

Personalized Karaoke

Xian-Sheng HUA, Lie LU, Hong-Jiang ZHANG

Microsoft Research Asia
{xshua; llu; hjzhang}@microsoft.com

Abstract

In this paper, a personalized Karaoke system, *P-Karaoke*, is proposed. In the P-Karaoke system, personal home videos and photographs, which are automatically selected from users' multimedia database according to their content, users' preferences or music, are utilized as the background videos of the Karaoke. The selected video clips, photographs, and lyrics that obtained from *Lyric Service* or manually labeling, are aligned with the music rhythm, connecting by specific content-based transitions. Additionally, photographs are converted into *motion photo clips* by a *Photo2Video* technology, which automatically converts a photograph or photographic series into a video by simulating camera motions. Furthermore, a *Query by Humming* (QBH) system can be integrated into P-Karaoke easily, which enables users to find their desired music/songs efficiently.

1. Introduction

Karaoke is a form of entertainment originally developed in Japan, in which amateur performers sing pop songs to the accompaniment of pre-recorded music. It involves using a karaoke machine which enables performers sing live, usually by following the words on a video screen that in sync with the music. Typically, video tapes, discs or machine that support Karaoke are pre-recorded and thus cannot change the video content.

In this paper, a personalized Karaoke system, *P-Karaoke*, is proposed, which enables users to use their favorite home videos and/or photographs as the background video. Figure 1 illustrates the architecture of the P-Karaoke system, which mainly consists of four stages, including Multimedia Data Acquisition, Content Analysis, Content Selection, and Composition.

As Figure 1 shows, P-Karaoke is built based on MyVideos [1] and MyPhotos [2], which are personal video and photograph management systems, respectively. The videos and photographs in these two systems are the two main inputs of the P-Karaoke system. "My Music" is the user's music/song database, while "My Lyrics" may be downloaded from the *Lyric Service* (to be explained in Section 3 in detail) on the Internet or manually labeled according to user's music/song database. After obtaining the required multimedia data, the system analyzes the content of the videos, photographs and music, and the output of the analyses will be employed when composing personalized Karaoke video. When a user submit a music/song request by inputting/selecting the title of a specific song, or by a *Query By Humming* (QBH) [3]

system, background videos are automatically composed from the analyzed personal video and/or photograph database according to their content, user's preferences or music. In particular, selected photographs are converted into video clips by a *Photo2Video* technology, which automatically converts a photograph or photographic series into a video by simulating camera motions. Simultaneously, lyric of the song are superimposed on the video, while video shot boundaries, music beats and the characters or syllables of the lyric are well aligned.

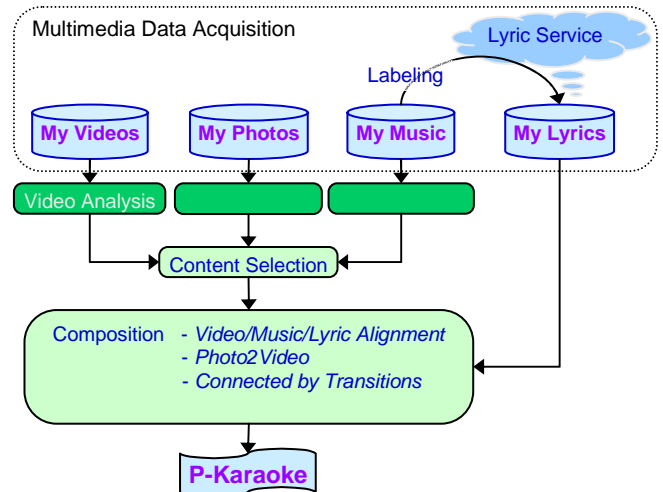


Figure 1. Architecture of P-Karaoke.

The rest of the paper is organized as follows. After presenting video, photograph and music analyses in Section 2, lyric acquisition and formatting are introduced in Section 3. Section 4 describes how to automate content selection, followed by video/music/lyric alignment and composition in Section 5. Conclusion and discussion are presented in Section 6.

2. Content Analysis

In this section, we will present how P-Karaoke analyzes the content of personal home videos, photographs and music. Temporal structure information and a set of metadata are extracted from these multimedia data, which will be employed for composing the Karaoke background video.

2.1 Video Analysis

Content analysis for home videos consists of two components: temporal structure parsing and attention (importance) detection. Based on the analysis results, P-Karaoke selects appropriate or "important" video segments/clips to compose the background video for Karaoke.

2.1.1 Temporal Structure Parsing

Videos are broken into shots, which are subsequently grouped into scenes and simultaneously subdivided into sub-shots. There are numerous shot detection algorithms reported in literatures and TREC VID [4]. In our system, we use an algorithm similar to the one proposed in [5]. For raw home videos, most of the shot boundaries are simple cuts, which are much easier to be correctly detected in comparison with professionally edited videos. Once transitions are detected, video temporal structure is further analyzed using by the following two approaches.

One approach divides the shots into smaller segments, namely, sub-shots, whose lengths are in a certain range (defined in Section 4). This is accomplished by detecting the maximum of the frame difference curve (FDC), as shown in Figure 3. A shot is cut into two sub-shots at the maximum peak, if the peak is separated from the shot boundaries by at least the minimum length of a sub-shot. Then the above process is repeated until the lengths of all sub-shots are smaller than the maximum sub-shot length.



Figure 2. Sub-shot boundary detection by finding local maximum of frame difference curve (three boundaries are found for this shot).

The other approach is to merge shots into groups of shots, i.e., scenes. There are many scene grouping methods presented in the literature [6][7]. In this paper, a hierarchical method that merges the most similar adjacent scenes/shots step-by-step into bigger ones is employed. The similarity measure is the intersection of quantized color histogram in HSV space [7]. The stop condition can be determined either by similarity threshold or the final scene numbers. We may also build higher level structure on scene, i.e., time, which is based on the time-code or timestamp [8] of the shots. In this level, shots/scenes that shoot in the same period are merged into one group.

2.1.2 Attention Detection

Generally, selecting appropriate or “important” video segments, or video summarization requires semantic understanding of the video content. Unfortunately, current computer vision and artificial intelligence technologies cannot accomplish it for unstructured home videos. However, if the objective is creating a compelling background video for Karaoke, it may not be necessary to understand the semantic content completely. Alternatively, we need only determine those parts of the video more “important” or “attractive” than the others. Assuming that the most “important” video segments are those most likely to hold a viewer’s interest, the task becomes how to find and model the elements, such as object motion, camera motion, specific objects/faces, static attention regions, audio and language, that are most likely to attract a viewer’s attention. This is the main idea of the work proposed by Ma et al.[9]. In our system, video segment selection is also based on this idea.

Based on attention detection, an *attention curve* is produced by calculating the attention/importance index of each video frame. Importance index for each sub-shot is obtained by averaging the attention indices of all video frames within this sub-shot.

2.2 Photograph Analysis

Photograph analysis consists of three components: quality filtering, grouping and focus detection. It is necessary to mention here that the background video of P-Karaoke could be videos from video database only, or photographs from photo database only, or a combination of them. Photo grouping is employed when using photographs only, while if we use both videos and photographs, each photograph is regarded as a video shot (which contain only one sub-shot, i.e., the shot itself), and then use video scene grouping to form groups. In that case, photo importance is the entropy of the quantized HSV color histogram.

2.2.1 Quality Filtering

Since most of the photographs are taken by unprofessional home users, there are frequently many low quality photographs in them which may be in the following cases,

- *Under or over exposed images*, e.g., the photographs that are taken when the exposal parameters are not well set. It can be detected by check whether the average brightness of the photograph is too low or too high.
- *Homogenous images*, e.g., floor, wall. They can be detected by checking whether the color entropy is too low. These photographs always have no salient object which user may have interest in.
- *Blurred images*. They are detected by the method in [10].

It is possible that some of these kinds of photographs could be enhanced or improved by image processing technologies, but this issue is not discussed in our paper. In the following sections, all processing are employed on the filtered photograph set.

2.2.2 Photograph Grouping and Selecting

A three-layer structure is used to group the photographs, namely, day, scene, and GoS (*Group of very Similar photographs*). The top layer, i.e., day, contains all photographs taken on a certain date, which can be obtained from the metadata of digital photographs or OCR results from analog photographs that have date stamps [11]. If none of these two kinds of information can be obtained, the date on file created is used. The middle layer, scene, represents a group of photographs that may be taken at the same place (scene). And the lowest layer, GoS, is a group of pictures which are very similar.

The top two layers, day and scene, will be used to determine transition types and support editing styles, as to be explained later. The lowest layer, GoS, is used for filtering out very similar photographs since photographers often take several photographs for the same or nearly the same object or scene. It will be boring if all of them appear in Karaoke, especially they are showed one by one.

In our system, photographs are firstly grouped into top-layer ‘day’ based on the date information. Then, a

hierarchical clustering algorithm similar to the approach in [12] with different thresholds is employed to group the lower two layers.

2.2.3 Focus Detection

Focus detection is the preparation step for Photo2Video, which will be described with more detail in Section 5. Focuses are the target areas in the photographs that the simulated camera will pan from/to, or zoom in/out. It is assumed in this system that the focuses of the simulated camera are those areas in the photographs that most likely attract viewers' attention. Typically human faces are more attractive than other objects, so firstly a face detector similar to the one in [13] is applied to capture dominant faces in the photographs. Faces are detected by the method proposed in [13]. In our system, we only count faces that not smaller than 100×100 pixels.

Other than faces, Li et al [14] defined a saliency-based visual attention model for static scene analysis. We adopt this approach to detection attended areas in the photographs. The saliency map obtained by this method is binarized in an adaptive manner to get separate attention areas/spots. Attention areas that have overlap with faces are removed.

Faces and attention areas with high confidence are taken as the attention focuses of the photographs. Users may also assign or modify the detected focus areas for a photograph.

2.3 Music Analysis

In order to align video shot (including photograph) boundaries with music beat, i.e., make the video transition happened at the beat positions of the incidental music, we segment the music into several music sub-clips, whose boundary is at the beat position. Each video shot is shown in one music sub-clip. It not only ensures that video shot transition is happed at the beat position, but also sets the duration of the video shot.

Instead of exact beat detection [15], in our real implementation, we only detected onset sequence [16], in case that beat information is not obvious at some part of the song. The strongest onset in a window is supposed as a beat. This is reasonable because there are several beat positions in a window (for example, such as 3s); thus, the most possible position of a beat is the position of the strongest onset.

To give a more comfortable perception, the music sub-clip should not be too short or too long. From our user study, the tolerable length of music sub-clip is about 3-5 seconds. Then, music sub-clip can be segmented by the following way: given the previous boundary, the next boundary is selected as the strongest onset in the window which is 3-5 seconds (the tolerable music sub-clip length) from the previous boundary.

The tolerable music sub-clip length can be set manually, it can also automatically set according to its tempo content, as our previous work [16] done. Thus, when the music tempo is fast, the length of music sub-clip is short; otherwise, the length of music sub-clip is long.

After music sub-clips are determined, video shot transition can be easily placed at the music beat position just by aligning the duration of a video shot and the corresponding music sub-clip.

3. Lyric Acquisition and Formatting

To finally generate a complete Karaoke, we should have the corresponding lyrics and align it with the selected song. However, it is very difficult, if not impossible, to automatically align lyrics with the song based on content analysis only. To avoid this issue, the time of each syllable in the lyrics have to be labeled. In our system, the *Lyric Service* is designed to provide labeled lyrics.

There are some available lyrics labeled by music fans on the Internet. However, most of them are designed as a plug-in of mp3 player and only labeled the start time and duration of each sentence. It is for lyric showing when listening mp3, but it is not accurate enough for Karaoke usage which requires "syllable-by-syllable rendering".

There are also many diverse formats on the internet, such as the lyrics labeled by sentence mentioned above. Traditional Karaoke machine uses the ST3 and KAR format which combine the lyrics with the midi music, and take the lyrics as one of midi channel.

In order to provide a more flexible and incorporate more information, we separate the lyric with the songs and defined a new lyrics format using XML, which can also easily converted from other formats.

Figure 3 illustrates the format of an excerpt of a lyric file used currently, which comprises most of key items, except for some general information (metadata) about the song, such as artist, album, year, composer and so on.

```

<Lyric>
  <Group type="solo" name="singer1">
    <Sentence start="" stop="">
      <syllable start="" stop="" value="" />
      .....
    </Sentence>
    <Sentence start="" stop="">
      .....
    </Sentence>
  </Group>
  <Group type="solo" name="singer2">
    .....
  </Group>
  <Group type="chorus" name="singer1, singer 2">
    .....
  </Group>
</Lyric>

```

Figure 3. An excerpt of a lyric file.

4. Content Selection

As aforementioned, the background video could be video segments from MyVideos only, photographs from MyPhotos only, or a combination of video segments and photographs. In Section 2.2, we have discussed content selection in the case of using photographs only. In this section, we only focused on the other two cases. Actually, if we use both videos and photographs, each photograph can be regarded as a shot (a sub-shot as well at the same

time), and photograph groups can be regarded as “scenes”. Thus, this case can be treated the same as the case we use videos only. Therefore, below we only discuss video content selection.

To ensure that the selected video clips and/or photograph are of satisfactory quality, a set of rules derived from studying professional video editing are followed. Firstly, using a long unedited video as Karaoke background is boring, as generally there are lots of redundant content and low quality segments in typical raw home videos. An effective way to compose a compelling video is to present a video that is as compact as possible, yet preserves the most critical features required to tell a story. In other words, the editing process should select segments with relatively higher “importance” or “excitement” value from the raw video. Secondly, for a given video, the most “important” segments according to an importance measure could concentrate in one or in a few parts of the time line of the original video. This may obscure the storyline in the edited video. In other words, the distribution of the selected highlight video should be as uniform along the time line as possible so as to preserve the original storyline.

These above two rules deal with how to select suitable segments that are representative of the original video in content and of high visual quality. In fact, content selection can be formulated as an optimization problem. The next issue is how to design the objective function. According to the two rules mentioned above, there are two computable objectives as listed below:

- (1) Selecting “important” sub-shots.
- (2) Selected sub-shots should be nearly uniformly distributed.

Of course, other computable objectives that may assist content selecting can be adopted here too.

The first objective is achieved by examining the average attention index of each sub-shot as described in Section 2.1. For the second objective, *Distribution Uniformity* is represented by normalized entropy of the selected shots distributed along the timeline of the raw home videos.

5. Video Composition

In this section, we will firstly introduce the scheme to align shot boundaries, music beats and lyric, then present how to convert photograph or photographic series into videos. Next, the methods for connecting shots with specific transitions and applying transformation effects on shots are introduced. And last, style supporting is presented.

5.1 Alignment

The first issue is to align shot transitions with music beats. To make the Karaoke background video more expressive and attractive, shot transitions had better occur exactly at music beats, i.e., at the boundaries between the music sub-clips. This alignment requirement is met by the following alignment strategy.

- (1) The minimum duration of sub-shots is made greater than maximum duration of music sub-clips. For example, we may set music sub-clip duration in the range between 3 and 5 seconds, while sub-shots duration in 5 to 7 seconds.
- (2) Since sub-shot durations are generally greater than music sub-clips, we can shorten the sub-shots to match their duration to that of the corresponding music sub-clips.

Another alignment issue is syllable-by syllable lyric rendering. As the time of each syllable has been clearly indicated in the lyric file, it is quite easy to accomplish this objective.

5.2 Photo2Video

Photo2Video is a technology developed to automatically convert photographs into video by simulating temporal variation of people’s study of photographic images using simulated camera motions [16].

When we view a photograph, we often look at it with more attention to specific objects or areas of interest after our initial glance at the overall image. In other words, viewing photographs is a temporal process which brings enjoyment from inciting memory or from rediscovery. This is well evidenced by noticing how many documentary movies and video programs often present a motion story based purely on still photographs by applying well designed camera operations. That is, a single photograph may be converted into a *motion photograph clip* by simulating temporal variation of viewer’s attention using simulated camera motions. For example, zooming simulates the viewer looking into the details of a certain area of an image, while panning simulates scanning through several important areas of the photograph. Furthermore, a slide show created from a series of photographs is often used to tell a story or chronicle an event. Connecting the motion photograph clips following certain editing rules forms a slide show in this style, a video which is much more compelling than the original images.

Focuses detected in Section 2.2 are areas in a photograph that most likely will attract a viewer’s attention or focus. These areas are used to determine the simulated camera motions to be applied to the image, based on a similar technology as Microsoft PhotoStroy [18]. One motion photo clip is regarded as one shot (one sub-shot as well).

5.3 Transitions and Effects

Twenty-seven transformation effects provided by Microsoft Movie Maker 2 [19] are used in our system, including grayscale, blurring, fading in/out, rotation, thresholding, sepia tone, etc. Sixty transition effects provided by Microsoft DirectX and Movie Maker are also employed in our system, including cross fade, checkerboard, circle, wipe, slide, etc.

The transformation and transition effects can be selected randomly in a specific effect set, or determined by the styles, as to be explained in detail later.

Simple rules for transition selection are also employed. For example, we use “cross fade” for the sub-shots/photographs in the same scene/group/day, use others randomly selected transitions as a new day/group/day comes out.

5.4 Style Support

As an extension of our system, we support different styles according to users’ preference. We may define as many styles as we want. Here we just use three style examples, namely, music video, day by day, and old movie, to show how we support different styles. For different showing style, different transformation effects and transition effects are selected. They are obtained from users’ suggestions, although they seem a little arbitrary. We can further improve them according to more users’ feedbacks.

5.4.1 Music Video

In this style, firstly we segment the music according to the tempo of the music. That is to say, if the music is fast, the music sub-clip will be shorter, and vice versa. Then video segments/photographs and music are fused together to get the background video by the following rules for transformation effects and transition effects.

- *Transformation Effects.* Apply randomly selected effects from the entirely effect set on half of randomly selected sub-shots.
- *Transition Effects.* Apply randomly selected transitions from the entirely transition set except “cross fade” between half of randomly selected sub-shots changes. Others, we use “cross fade”.

5.4.2 Day by Day

In this style, when a new day comes out, we add a man-made photograph before the first sub-shot of the day to illustrate the creating date of the sub-shots coming next. The rules for transformation effects and transitions are defined below.

- *Transformation Effects.* A “fade in” effect is added on the first sub-shots of each day, while a “fade out” effect is added on the last sub-shots of each day. Others, we do not use effects.
- *Transition Effects.* Use “fade” between sub-shots those are in the same day, and use randomly selected effects when a new day begins.

5.4.3 Old Movie

Sepia tone or grayscale effect is applied on all sub-shots, while only “fade right” transitions are used between them.

6. Conclusion and Discussion

In this paper, a personalized Karaoke system, *P-Karaoke*, is proposed, which enables users to use their favorite home videos and/or photographs as the background video. In our system, photographs are converted into *motion photo clips* by simulating camera motions; video shots with higher importance are selected as Karaoke video content; and lyrics are obtained from *Lyric Service* or manually labeling. These three kinds of data are finally aligned with the music rhythm, connecting by specific content-based transitions, which are determined based on the content of the

corresponding clips or photographs. Furthermore, a *Query by Humming* (QBH) system can be integrated into *P-Karaoke* easily, which enables users to find their desired music/songs efficiently. The results are interesting and compelling.

There are a number of possible improvements for this system. For example, face detection and tracking may assist to create music videos that have a “central character” or “leading actor”. In addition, semantic classification of video shots, such as indoor vs. outdoor, cityscape vs. landscape, beach, sun rising/falling, moon night, etc., may also facilitate semantic content selection.

References

- [1] Y. Wang, P. Zhao, D. Zhang, M. Li, and H.J. Zhang, “MyVideos – A system for home video management”, *ACM Multimedia 2002*.
- [2] Y. Sun, H.J. Zhang, L. Zhang, and M. Li, “MyPhotos – A system for home photo management and processing”, *ACM Multimedia 2002*.
- [3] L. Lu, H. You, H.J. Zhang, “A New Approach to Query by Humming in Music Retrieval”, *ICME 2001*, pp776-779. 2001.
- [4] TREC Video Retrieval Evaluation. <http://www-nlpir.nist.gov/projects/trecvid/>.
- [5] H.J. Zhang, A. Kankanhalli, and S. W. Smoliar, “Automatic Partitioning of Full-Motion Video,” *Multimedia Systems*, 1, 10-2, 1993.
- [6] J.R. Kender, and B. L. Yeo, “Video Scene Segmentation via Continuous Video Coherence,” *Proc IEEE Intl Conf on Computer Vision and Pattern Recognition 1998*, 367-373.
- [7] T. Lin, and H.J. Zhang, “Video Scene Extraction by Force Competition,” *ICME 2001*.
- [8] P. Yin, X.S. Hua, and H.J. Zhang, “Automatic Time Stamp Extraction System for Home Videos,” *ISCAS 2002*.
- [9] Y.F. Ma, L. Lu, H.J. Zhang, and M.J. Li, “A User Attention Model for Video Summarization,” *ACM MM 2002*, 533-542.
- [10] T. M. Cannon, “Blind Deconvolution of Spatially Invariant Blurs with Phase,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, February 1976.
- [11] X.R. Chen, “Photo Time Stamp Recognition,” *Microsoft Research Technical Report*, 2001.
- [12] J. Platt, “AutoAlbum: Clustering Digital Photographs using Probabilistic Model Merging”, *IEEE Workshop on Content-Based Access to Image and Video Libraries 2000*.
- [13] S.Z. Li, et al, “Statistical Learning of Multi-View Face,” Detection. *Proceeding of ECCV 2002*.
- [14] Y. Li, Y.F. Ma, and H.J. Zhang, “Salient region detection and tracking in video,” *ICME 2003*.
- [15] Eric D. Scheirer, “Tempo and beat analysis of acoustic musical signals,” *Journal of the Acoustical Society of America*, 103(1):588-601, 1998.
- [16] X.S. Hua, L. Lu and H.J. Zhang, “Photo2Video”, *ACM Multimedia 2003*.
- [17] X.S. Hua, L. Lu, and H.J. Zhang, “Content-Based Photo Slide Show with Incidental Music,” *ISCAS2003*.
- [18] Microsoft Plus! Digital Media Edition. <http://www.microsoft.com/windows/plus/dme/>.
- [19] Microsoft, Movie Maker 2, <http://www.microsoft.com/>.