

# CONTENT BASED PHOTOGRAPH SLIDE SHOW WITH INCIDENTAL MUSIC

Xian-Sheng Hua, Lie Lu, Hong-Jiang Zhang

Microsoft Research Asia

No. 49 Zhichun Road, Beijing 100080, P.R. China

{i-xshua, llu, hjzhang}@microsoft.com

## ABSTRACT

In this paper, we proposed a new scheme to generate photograph slide show with incidental music based on the content of the photographs and music. Unlike existing photograph slide show software, which show the photographs one by one (perhaps with some fancy transition effects) with music in the background, our algorithm utilizes content analysis of the photographs and the incidental music. Firstly, low-quality photographs, such as blurred or under exposed, are filtered out by a photograph quality detector; and the rest of the photographs are grouped into a three-layer construction, namely, day, scene, and group of similar photographs (GoS). Incidental music is segmented based on its beats, and then aligned with photographs, which makes the photograph transitions occur at the beat position of the music. Various transition effects and transformation effects are applied based on the structure and content of the photographs. According to different users' preferences, several editing styles are also supported in our algorithm. The experimental results are very impressive according to the user study.

## 1. INTRODUCTION

Nowadays, as photograph becomes a major data type in personal computers, how to show or browse these photographs more efficiently and more enjoyably is an important issue. Many commercial software or shareware can generate photograph slide show with music [1][2][3][4][5]. However, all of them simply put the photographs and music together without any analysis on their content. Photographs are simply showed one by one and music is played in the background. It makes the showing results a little flat, even if many fancy transitions are added between photographs.

In ACDSee [1], each image being shown for some predetermined time before going on to the next without any transformation effect or transition, and the music is only used as the background. Transitions are employed in PhotoJam [2], but no transformation effects are applied on the photographs. Music content is also not taken into consideration in PhotoJam. Many other software programs, such as Digital Photo Slide Show [3], My Photo Slide Show [4], Ulead DVD PictureShow [5], also do not analyze the structure and content of the photographs and music.

If we extract some content information from photographs and music, then re-organize the photographs and match photographs and music based on their content, the photograph show will be more attractive and enjoyable. Our work presented in this paper is an attempt in this direction.

To achieve more reasonable and pleasing results, the following principles that suggested by many users should be taken into consideration.

- (1) Low quality photograph should not be selected. Here low quality photographs include those blurred or under/over exposed.
- (2) For those very similar photographs, only one of them is needed to be shown in a showing session.
- (3) Music should be suitably matched with photograph show. For example, photograph transitions should occur at the music beat, which can make the photograph showing more pleasing.

Based on these principles, content-based analysis on photograph and incidental music is necessary. Figure 1 illustrates the processing flow of our system. Firstly, photographs are grouped into a three-layer structure, which is named as, from high to low, day, scene and GoS (Group of very Similar photographs) after quality filtering, while incidental music is segmented into sub-music clips after beat detection. Then we select one photograph from each GoS as the photograph set we used for generating slide show. Each photograph is connected by specific transitions, which occur at the strong beats of the music, based on the structure and content of the photographs. Transformation effects are also applied on each photograph, such as grayscale, blurring, fading in/out, rotation, thresholding, etc. According to different users' preference, several editing styles are also supported in our algorithm.

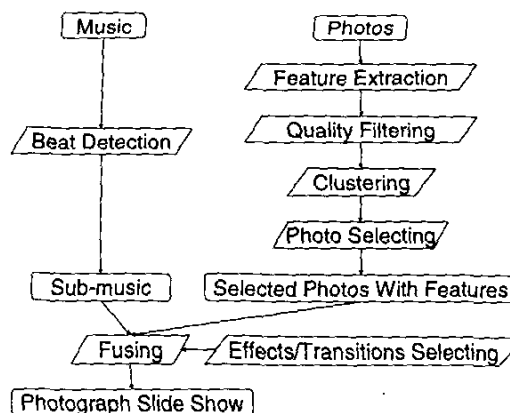


Figure 1. Block diagram of our algorithm – Photograph Slide Show with Incidental Music.

The rest of this paper is organized as follows. In Section 2 we present how music is analyzed. In Section 3 we then describe how we analyze photographs. Based on the analysis results, schemes that fuse photographs and music together to get the final output are presented in Section 4. Experimental evaluations are presented in Section 5. Finally we summarize our system in Section 6.

## 2. MUSIC ANALYSIS

In order to make the photograph transition happened at the beat positions of the incidental music, it is necessary to detect the beat series in the music. To align the photograph transition and music beats more conveniently, we segment the music into several sub-music clips based on the detected beat. The boundary of the sub-music is at the beat position. These sub-music clips are used as the basis of photograph slide show, i.e., one photograph is shown in one sub-music clips, and then the next photograph is shown in the next sub-music. It not only ensures that photograph transition is happened at the beat position, but also sets the duration of showing one photograph.

The length of the sub-music is randomly set based on the beat positions in a tolerable range. In general, when the music tempo is fast, the length of sub-music should be short; otherwise, the length of sub-music could be long. To achieve this objective, an adaptive sub-music length decision scheme is also proposed.

### 2.1 Sub-Music Segmentation

Instead of beat detection using complex algorithm [6], a much simpler approach is used in our scheme. In our real implementation, we did not detect exact beat series, we only detected onset, since beat information is not obvious, especially for those no-drum light music, which is always selected to accompany the photograph show. The strongest onset in a window is supposed as a beat. This is reasonable because there are many 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 detect onset sequence, an octave-scale filter-bank is applied to divide the music frequency domain into several sub-bands:

$$\left[0, \frac{\omega_0}{2^n}\right), \left[\frac{\omega_0}{2^n}, \frac{\omega_0}{2^{n-1}}\right), \dots, \left[\frac{\omega_0}{2^2}, \frac{\omega_0}{2^1}\right) \quad (1)$$

wherein  $\omega_0$  refers to the sampling rate and  $n$  is the number of sub-band filters. In real implementation, 7 sub-bands are employed.

After amplitude envelope is extracted from each sub-band, a Canny operator is applied to estimate its difference curve. Then, the sum of the difference curves in each sub-band is used to extract onset sequence. Each peak in the sum curve is considered as an onset, and peak value as the onset strength.

To give a more comfortable perception, the sub-music should not be too short or too long. From our user study, the tolerable length of sub-music clip is about 2-6 seconds. Then, sub-music 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 2-6 seconds (the tolerable sub-music length) from the previous boundary.

In above sub-music segmentation, the sub-music boundary is determined by the strongest onset in a constant window, although the sub-music length is a little randomly. Thus, the length of sub-music is not relevant to its tempo content. In fact, it is more reasonable if length of sub-music can be automatically matched with its tempo content. For example, when the music tempo is fast, the length of sub-music could be short; otherwise, the length of sub-music could be long. To achieve this objective, we can adjust the range of sub-music length.

Suppose the tempo of the music is constant, then the range of sub-music length can be experientially set as,

$$\begin{aligned} SubMusicLen_{min} &= \min\{\max\{2 * Tempo, 2\}, 4\} \\ SubMusicLen_{max} &= SubMusicLen_{min} + 2 \end{aligned} \quad (2)$$

where  $SubMusicLen_{min}$  and the  $SubMusicLen_{max}$  is the lower bound and upper bound of the sub-music length. Music tempo can be obtained from the auto-correlation of the onset sequence. In general, the tempo of music is about 1-2 second. It coarsely represents how fast/slow of this music clip. Thus, for a music piece with 60bpm, the tempo is 1 second; and its sub-music length is in 2 - 4s.

After sub-music clips are determined, photograph transition can be easily placed at the music beat position just by aligning the photograph show duration and the corresponding sub-music length.

Since the beats or strong onsets of a particular music are fixed, the music can be analyzed in advanced and recorded with its beat positions. When a photograph slide show is to be generated, a suitable music (e.g., tempo or number of sub-music clips) can be chosen from the music database.

## 3. PHOTOGRAPH ANALYSIS

Not all photographs will be shown in our implementation. The low quality photographs will be filtered out firstly. The left photographs will be grouped into a three-layer structure, according to the color similarity and time similarity.

### 3.1 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 photographs*. It can be detected by the method in [7].

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

### 3.2 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 [8]. 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 the slide show, especially they are showed one by one.

### 3.2.1 Photograph Grouping

In our system, photographs are firstly grouped into top-layer 'day' based on the date information. Then, a hierarchical clustering algorithm with different thresholds is used to group the lower two layers.

These two layers' grouping could be also time-constrained or not. For time constrained grouping, each group contains photographs in a certain period of time. There is no time overlap between different groups. In [9], the authors use the time and order of photograph creation to assist in clustering (photograph groups consist of temporally contiguous photographs). A content-based clustering algorithm using best-first probabilistic model merging, which is fast and yields clusters that are often semantically meaningful, is also applied for clustering in [9]. We use a method similar to the one proposed in [9] to accomplish this task.

If no time constraint is needed, photograph can be grouped only according to their content similarity. Here we use a simple hierarchical clustering method for grouping. Color histogram intersection is used as similarity measure of two photographs or two photograph clusters.

We use HSV color cone [10] that is quantized by a 3D Cartesian coordinate system with 20 values for X and Y, 10 values for Z (the lightness), respectively. Let  $h = \{h_i, 0 \leq i < N\}$  and  $g = \{g_i, 0 \leq i < N\}$  denote the quantized HSV color histogram, the intersection of them is defined as

$$d(h, g) = \frac{\sum_{i=0}^{N-1} \min(h_i, g_i)}{\sum_{i=0}^{N-1} h_i} \quad (3)$$

where  $N$  is dimension of the histogram and equal to 4000 here. For photograph clusters, average histogram is used for computing similarity.

Two thresholds,  $T_{scene}$  and  $T_{GoS}$  are set to determine whether two photographs are in the same scene or GoS, respectively. If the similarity of two photographs or photograph clusters is larger than  $T_{scene}$  or  $T_{GoS}$ , they are merged into one cluster.

### 3.2.2 Photograph Selecting

As mentioned above, for very similar photographs, it will be boring if they appear adjacently. To deal with this problem, we only select one photograph from the one GoS in the final output. In the time-constrained classification, we select the last photograph from the GoS since the last photograph is more likely the most satisfied one when people taking photographs for the same object or scene. For others, we just randomly select one of them. Of

course, users may also indicate the photographs that must be included in the slide show.

## 4. SLIDE SHOW GENERATING

Photographs are connected by transitions at the beats of the music. Transformation effects are also applied on photographs to get more enjoyable output.

### 4.1 Transformations and Transitions

Twenty-seven transformation effects provided by the latest version of Microsoft Movie Maker [11] 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 used 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 Section 4.2.

Simple rules for transition selection are also employed. For example, we use "cross fade" for the photographs in the same scene, use others randomly selected transitions as a new day comes out.

### 4.2 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 photograph, 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.

#### 4.2.1 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 sub-music will be shorter, and vice versa. Then photographs and music are fused together to get the slide show 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 photographs.
- *Transition Effects.* Apply randomly selected transitions from the entirely transition set except "cross fade" between half of randomly selected photograph changes. Others, we use "cross fade".

#### 4.2.2 Day by Day

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

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

#### 4.2.3 Old Photograph

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

One thing need to be mentioned here is, when the number of the photograph is not equal to the number of sub-music, several disposal ways can be selected freely. If the number of photographs is less than the number of sub-music clips, we may fade out the music, or repeat the photographs. If the number of photographs (after quality filtering and photograph selecting) is more than the number of sub-music clips, we can repeat the music, or select some of the photographs.

### 5. EXPERIMENTS

We've applied our algorithms on various kinds of photographs and music, and got very impressive results.

Figure 2 illustrates the prototype of the system. A photograph folder and a music file are selected by pushing corresponding buttons. All the photographs in that folder are displayed in the left part of the program window. Before generating slide show, styles can be indicated. The results of the system can be viewed as a slide show or can be output as a video file. All the photographs which are selected in the final output are marked with a red box. Users can also indicate specific photographs that must appear in the slide show before generating. In Figure 2, we can see that only one photograph of the first two similar photographs is selected, and the tenth photograph is filtered out because it is too dark.

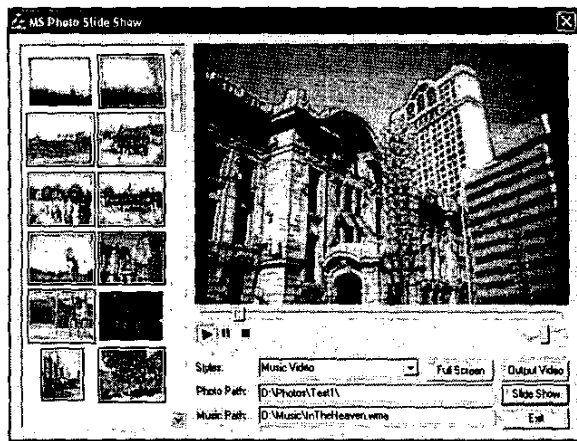


Figure 2. Prototype of Photograph Slide Show.

Ten users are invited to do the user study. We compare our results with ACDSee and PhotoJam. Three sets of photographs and music are used in the user study. Photographs in Test1, are scanned analog photographs of sight seeing, photographs in Test2 and

Test3 are digital photographs of outdoor events and indoor conference, respectively. So there are nine results in total. All users are required to give a satisfaction score to each result. The score of the first photograph slide show generated by ACDSee is fixed to 0.50 thus the users can take it as an example to giving scores for other results. Average satisfaction values are listed in Table 1, which shows our system have much higher satisfaction.

Table 1. Results of user study of our system.

System	Test1	Test2	Test3	Average
ACDSee	0.50	0.48	0.45	0.48
PhotoJam	0.73	0.69	0.62	0.68
Ours	0.85	0.88	0.86	0.86

### 6. SUMMARY

In this paper, we presented a novel scheme to generate an impressive photograph slide show aligned with the incidental music. The scheme fully utilizes the content analysis of the photographs and the incidental music. Photographs are reorganized with or without time constrain, and matched with incidental music appropriately. Various transition effects and transformation effects are applied based on the structure and content of the photographs. According to different user's preference, several editing styles are also supported in our algorithm. Experimental results showed our approach is very promising.

### 7. REFERENCES

- [1] ACDSee, <http://www.acdsystems.com/English/index.htm>.
- [2] PhotoJam, <http://www.shockwave.com/sw/content/photojam>.
- [3] Digital Photo Slide Show, <http://www.digitalphotoslideshow.com/>.
- [4] My Photo Slide Show, <http://www.copseystrain.com/myphotoslideshow/>.
- [5] Ulead DVD PictureShow, <http://www.ulead.com/dps/runme.htm>.
- [6] Eric D. Scheirer, “Tempo and beat analysis of acoustic musical signals,” *Journal of the Acoustical Society of America*, 103(1):588--601, 1998.
- [7] T. M. Cannon, “Blind Deconvolution of Spatially Invariant Blurs with Phase,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, February 1976.
- [8] X.R. Chen, “Photo Time Stamp Recognition,” *Microsoft Research Technical Report*, 2001.
- [9] J.Platt, “AutoAlbum:Clustering Digital Photographs using Probabilistic Model Merging”, *IEEE Workshop on Content-Based Access to Image and Video Libraries 2000*.
- [10] W. Y. Ma and H. J. Zhang, “Content-based image indexing and retrieval”, in *Handbook of Multimedia Computing*, Borko Furht, ed. CRC Press, 1998.
- [11] Microsoft, Movie Maker, <http://www.microsoft.com/>.