Tiling Slideshow

Jun-Cheng Chen[1], Wei-Ta Chu[1], Jin-Hau Kuo[1], Chung-Yi Weng[1], and Ja-Ling Wu[1,2]
[1]Department of Computer Science and Information Engineering
[2]Graduate Institute of Networking and Multimedia
National Taiwan University
{pullpull,wtchu,david,chunye,wjl}@cmlab.csie.ntu.edu.tw

ABSTRACT
This paper presents a new medium, called tiling slideshow, to display photos in a tile-like manner, coordinating with the pace of background music. In contrast to the conventional photo slideshow, multiple photos that have similar characteristics are well arranged and displayed at the same layout. Motivated by the concepts of technical writing, each displaying layout is composed of a larger topic photo and several small-size supportive photos. Based on this idea, the proposed tiling slideshow system consists of three major components: image clustering, music analyzer, and layout organizer. Given the limited displaying space, we consider the context and relationship between photos and model the layout organization as a constrainted optimization problem. Experiments on real consumer photograph collections show that the novel displaying method gives users more pleasant browsing experience than the methods that focus only on single photograph display.

Categories and Subject Descriptors
H.5.1 [**Multimedia Information Systems**]: animations. H.3.1 [**Content Analysis and Indexing**]: abstracting methods, indexing methods.

General Terms
Algorithms, design, experimentation.

Keywords
Slideshow, photo clustering, music analysis, and image content analysis.

1. INTRODUCTION
Digital camera has become an indispensable commodity for each family or individual in recent years. With the advance of the technology of digital storage, people can take pictures at will and have been more accustomed to record everything by photographs rather than text. Nevertheless, large amounts of photos without appropriate organization draw a potential problem in information access. People have to spend much time on browsing and often get lost in large volume of photo collections. Therefore, either organization or access issues pose urgent needs in advanced photos analysis and presentation techniques.

Studies on content-based image retrieval [1] provide an approach for efficiently indexing and accessing image collections. However, it's often the case that users have no initial examples for content-based searching, and the retrieval results are not presented in an organized manner. Digital photo album [2] or the so-called photo table of content [3] are therefore developed to facilitate efficient browsing or photo management. Commercial photo browsers, such as ACDSee [4] and Picasa [5], provide thumbnail functionalities to scale down photos so that users can browse multiple photos at a glance. Although the aforementioned tools/techniques provide some ways for managing and accessing image/photo collections, some critical problems that significantly impede users' browsing experience are worthy of further investigation:

(1) Popularity of photo capturing devices creates massive disordered photos, which stuff user's storage and make photo browsing and access tedious. Therefore, how to automatically organize photos by time, content, or topics is urgently demanded. Some studies have been conducted for photo clustering based on temporal context [6][7], while relatively few works have exploited clustered data for creating organized presentation.

(2) One of the most popular ways to present photos is the photo slideshow. It's provided in many photo management systems as an indispensable function. However, for large amounts of photos, sequentially browsing often takes much time and makes users weary. Some works have been conducted to improve browsing experience by cooperating with music [8], but the lengthy browsing time is still not addressed.

(3) Since most photos are taken by amateurs who are not familiar with photography, many photos are suffered from quality degradation, such as blur derived from hand shaking and underexposure (overexposure) caused by bad light or shutter control. Although some isolated studies for these issues have been proposed [9], techniques of quality estimation and photo filtering are not widely applied as a factor in photo presentation and management.

(4) In conventional photo slideshows, photos are displayed one-by-one, according to alphabetical or temporal order.

MM'06, October 22–28, 2006, Santa Barbara, USA.

Therefore, photos taken in the same scene or having the same topic are separated into different slots, and the browsing experience is cut off. Hua et al. proposed an approach to convert two-dimensional (2D) photos into three-dimensional (3D) videos [10] to create a new browsing manner. However, this system deals less with photo organization and doesn't emphasize the browsing experience via audio-video synchronization.

In this paper, we propose a system that automatically generates music-driven photo slideshows, in which multiple photos having similar characteristics would be displayed in the same frame, and the demonstration of photos proceeds as the pace of the incidental music. Because a frame is tiled by multiple photos, we called the proposed presentation *tiling slideshow*.

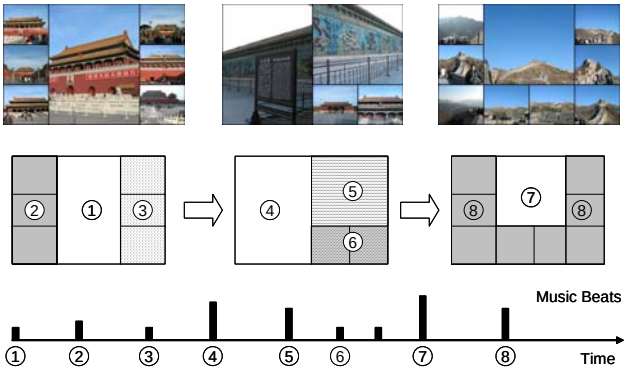

Figure 1. An example of tiling slideshow.

As the example shown in Figure 1, photos are displayed at different frames according to the results of photo clustering. For example, photo sets 1~3 and 4~6 are categorized into two different clusters and are displayed at two frames. In each frame, the time instant of each photo's occurrence is determined by the beats of the incidental music. As a strong beat occurs, e.g. time instant (4) in Figure 1, the displaying content switches to another frame, in which different numbers of photos would be displayed.

The idea of tiling multiple photos into the same frame emphasizes the atmosphere of viewing experience because the coherence of a frame is elaborately maintained. Collaborative presentation of photos that is synchronous to music beats even improves the enjoyment of photo browsing.

To accomplish the aforementioned ideas, we have to combat with several challenges. The issues we discussed are summarized as follows.

- Photo filtering: based on the metadata in EXIF (Exchangeable image format) [11], we first perform orientation correction. To remove photos that are suffered from significant quality degradation, results of blur [9][12] and underexposure/overexposure [13] detection are the clues for quality estimation.
- Photo clustering: the proposed system automatically organizes photo collections by using temporal [6] and content characteristics. We hierarchically integrate them to perform finer clustering so that photos having similar context are grouped together and are displayed at the same frame.
- Layout organization and tiling: photos in the same cluster are arranged in the same frame. Because the space of a frame is fixed, how to arrange these photos, such as locations and sizes, is vital to the final presentation. We model this issue as a constrainted optimization problem and devise a mechanism to evaluate the importance of each photo. Smart cropping based on the salience analysis [14] is also applied to shrink the occupied space of an image.
- Music beat analysis: we detect beats [15] to estimate the pace of the incidental music. Since the synchronization between music and photo occurrence plays an important role in our work, we use beats to drive the progress of presentation.

The rest of this paper is organized as follows. Section 2 describes the system overview, which mainly consists of visual analysis, music analysis, and the composition module. Visual analysis related techniques are described in Section 3, and beat detection of music signals are described in Section 4. Details of the composition module are addressed in Section 5. Section 6 shows the evaluation results, and Section 7 concludes this paper.

2. System Overview

2.1 Essential Idea

The goal of the proposed tiling slideshow system is to generate descriptive presentation via elaborate arrangement of photographs. According to the guidelines of writing, a solid paragraph contains a topic sentence, which identifies the main idea of this paragraph, and several supportive sentences, which provide supportive details of the main idea. Many paragraphs are therefore concatenated to convey the whole narration of an article. Likewise, we advocate that a journey or an event can be reproduced by many *photographic paragraphs*, which are frames shown in Figure 1 and are composed of a topic photo (with larger size) and several supportive photos (with smaller size). Accompanying with the incidental music, the tiling slideshow portrays an audiovisual composition that presents 2D photos in a 3D manner. It is conceptually viewed as a *photographic story*, which is composed of many photographic paragraphs.

2.2 System Overview

The system consists of three main stages, as shown in Figure 2. In the preprocess stage, photos with serious quality degradation are filtered out. Visual quality is estimated via the clues of motion blur and underexposure/overexposure. In the analysis stage, photos are first clustered according to temporal context. Based on the results of time-based clustering, visual features such as color layout and dominant color are used to perform finer content-based classification. The clues of content characteristics will be used in the layout arrangement and importance measurement of each photo. In terms of music content, we perform beat detection to characterize beats information. A music clip can be segmented according to strong beats. There are two phases, say spatial and temporal compositions, in the final stages. In the spatial composition, we perform some manipulations on a cluster of photos and tile them into a frame. The manipulations, such as scale down, cropping, and location assignment, are determined by the designed importance metrics. In the temporal composition, occurrence of parts of a frame and switching between frames are determined by the detected music beats. They are temporally synchronized to make coordinate effects.

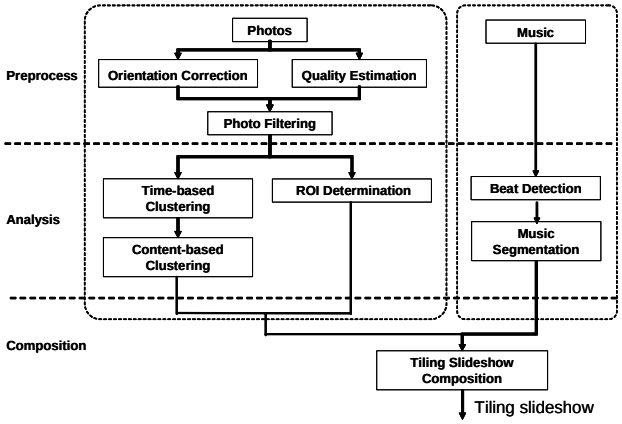

Figure 2. System flowchart of the proposed titling slideshow.

3. Visual Processing

3.1 Orientation Correction

The orientation problem derives from the inconsistency between user's intuition in browsing and the taken angle of a photo. Recently, some studies [16][17] have been conducted for automatic orientation correction. Four orientations, i.e. north (0˚), west (90˚), south (180˚), and east (270˚), are automatically identified. Vailaya et al. [16] proposed a Bayesian learning methodology, which models the characteristics of features extracted by PCA-based and LDA-based mechanisms. Luo and Boutell [17] integrated low-level features with semantic cues, such as sky, water, grass, and faces, to deal with this problem. Several semantic detectors were elaborately designed to facilitate orientation correction for unconstrained consumer photos.

Although the aforementioned methods show promising experimental results, sophisticated modeling and/or feature extraction are not computation tractable. Moreover, for the application of photo slideshow, any misclassification would cause apparent annoyance in browsing. We have an urgent need to perfectly correct mis-orientation. Fortunately, more and more digital cameras are equipped with orientation sensors and simultaneously store orientation information as EXIF metadata when shooting [11]. It is easy and reliable to be used in correcting orientations of photographs.

3.2 Photo Filtering

In general, most users are not familiar with photography. Therefore, the taken photographs often suffer from unwanted defects, which may be in one or two of the following cases:

- *Blur photographs*, which are often caused by hand-shaking or out of focus. Blur recovery of an image without any extra information is still a very challenging and hard problem. Instead of removing blur effects from photographs, we prefer to know whether the photograph is blurred or not and, if yes, how much the blur extent it is. We adopt a wavelet-based blur detection method [9], which detects the occurrence of blur and blue extent based on the statistics of edge characteristics. With this information, photographs with severe blur degradation can be filtered out or kept for further image processing.
- *Underexposure and overexposure photographs*, which are often due to shooting with incorrect exposal camera parameters. To our best knowledge, little literature [13] has been reported on detecting or correcting the ill-conceived photos. We devise a simple detection method based on intensity characteristics. Given a photo, the darkness and brightness of each pixel are estimated as follows:

$$darkness_{ij} = \begin{cases} 1, & \text{if } P_{ij} < T_d \\ 0, & \text{otherwise} \end{cases}$$

$$brightness_{ij} = \begin{cases} 1, & \text{if } P_{ij} > T_b \\ 0, & \text{otherwise} \end{cases}$$

  where P_{ij} is the intensity value of the pixel at position (i,j), and T_d and T_b ($T_b > T_d$) are designated thresholds, which are derived from the statistical characteristics of pixel intensities in underexposure and overexposure photos. When the number of the darkness (brightness) pixels in a photo is larger than 90 percent of the total pixels, this photo is claimed as an underexposure (overexposure) photo.
- *Duplicate photographs*, which denote the same scene is taken in multiple photos. To ensure the diversity of presentation, very similar photos should be removed. To detect duplication, photos are first down-sampled into 8×8 signatures. Dissimilarities between signatures are defined as the Euclidean distances of RGB values. When the dissimilarity between two photos is significantly low, they are claimed as duplicates. In addition, duplicate photos are often taken within a short time period, because people often take several photos at the same site or for the same objects. Therefore, we only perform duplicate detection for the photos that are temporally adjacent to each other. Duplication detection is generally a challenging recognition problem. It can be elaborately implemented by more effective features such as SIFT (scale invariant feature transform) [18] or more sophisticated modeling techniques. In our work, we simply apply a simplified method for the reason of computation complexity.

Low-quality photographs should be filtered out because they have less image appeal. In the following sections, all processes are applied to the filtered photo sets.

3.3 Content Analysis

Each photograph records one event, and a group of photographs narrates a story. Tiling slideshow is suitable for integrating many individual events into stories. To realize the idea of photographic story, as described in Section 2.1, we have to organize photos such that the photos in the same cluster are semantically related and will be displayed in the same frame (photographic paragraph). In general, semantically related photographs have high coherence on temporal and spatial context. Therefore, we hierarchically organize photos from the perspectives of time-based and content-based clustering.

3.3.1 Time-based Clustering

One of the most intuitive and effective ways for organizing photos is to exploit time information. Because "pictures are not taken in a vacuum [7]," contexts of photographs convey certain relationships between them. From the perspective of temporal context, photos taken within a certain time period usually share the same topic and record the same semantic events, such as a birthday party. Therefore, the time information embedded in

photos does a great help to effective organization since photos within the same cluster share certain degrees of semantic correlation.

To achieve this job, the time-based clustering algorithm proposed in [3] is adopted in our work. Photos are first sorted by their creation time. This algorithm dynamically determines noticeable time gaps through checking the temporal context of photos in a sliding window, say 10 photos. As shown in Figure 3, the photos in the top row were temporally close to each other and are categorized into the same cluster. On the other hand, the photos in the bottom row are divided into two clusters because a notable time gap exists. This method dynamically determines time gaps that would vary significantly and reveals the change of shooting pace.

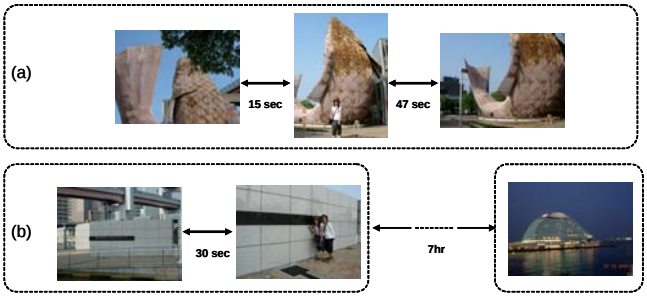

Figure 3. Examples of time-based clustering aided by temporal context. The photos in the top row are categorized into the same cluster, while the photos in the bottom row are in two clusters.

3.3.2 Content-based Clustering

In addition to the time-based clustering, the spatial context of the photos in the same cluster is exploited to conduct finer organization as well. In the same frame, the adjacent "tiles" are expected to be visually similar to each other. The homogeneity of adjacent tiles emphasizes the tone of photographic paragraphs. Therefore, we further perform content-based categorization for the photos in the same time-based cluster.

We exploit dominant color and color layout descriptors defined in MPEG-7 [20] as the spatial context. Dominant color descriptor represents the statistical characteristics of an image. Color layout descriptor specifies a spatial distribution of colors and roughly describes the structure of whole image. The distances between two photos in terms of dominant color and color layout are respectively calculated. They are both normalized to the range [0,1]. The spatial context distance $d(.)$ between two photos is then defined as the average of normalized dominant color and color layout distances.

Let P_i (i = 1,2,…,N) be the ith photo in a time-based cluster Ψ. To measure the homogeneity in Ψ, we define a metric

$$D_s = \max_{\substack{P_i,P_j \in \Psi \\ P_i \neq P_j}} d(P_i, P_j). \tag{1}$$

If D_s is larger than a threshold ζ (which is set as 0.5), at least two photos in the time-based cluster are very dissimilar. In this case, we perform content-based clustering for finer organization. The threshold ζ can be adjusted based on user preferences. If larger (smaller) value is set, the final slideshow video would prefer to the results with more (less) photos displaying in the same frame.

After performing content-based clustering, these N photos are clustered into m classes $C_1,\ldots,C_m$, where C_m has n_m photos. In order to measure the goodness of clustering results, we define the within-cluster distance S_w as follows:

$$S_w = \max_{g=1,\ldots,m} \frac{1}{n_g(n_g-1)} \sum_{P_i \in C_g} \sum_{\substack{P_j \in C_g \\ P_i \neq P_j}} d(P_i, P_j), \tag{2}$$

where both P_i and P_j are categorized in the cluster C_g. The within-cluster distance denotes the average distance between the photos in the same content-based cluster. We find the maximal value to represent the worst case in this clustering situation.

On the other hand, we define the between-cluster distance S_b as:

$$S_b = \min_{\substack{g \neq h \\ g=1,\ldots,m \\ h=1,\ldots,m}} \frac{1}{n_g n_h} \sum_{P_i \in C_g} \sum_{P_j \in C_h} d(P_i, P_j), \tag{3}$$

where P_i and P_j belong to different content-based clusters. The between-cluster distance denotes the average distance between different clusters. Likewise, the minimum value is calculated to represent the worse case in this clustering situation.

We hope that the between-cluster distance can be as large as possible, and the within-cluster distance can be as small as possible. Conceptually, we want to maximize the homogeneity of photos in the same cluster and the heterogeneity between clusters. Therefore, we jointly consider these two distances and find the clustering situation that has the largest S_b/S_w.

Given a photo set $\Psi=\{P_1,P_2,\ldots,P_N\}$ belonging to the same time-based cluster, the content-based clustering algorithm can be expressed as:

CONTENT_BASED-CLUSTERING(Ψ)

1 For $i = 1$ to n
2 G_i = ADJ-GRAPH(Ψ, i)
3 Calculate S_b and S_w based on clustering result G_i
4 $R_i = S_b / S_w$
5 Return G_i with maximum R_i

ADJ-GRAPH(Ψ, k)

1 Initialize a graph G with N disjunct nodes
2 For each node in G
3 Connect node i and node j if they are mutually k nearest neighbors
4 Return the adjacency graph G

Note the $R_i=S_b/S_w$ denotes the joint influence of homogeneity within a cluster and heterogeneity between clusters in a specific clustering situation, which is indicated by the adjacency graph G. In the graph G, each node denotes a photo, and nodes that are connected belong to the same content-based cluster. This algorithm evaluates possible cluster situations and determines how to cluster photos on the basis of the visual conformance.

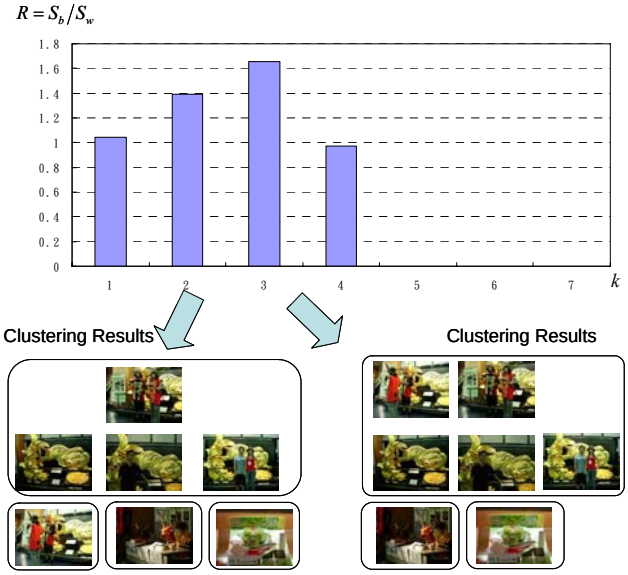

Figure 4. Examples of content-based clustering.

Figure 4 shows an example illustrating the trend of R and the parameter k for the nearest neighbors. Real samples are also shown to demonstrate the relationship between R and content-based clustering results. Photos in the same block are categorized as the same content-based cluster. In this figure, R is maximal when k=3, and the content-based clustering result is better than that in the case of k=2. This figure demonstrates the positive correlation between the designed metric and the clustering results. After this process, the average number of photos per cluster is 5.88 in our experiments.

3.3.3 Region of Interest Determination

The essential idea of tiling slideshow is to arrange photos of the same cluster into the same layout such that sympathy can be aroused to mold a new browsing experience. Because the frame space is smaller than that of multiple photos, it is inevitable to shrink photos into smaller tiles to fit in one frame. The simplest way is to directly resize photos according to their aspect ratios. However, blind resizing often causes significant information loss because the details of important objects would be rudely shrunk. In order to make information loss as small as possible, we prefer cropping a region from the original photo, which retains the important and attractive part.

In this work, region-of-interest (ROI) is derived based on user attention model. User attention modeling [14] has been proposed to evaluate the attentive values of visual data. Two approaches, i.e. top-down and bottom-up, have been addressed in the literature [24].

- Top-down approach: Human faces are highly semantic and often more attractive than other objects. We exploit OpenCV face detector to find face regions, as shown in Figure 5(b).
- Bottom-up approach: We apply the contrast-based visual attention model to perform salience analysis, as proposed in [14]. The model assumes that a human perception is closely related to the strength of the contrast between objects and backgrounds. We adopt this approach to generate a saliency map (c.f. Figure 5(d)), in which high intensity parts indicate more attentive regions. On the basis of the saliency map, the center of gravity and the ranging variance of the saliency map are evaluated to determine the position and width/height of the attentive region (c.f. Figure 5(e)).

According to these two approaches, we determine region-of-interest (ROI) that includes the attentive regions. The detected attentive region will be the unit for cropping and/or resizing.

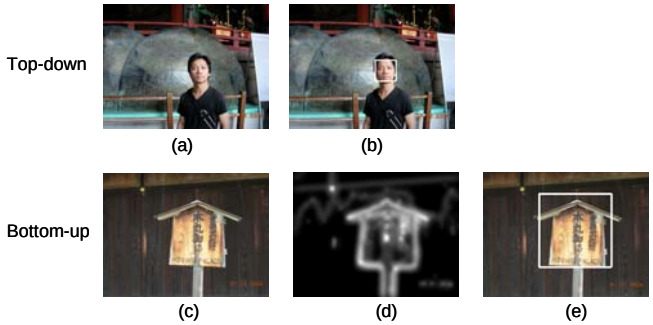

Figure 5. The process of ROI determination. (a)(c) The original photo, (b) face detection result, (d) saliency map derived from color contrast, (e) attentive region derived from the saliency map.

4. Music Analysis

Music plays an important role in multimedia presentation. Accompanying with the pace of the incidental music, the tiling slideshow not only concerns about how to construct solid photographic paragraphs, but also put efforts on how to concatenate them as an affective photographic story. To achieve this goal, we analyze beats information in music and segment the music into sub-units. The coordination between visual and aural media is maintained by aligning each tiling frame with a corresponding sub-unit.

Beats correspond to the sense of equally spaced temporal units [15]. To coordinate visual and aural media, we try to match the occurrence of photos and frame switching with music beats. In our work, the beat tracking algorithm proposed in [15] is applied to extract music beats. This method analyzes music signals in different frequency bands and estimates beats information after envelope extraction. Beat information serves as the timer for photo presentation. Furthermore, the timing of two-level of presentation, i.e. frame switching and photo display in a frame, should be further designed as follows.

1) Timing for frame switching:

In addition to music beats, we also consider energy dynamics to determine the timing for frame switching. The sound energy difference is calculated based on root mean square (RMS) values [19] of adjacent audio frames. In the example of Figure 6, if the starting time of frame 1 is t_1, we check the sound energy differences in the range from (t_1+r_1) to (t_1+r_2), and the largest energy difference in this range is detected. To guarantee the coordination between visual and aural media, the timestamp of the nearest beat to the largest energy difference is set as the timing for frame switching, like timestamp t_4 in Figure 6. In our implementation, r_1 and r_2 are set as 4 and 6 seconds, respectively. This method prevents too short displaying frames and simultaneously considers multimodal coordination.

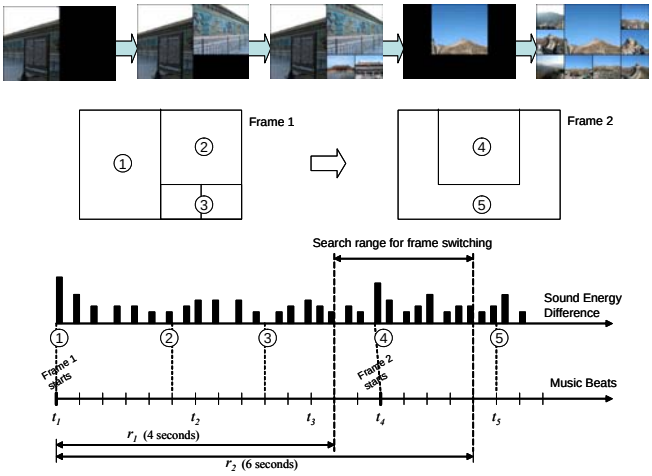

Figure 6. An example of determining the timing for frame switching and photo display.

2) Timing for photo display in a frame:

In each frame, we have to determine the occurrence timestamp of each photo (or a photo set shown in frame 2). Many variations can be used for this task. One of them is to unequally dispatch displaying time according to the allocated area of each photo. In our implementation, we prefer to averagely distribute the displaying duration to each photo. We find the timestamps of the music beats that are nearest to the averagely distributed points. With this elaborate design, the proposed scheme synchronizes visual slideshow with the pace of the incidental music and constructs a new type of photo presentation.

5. Tiling Slideshow Composition

A tiling slideshow displays photos that are in the same time-based cluster at the same frame and elaborately arranges tiles based on content-based characteristics. We face several challenging problems in the composition process:

Problem 1: Given a time-limited music clip, we often have to select a subset of photo clusters for displaying. As a frame lasts for 4~6 seconds, the slideshow only affords at most 60 clusters of photos if the user selects a 4-min music clip. The importance of a photo cluster is, therefore, defined to be the metric for cluster selection.

Problem 2: Given a cluster of photos, we hope to reasonably manipulate them such that more important or attractive photos occupy larger space, and photos having similar characteristics are located closely. On the basis of predefined layout sets, which define the arrangement of included tiles, we have to devise a method to select the most appropriate layout to be the display platform.

Problem 3: Once the matching between photos and tiles are determined, we have to resize or crop the original photos to fit in with the limited region. We model this issue as a constrainted optimization problem.

The following sub-sections describe the details of the composition process.

5.1 Photo Importance

5.1.1 Cluster-based Importance

The basic unit for frame composing is a content-based cluster described in Section 3.3. We estimate the importance of each cluster by two features: *PPM* (photos per minute) and *PC* (photo conformance). Given a time-based cluster Ψ and the included content-based clusters C_g, g=1,2,…,m, we define the *PPM* as:

$$PPM(C_g) = N(\Psi)/Time\_Duration(\Psi), \tag{4}$$

where $N(\Psi)$ denotes the number of photos in Ψ, and *Time_Duration*(.) returns the time difference between the first photo and the last photo in the cluster, which is temporally sorted.

Photo conformance is defined as

$$PC(C_g) = 1 - \frac{1}{n_g(n_g-1)} \sum_{P_i \in C_g} \sum_{\substack{P_j \in C_g \\ P_i \neq P_j}} d(P_i, P_j). \tag{5}$$

It's conceptually similar to the within-cluster distance.

These two features are concatenated as a feature vector $\vec{x}=(PPM(C_g), PC(C_g))$. The cluster-based importance CI_g for the gth cluster is estimated as:

$$CI_g = E(\vec{x}_g) + \frac{1}{2(m-1)+m\lambda} \sum_{g=1}^{m} \left| \vec{x}_g - E(\vec{x}_g) \right|, \tag{6}$$

where $E(\vec{x}_g)$ is the mean of the feature vectors in gth cluster, and the parameter $\lambda > 0$ is a predefined constant. This is a nonlinear fusion scheme. The greater a feature is, the more greatly it affects the returned value [22]. The calculated cluster-based importance is the metric for solving the **problem 1**.

5.1.2 Photo-based Importance

The photo-based importance is defined to facilitate space allocation in a frame. It is calculated based on two features, *FR* (face region) and *AV* (attention value).

For one photo P_i, the face region feature is defined as

$$FR(P_i) = \sum_{j}^{n_f} Area(Face_j) / Area(P_i), \tag{7}$$

where n_f faces are detected in this photo and the feature *FR* is calculated as the ratio of sum of face regions to the total area of P_i.

On the other hand, the attention value of P_i is defined as

$$AV(P_i) = \sum_{x} \sum_{y} Sa(x,y) \times G(x-m_1, y-m_2), \tag{8}$$

where $Sa(x,y)$ denotes the salient value of the pixel at (x,y), and (m_1,m_2) is the position of the centroid of the saliency map, which is generated in Section 3.3.3. The attention value of P_i is weighted sum of saliency values, in which the weighting function is a zero-mean Gaussian centered at (m_1,m_2).

The linear weighting method is applied to calculate the photo-based importance. Because the semantic concept of face is clear than the low-level attention value, higher weight is employed for face region. Both features should be normalized before fusing them together, and the fusion equation is

$$PI_i = W_{face} \times FR(P_i) + W_{attention} \times AV(P_i) \tag{9}$$

The calculated photo-based importance is treated as the metric for solving the **problem 2**.

5.2 Cluster Selection

Given a user-selected music clip, it is divided into smaller segments according to the information described in Section 4. A (time-based) cluster of photos are, therefore, displayed in each music segment (4~6 seconds). Nevertheless, it's often the case that thousands of photos (and therefore hundreds of photo clusters) cannot be completely displayed because a time-limited music clip (e.g. 4 minutes) is divided to just tens of segments. To solve this problem, photo clusters are sorted based on the cluster-based importance in descending order, and the first N_s clusters are picked for presentation if only N_s music segments are available.

5.3 Template Determination

Once the clusters to be displayed are determined, the problem now is how to select appropriate layouts for presentation. We define several templates for showing different numbers of photo in a frame. Figure 7 shows some examples of templates for showing 3, 4, or 5 photos. Intuitively, if the number of photos in a selected cluster is 4, we just select the templates with 4 cells for presentation. To enrich the variety of displaying layout, several templates with 4 cells are designed. Therefore, we have to devise a method to determine which 4-cell template is appropriate for showing the given photo cluster, and determine which photo should be put into which cell.

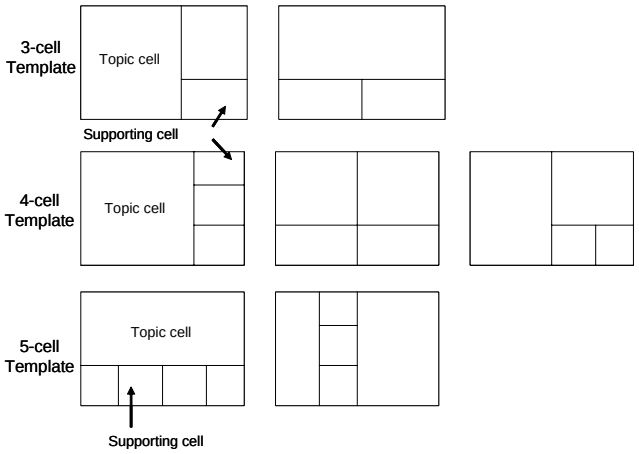

Figure 7. Examples of different kinds of templates.

5.3.1 Template Importance

Each tiling template consists of at least one topic cell and several supportive cells. Because the area of topic cells is larger than that of supportive cells, it is obvious that the photos in topic cells would be more attractive to humans. To describe the importance among cells, we define the cell importance value in the following.

Given a tiling template T and its cells Tc_i, i=1,2,…,k, the importance value of each Tc_i is the ratio of the area of Tc_i over the area of T, that is

$$Ic_i = Area(Tc_i) / Area(T). \tag{10}$$

Importance values of the cells in the same template are packed into a vector in descending order. The corresponding vector to the ith template is called its template importance vector TV=(Ic_1, Ic_2, …, Ic_k).

5.3.2 Template Determination

Given a set of h-cell templates, Γ=$\{T_{h,1}, T_{h,2}, \ldots, T_{h,s}\}$, a photo cluster that contains h photos should be mapped to one of these templates. To choose the most suitable tiling template for the cluster of photos, we generate a photo importance vector PV by packing photo's importance in descending order, which is calculated in Section 5.1.2. The included angle between photo-based importance vector and template importance vector is defined as the metric of template determination:

$$T_{h,i} = \arg \min_{i=1,2,\ldots,s} \left(\frac{PV \cdot TV_{h,i}}{\| PV \| \| TV_{h,i} \|}\right), \tag{11}$$

where $TV_{h,i}$ is the corresponding template importance vector of the template $T_{h,i}$. The minimum included angle between two vectors denotes the best match between photos and templates. Because both importance vectors are sorted in descending order, which photo should be put into which cell is also determined by the equation (11). That is, more important photos should be put into larger cells. This process solves the prescribed **problem 2**.

5.4 Composition

The final task for generating the tiling frame is to put photos into the cells of the determined template. This process is like to stick tiles on the wall, and that's why we call this work tiling slideshow. Nevertheless, the ratio of width to height of each cell is often different from that of the selected photo. Moreover, the resolution of photos taken by current digital cameras is at least two million pixels (about 1600×1200), which is significantly larger than the targeted resolution, 720 × 480, in DVD. Therefore, it's unavoidable that we should resize the photos to fit into the layout. In this subsection, we describe how to crop a photo such that it can be perfectly put in a designate cell.

5.4.1 Region Selection

In order not to significantly distort the content of each photo, we model our cropping operation as the following constrainted optimization problem. Given a photo, we want to find the region R, which has the same aspect ratio as the designate cell and possesses the largest content value. This selection can be formulated as:

$$\max_R C(R), \text{ such that } g(R) = g(R_c), \tag{12}$$

where $g(R)$=w_R/h_R and $C(R)$ denotes the estimated content value of the region R (with width w_R and height h_R). R_c is the region of a specific cell, and g(.) returns the aspect ratio of a region.

As described in Section 3.3.3, we can evaluate user attention from top-down or bottom-up approaches. We evaluate the region R's content value either in top-down or bottom-up manners according to the occurrence of the face region.

1) The top-down case:

$$C(R) = \sum_{x=r_1}^{r_1+w_R-1} \sum_{y=r_2}^{r_2+h_R-1} IMP_{top-down}(x, y), \tag{13}$$

where $IMP_{top\text{-}down}$ denotes the importance map that is generated by applying a zero-mean 2D Gaussian to the centroid of the largest face region. Every pixel in this photo has its corresponding importance value derived from the Gaussian distribution. In this case, the maximal $C(R)$ occurs at the region located at the center

of the largest face region, which is a reasonable and appropriate region to be preserved.

2) The bottom-up case:

$$C(R) = \sum_{x=r_1}^{r_1+w_R-1} \sum_{y=r_2}^{r_2+h_R-1} IMP_{bottom-up}(x, y). \quad (14)$$

If there is no face in the photo, the importance map is derived from visual features, and is the same as the saliency map described in Section 5.1.2. That is,

$$IMP_{bottom-up}(x, y) = G(x - m_1, y - m_2), \quad (15)$$

where (m_1,m_2) is the centroid of the saliency map. In these formulations, we try to find a region which cause minimum information loss after cropping. The effects of resizing can be neglected because all candidate regions have the same scaling factors and don't change the optimization result.

5.4.2 Implementation

The formulation described above is in a full-search manner. However, we can exploit some tricks to save computation in practical implementation. In top-down cases, we first find the centroid of the largest face region. Starting from this position, we expand the region towards four directions (top, down, left, right) according to the aspect ratio of the targeted cell. The expansion stops when at least two boundaries of this region reach the boundaries of the photo. After this process, the selected region has the maximal content value, constrained by aspect ratio consideration. Finally, the selected region is resized to stick on the targeted cell. Likewise, the only difference in bottom-up cases is that we start expansion from the centroid of the saliency map.

Through the processes of cluster selection, template determination, and image manipulation, we can determine the styles of display, the positions and sizes of photos. After determining the timing for displaying a photo or switching frame by the music beats, a tiling slideshow is finally generated. To facilitate more gorgeous presentation, we also include transition effects such as fade-in to display photos.

5.4.3 Discussion

In addition to the content-based features we used for content value estimation or template determination, more high-level information or user intervention can be involved in our system. Although traveling photos are used in our experiments, the tiling slideshow can be generated for different types of consumer photos. Moreover, since the image annotation is often widely applied in photo sharing community like Flickr [23], semantic metadata or user preference can be considered in the template determination process in the near future.

6. EVALUATION

As objective evaluation of the proposed system is difficult, we evaluate the performance of our system through subjective experiments. Three photo sets, which are taken by three different amateurs, are used for evaluation. Detailed information of the evaluation photo sets is listed in Table 1. Note that photo sets 1 and 2 respectively combine photos in multiple trips taken by the same person. The incidental music clips are all pop music, while that for sets 1 and 2 are pure music (without vocal performance) and that for the third photo set contains singing.

Table 1. Information of evaluation photo sets.

	Photo set 1	Photo set 2	Photo set 3
# of photos	780	522	1257
Time span	07/01/2003 ~ 04/04/2006	01/11/2005 ~ 07/12/2005	07/19/2004 ~ 07/31/2004
Length of the incidental music	3 minutes 31 seconds	4 minutes 38 seconds	4 minutes 6 seconds
Geographic spread	Osaka, Kyoto, Kobe (Japan)	Melbourne, Brisbane (Australia)	Osaka, Kyoto, Kobe (Japan)
	Nagoya, Tokyo (Japan)	Amsterdam (Netherlands)	

We compare the satisfaction of tiling slideshow with the slideshow generated by the commercial software ACDSee [4] and Photo Story [21]. In [4], conventional slideshow generated by sequentially switching photos one-by-one is provided. It has no ability to accompany the slideshow with music. On the other hand, Photo Story generates camera motion effects on single photo and sequentially switching photos as well. Accompanying the slideshow with music is affordable in Photo Story.

Twenty-seven evaluators are invited to join the user study. In our experiments, the results generated by ACDSee, Photo Story, and tiling slideshow are projected on a 60-inch screen. Evaluators spend the same time length as incidental music in viewing the slideshows. For example, they spend 3 minutes 31 seconds in viewing the slideshow generated based on photo set 1.

These evaluators are asked to give scores from 1 to 10 to show their satisfactions (higher score means better satisfaction) with respective to the following perspectives:

- Question 1: Richness

  (How do you feel the photo variety in a time unit?)
- Question 2: Fun

  (Do you think it's a funny presentation?)
- Question 3: Experience

  (Do you think the sequence helps you experience this trip?)
- Question 4: Acceptance

  (If you have this tool, are you willing to use it to generate your own slideshow?)
- Question 5: Atmosphere

  (How do you feel the audiovisual effects of this slideshow?)

The scores of the slideshows generated by ACDSee are fixed as 5 to be the baseline. The results of subjective tests are illustrated in Figure 8. Generally, the titling slideshow has significantly better satisfactions than others. The performance differences between Photo Story and tiling slideshow in question 5 are relatively small. We think that it's because the beat estimation of music clips is not

good enough, and the synchronization between visual and aural media doesn't always act as humans expected.

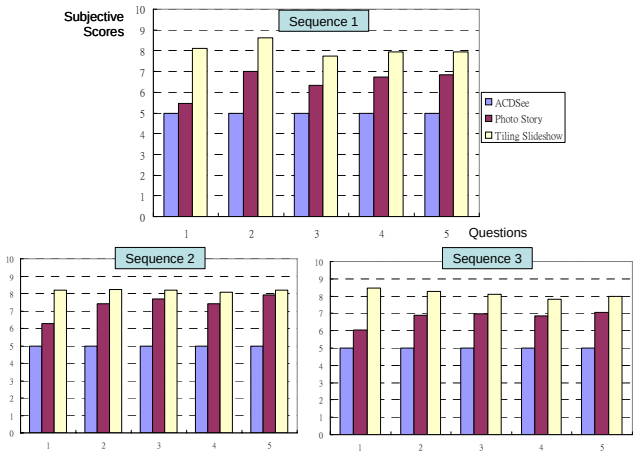

Figure 8. Results of subjective evaluation for three photo sets.

For tiling slideshow, we further evaluate the performances in terms of content-based clustering and layout determination. Two questions are asked:

- Question 6: How do you feel the visual coherence of photos in the same frame?
- Question 7: How do you feel the layout of display?

Figure 9 shows the evaluation results of these three tiling slideshows. In general, because the layout varies widely in presentation, the results in three sequences are similar for question 7. On the other hand, because the size of photo set 3 is larger, and photos in which are more complex, the performance of content-based clustering is slightly worse than others. Moreover, humans are sensitive to the abnormal situation that important objects (e.g. faces and human bodies) are cropped. It's also the reason why the third tiling slideshow obtains worse judgment. Figure 10 shows some snapshots of these three tiling slideshows.

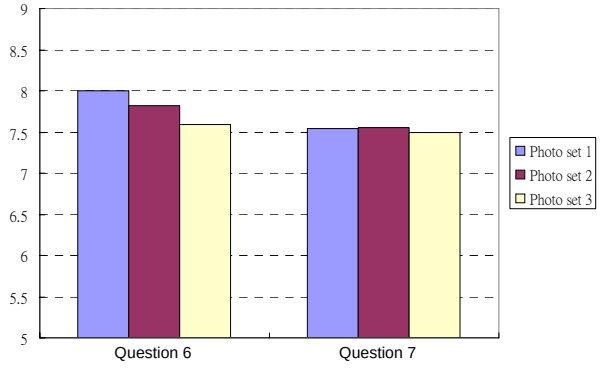

Figure 9. Evaluation results of three different tiling slideshows.

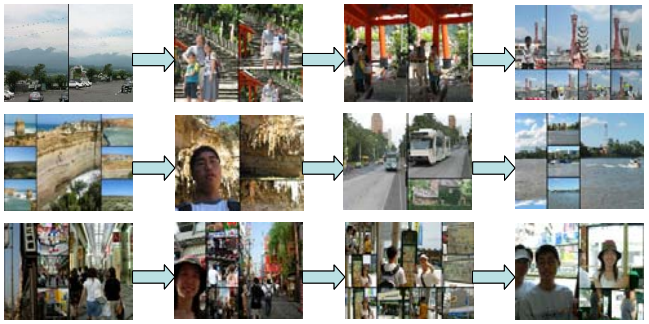

Figure 10. Some snapshots of the evaluated tiling slideshows.

7. CONCLUSION

The proposed tiling slideshow system automatically generates a composite audiovisual presentation. Photo collections are first organized according to temporal and spatial contexts, which are the units for agglomerative presentation. For a cluster of photos, the cluster-based importance and photo-based importance are estimated. They are the metrics for cluster selection and smart photo manipulation, respectively. Accompanying with incidental music, the tiling slideshow not only switches frames with music pace, but also displays each "tile" according to beat information. The results of subjective evaluation demonstrate the satisfaction of this new kind of presentation.

Many related issues are still worthy to investigate. For example, more elaborate features can be used in content-based clustering. The accuracy of music beat detection is also one direction for improving audio-video synchronization. Furthermore, the tiling process can be customized via user's feedback or cooperating with some useful metadata.

8. REFERENCES

[1] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., and Jain, R. Content-based image retrieval at the end of early year. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349-1380, 2000.

[2] Geigel, J., and Loui, A. Using genetic algorithms for album page layouts. *IEEE Multimedia*, vol. 10, no. 4, pp. 16-27, 2003.

[3] Platt, J.C., Czerwinski, M., Field, B.A. PhotoTOC: automating clustering for browsing personal photographs. In *Proceedings of IEEE Pacific Rim Conference on Multimedia*, pp. 6-10, 2003.

[4] ACDSee, http://www.acdsystems.com

[5] Picasa, http://picasa.google.com/

[6] Cooper, M., Foote, J., Girgensohn, A., and Wilcox, L. Temporal event clustering for digital photo collections. In *Proceedings of ACM Multimedia*, pp. 364-373, 2003.

[7] Luo, J., Boutell, M., and Brown, C. Pictures are not taken in vacuum – an overview of exploiting context for semantic scene content understanding. *IEEE Signal Processing Magazine*, vol. 23, no. 2, 2006.

[8] Hua, X.-S., Lu, L., and Zhang, H.-J. Content-based photograph slide show with incidental music. In *Proceedings*

of IEEE International Symposium on Circuits and Systems, vol.2, pp. 648-651, 2003.

[9] Tong, H., Li, M., Zhang, H.-J., and Zhang, C. Blur detection for digital images using wavelet transform. In *Proceedings of IEEE International Conference on Multimedia & Expo*, pp. 17-20, 2004.

[10] Hua, X.-S., Lu, L., and Zhang, H.-J. Automatically converting photographic series into video. In *Proceedings of ACM Multimedia*, pp. 708-715, 2004.

[11] Digital Still Camera Image File Format Standard. Japan Electronic Industry Development Association, 1998.

[12] Ben-Ezra, M., and Nayar, K. Motion-based motion deblurring. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, vol. 26, no. 6, pp. 689-698, 2004.

[13] Yan, W.-Q., and Kankanhalli, M.S. Detection and removal of lighting & shaking artifacts in home videos. In *Proceedings of ACM Multimedia*, pp. 107-116, 2002.

[14] Ma, Y.-F., Lu, L., Zhang, H.-J., and Li, M.J. A user attention model for video summarization. In *Proceedings of ACM Multimedia*, pp. 533-542, 2002.

[15] Scheirer, E.D. Tempo and beat analysis of acoustic musical signals. *Journal of Acoustical Society of America*, vol. 103, no. 1, pp. 588-601, 1998.

[16] Vailaya, A., Zhang, H., Yang, C., Liu, F.-I., and Jain, A.K. Automatic image orientation detection. *IEEE Transactions on Image Processing*, vol. 11, no. 7, pp. 746-755, 2002.

[17] Luo, J., and Boutell, M. Automatic image orientation detection via confidence-based integration of low-level and semantic cues. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, vol. 27, no. 5, pp. 715-726, 2005.

[18] Lowe, D.G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[19] Panagiotakis, C., and Tziritas, G. A speech/music discriminator based on RMS and zerocrossings. *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 155-166, 2005.

[20] Chang, S.-F., Sikora, T., Purl, A. Overview of MPEG-7 standard. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 688-695, 2001.

[21] Photo Story 3, http://www.microsoft.com.

[22] Ma, Y.-F., Hua, X.-S., Lu, L., and Zhang, H.-J. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 907-919, 2005.

[23] Flickr, http://www.flickr.com/

[24] Itti, L., Koch, C., and Niebur, E. "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, 1998.