

P-Karaoke: Personalized Karaoke System

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

Microsoft Research Asia

No.49 Zhichun Road, Haidian District, Beijing 100080, China

{xshua, llu, hjzhang}@microsoft.com

ABSTRACT

In this demonstration, 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, music and lyrics are well aligned to compose a Karaoke video, connecting by specific content-based transitions.

Categories and Subject Descriptors

H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems — video; I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—video analysis.

General Terms

Algorithms, Experimentation.

Keywords

Video content analysis, music content analysis, video editing.

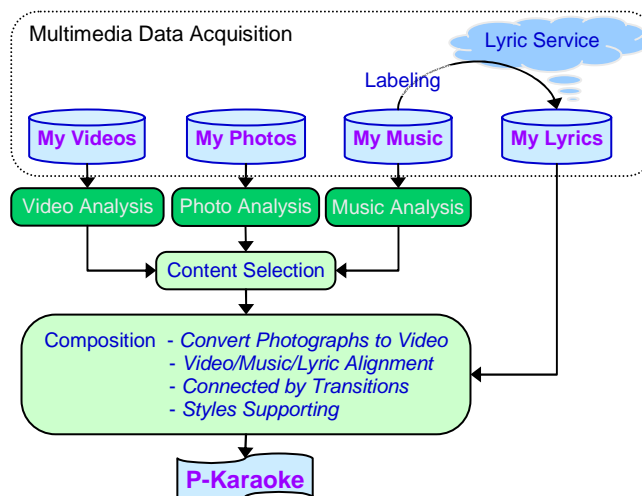


Figure 1. Architecture of P-Karaoke System.

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, following the words on a video screen. Most common Karaoke machines are audio mixers with microphone input built-in with CD, VCD, Laser Disc, or DVD players which play with special media that encode the original song in one audio track and music in another track. The video media also enable the display of the lyrics graphically on screen in sync with the music. Typically video tapes, discs or machine that support Karaoke are prerecorded and thus cannot change the video content.

In this demonstration, we present a personalized Karaoke system, *P-Karaoke*, which enables users to use their favorite home videos and/or photographs as the background video, accompanying with personal favorite songs.

2. P-KARAOKE SYSTEM DESIGN

Figure 1 illustrates the architecture of the P-Karaoke system, which mainly consists of four modules, including Multimedia Data Acquisition, Content Analysis, Content Selection, and Composition. As Figure 1 shows, P-Karaoke is built on MyVideos [1] and MyPhotos [2], which are personal video and photograph management systems, respectively. "My Music" is the user's music/song database, while "My Lyrics" may be downloaded from the *Lyric Service* on the Internet or manually labeled. After obtaining the required multimedia data, the system analyzes the content of the videos, photographs and music. The highlights of

the videos and/or photograph are selected based on their content, user's preferences, and aligned with the music and lyrics. Furthermore, each photograph is converted to a *motion photograph clip* by simulating camera motion, and the boundaries between two adjacent shots are connected by specific transitions according to their contents.

3. SYSTEM IMPLEMENTATION

In this section, we will present in detail the implementations of the latter three modules in the P-karaoke system.

3.1 Content Analysis

3.1.1 Video Analysis

Content analysis for home videos consists of two components: temporal structure parsing, and attention (importance) detection.

In temporal structure parsing, videos are broken into shots, which are subsequently grouped into scenes based on the similarity in HSV space, and simultaneously, subdivided into sub-shots based on *frame difference curve* (FDC)

To find those parts of the video more "important" or "attractive" than the others, an importance index of each video sub-shot are calculated based on *Attention Model* [3], which involves in object motion, camera motion, specific objects/faces, static attention regions, audio and language.

3.1.2 Photograph Analysis

Photograph analysis consists of three components: quality filtering, grouping and focus detection.

Since most of the photographs are taken by unprofessional home users, there are frequently many low quality photographs in them

such as under or over exposed images, homogenous images and blurred images. They are removed before further processing.

After filtering, photographs are grouped into a three-layer structure namely, day, scene, and GoS (*Group of very Similar photographs*), which provide the information to determine the transition types and editing styles in final composition. Very similar photographs are filtered out since it will be boring if all of them appear in the Karaoke video, especially they are showed one by one.

In our system, each single photograph is converted into a *motion photograph clip* by simulating temporal variation of viewer's attention using simulated camera motions, such as panning and zooming. In order to decide the target areas in the photographs that the simulated camera will pan from/to, or zoom in/out, human faces detector and static visual attention model [3] are employed to detect the focuses or attended areas in the photographs.

3.1.3 Music Analysis and Lyric Service

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

The tolerable sub-music length can also automatically set according to its tempo content [4]. Thus, when the music tempo is fast, the length of sub-music is short, and vice versa.

The corresponding lyric file of a song is provided by the *Lyric Service* on the Internet or manually labeling. In the lyric file, the time signature of each syllable is labeled. This information enables the P-Karaoke system provides "syllable-by-syllable" lyric rendering in sync with the music.

3.2 Content Selection

Based on the analysis results, P-Karaoke selects appropriate or "important" video segments and/or motion photograph clips to compose the background video for the Karaoke. To ensure that the selected video clips and/or photograph are of satisfactory quality, a set of rules derived from studying professional video editing is employed. First, 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. Only the segments with relatively higher "importance" or "excitement" value are selected from the raw videos. Second, for a given video, the most "important" segments according to an importance measure might 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. Therefore, in our system, the distribution of the selected highlight video segments along the time line is kept as uniform as possible so as to preserve the original storyline.

Content selection under the above two constraints is formulated as an optimization problem and solved by a GA algorithm.

3.3 Video Composition

One alignment issue is to align shot boundaries and music beats. To make the Karaoke background video more expressive and attractive, shot transitions occur exactly at music beats, i.e., at the boundaries between the music sub-clips. Another alignment issue is syllable-by-syllable lyric rendering. As the time signature of

each syllable has been clearly indicated in the lyric file, it is quite easy to accomplish this objective.

Twenty-seven transformation effects provided by Microsoft Movie Maker 2 [5] are used in our system, including grayscale, blurring, fading in/out, rotation, 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 or determined by the supported styles,

As an extension of our system, we support different editing styles according to users' preference. Three example styles are *music video*, *day by day*, and *old movie*. For different showing style, different transformation effects and transition effects are employed.

4. PROTOTYPE

Figure 2 shows a screenshot of the P-Karaoke prototype. In P-Karaoke, users can browse, manage, analyze and select their videos, photograph, music and lyrics. After users indicate source videos and/or photographs, and select desired music/song, the system automatically compose a background video for the music, while lyric is superimposed on the video frames and rendered syllable-by-syllable in sync with the music. The users are also provided with a collection of editing styles, each is implemented by a set of specific editing rules. Composed video can also be saved as a video file or script file for future access.

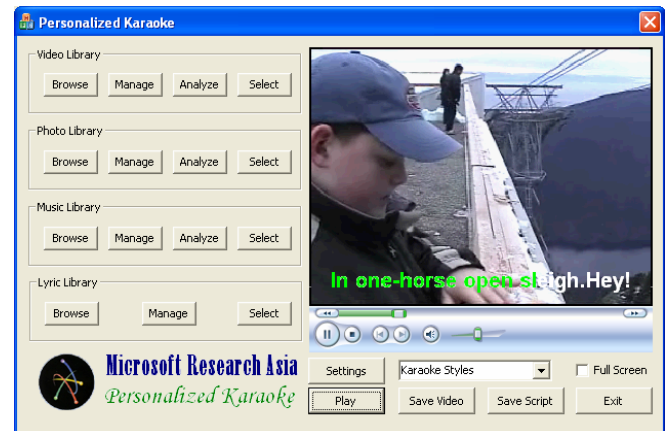


Figure 2. Prototype of P-Karaoke System.

In this system, the content analysis procedure can be processed in about 1/6 of real time on a Dell P4 1GHz computer for MPEG1 format raw videos. Content selection and alignment will cost less than 5 seconds for a normal music (say, 5 minutes). Video composing and rendering are processed in real time.

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] Y.F. Ma, L. Lu, H.J. Zhang, and M.J. Li. A User Attention Model for Video Summarization. *ACM MM 2002*, 533-542.
- [4] X.S. Hua, L. Lu, and H.J. Zhang. Content-Based Photo Slide Show with Incidental Music. *ISCAS2003*.
- [5] Microsoft. Movie Maker 2. <http://www.microsoft.com/>.