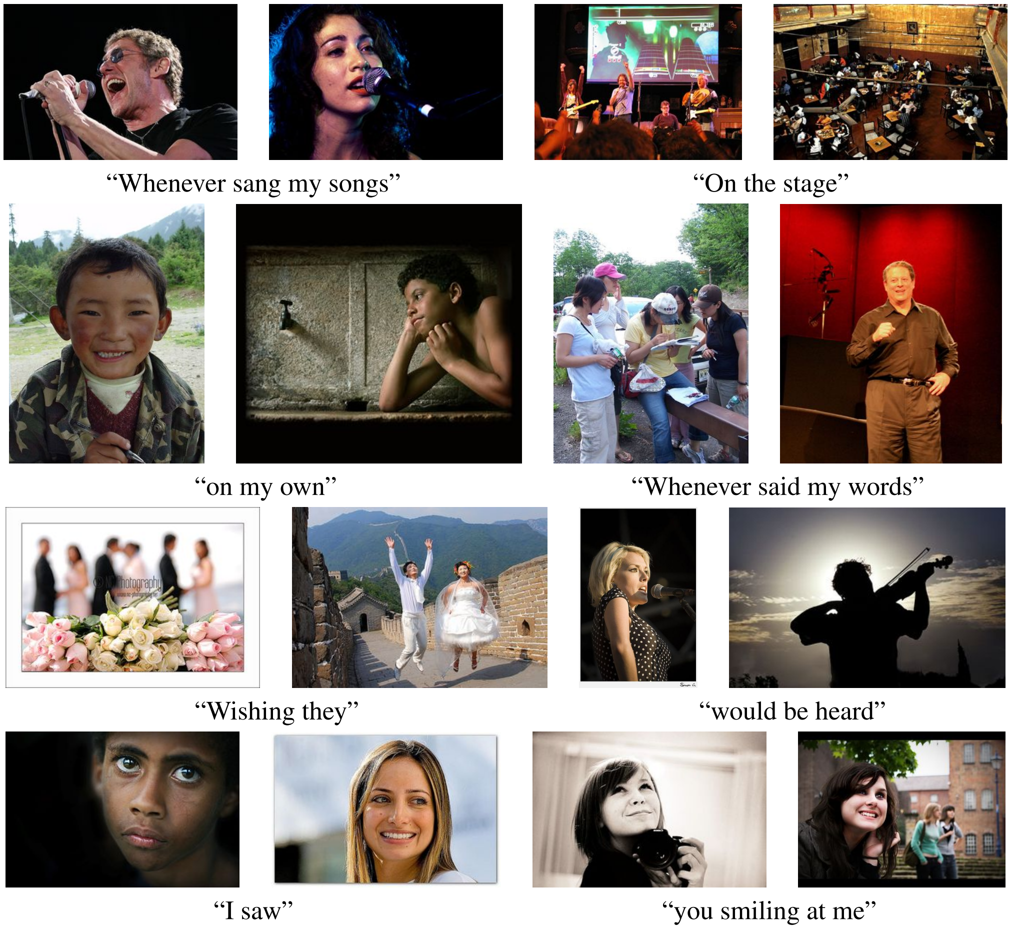
[6] X.-S. Hua, L. Lu, and H.-J. Zhang. P-karaoke: personalized karaoke system. In *MULTIMEDIA '04: Proceedings of the 12th Annual ACM International Conference on Multimedia*, pages 172–173, New York, NY, USA, 2004. ACM.

[7] X.-S. Hua, L. Lu, and H.-J. Zhang. Photo2videoła system for automatically converting photographic series into video. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(7):803–819, 2006.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[9] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, CUIDADO I.S.T., 2004.

[10] D. A. Shamma, B. Pardo, and K. J. Hammond. Musicstory: a personalized music video creator. In *MULTIMEDIA '05: Proceedings of the 13th Annual ACM International Conference on Multimedia*, pages 563–566, New York, NY, USA, 2005. ACM.



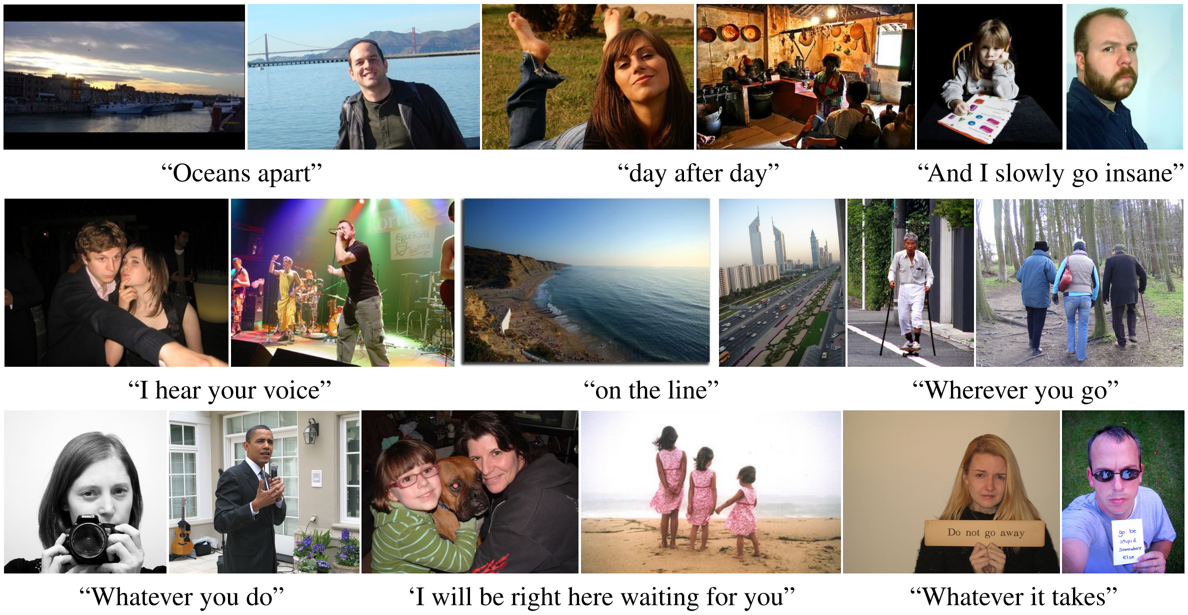
“Whenever sang my songs” “On the stage”

“on my own” “Whenever said my words”

“Wishing they” “would be heard”

“I saw” “you smiling at me”

Figure 1. A music slide show generation experiment using the song “Eyes On Me”.



“Oceans apart” “day after day” “And I slowly go insane”

“I hear your voice” “on the line” “Wherever you go”

“Whatever you do” ‘I will be right here waiting for you” “Whatever it takes”

Figure 2. A music slide show generation experiment for the song “Right Here Waiting”.