Tiling Slideshow: An Audiovisual Presentation Method for Consumer Photos

Wei-Ta Chu[1], Jun-Cheng Chen[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

{wtchu,pullpull,wjl}@cmlab.csie.ntu.edu.tw

ABSTRACT

A new type of audiovisual presentation is proposed to displays well-organized photos in a tile-like manner. Multiple photos that have similar characteristics are well manipulated to construct a frame. Displays of photos are accompanied with the pace of incidental music. This kind of presentation shows visual and aural coordination so that user's sympathetic responses can be aroused to mold a new browsing experience.

1. INTRODUCTION

Digital camera has become an indispensable commodity for each family or individual in recent years. With the advance of digital storage technology, 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 many potential problems in information access. People have to spend much time in browsing and often get lost in massive photo collections. Therefore, we have urgent needs in advanced analysis and presentation techniques that facilitate efficient photo organization and affective photo browsing.

Some commercial photo browsers and research projects [1][2][3] have provided thumbnails or photo management functionalities. Although they provide certain ways for managing and accessing image/photo collections, some critical problems still significantly impede users' browsing experience:

- Large amounts of disordered photos stuff user's storage and make photo browsing and access tedious. One of the most popular ways to present photos is slideshow. However, sequentially browsing often takes much time and makes users weary.

- Consumer photos taken by amateurs are often suffered from quality degradation. Techniques of quality estimation and photo filtering should play an important role in photo presentation and management.

- Conventional photo slideshows display photos one-by-one, according to alphabetical or temporal order. Therefore, photos taken in the same scene or having the same topic are separated into different slots, and the browsing experience is cut off.

We propose a system that automatically generates audiovisual slideshows, in which multiple photos with similar characteristics are displayed at the same frame, and the presentation proceeds following the pace of the incidental music. As the example shown in Figure 1, the timestamp 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. Because the appearance of the final result is like to stick vary-sized tiles on a wall, we call the proposed presentation *tiling slideshow*.

The goal of the proposed tiling slideshow system is to generate descriptive presentation via elaborate arrangement of photographs. According to the guidelines of technical writing, a solid paragraph contains a topic sentence that identifies the main idea and several supportive sentences that provide supportive materials. Many paragraphs are 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 the frames shown in Figure 1. Each frame is composed of a larger-sized "topic photo" and several smaller-sized "supportive photos." Tiling multiple photos into the same frame emphasizes the atmosphere of viewing experience. Photo presentation that is synchronous to music beats even improves the enjoyment of browsing.

On the basis of this idea, we propose a system that integrates visual and music analyses and automatically composes a vivid audiovisual presentation, as shown in Figure 2. The issues we discussed are summarized as follows.

- Photo processing: we perform orientation correction and remove ill-quality photos that are caused by blur and/or underexposure/overexposure.
- Photo organization: the proposed system automatically organizes photo collections by using temporal and content characteristics. We integrate them to perform finer clustering so that photos at the same scenic spots or presenting the same event are grouped together.
- Music beat analysis: we detect music beats and use them to drive the progress of presentation.
- Temporal and spatial composition: from temporal perspective, photo presentation and frame switching are synchronous to music beats. From spatial perspective, photos having similar characteristics are elaborately manipulated and arranged at the same frame.

2. Visual Processing

2.1 Photo Preprocessing

Before stepping further to elaborate organization, we try to correct photo orientation and filter out ill-quality photos. These preprocesses prevent users from a great deal of tedious work.

- *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 [4][5] have been conducted for automatic orientation correction. However, the reported methods are often sophisticated and are not computationally tractable. For the application of photo slideshow, 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 [6] when shooting. It is easy and reliable to be used in correcting photo orientation.

- *Blur Detection*

Most consumers are not familiar with photography, and the taken photographs often suffer from unwanted defects. Blurred photos are often caused by hand-shaking or out of focus. In this system, we adopt a wavelet-based method [7] to detect the occurrence of blur through edge characteristics in different resolutions. With this information, photographs with severe blur degradation can be filtered out.

- *Underexposure and Overexposure Detection*

Photos with bad exposure condition are often due to incorrect exposal camera parameters. We devise a simple detection method based on intensity characteristics [8]. When the number of the darkness (brightness) pixels in a photo is larger than a predefined threshold, this photo is claimed as an underexposure (overexposure) photo.

2.2 Photo Organization

To realize the idea of photographic paragraph, we have to organize photos so that photos in the same cluster are semantically related and are displayed at the same frame. Therefore, we organize photos based on time and content characteristics.

2.2.1 Time-based Clustering

From the perspective of temporal context, photos taken within a certain time period usually share the same topic and record the same semantic events. Based on this idea, the time-based clustering algorithm proposed in [2] is adopted in our work. Photos are first sorted by their shooting time. This algorithm dynamically detects noticeable time gaps through checking the timestamps of photos in a sliding window. These time gaps present changes of shooting pace and reveal that photos in different places or describing different events are taken.

2.2.2 Content-based Clustering

From content-based perspectives, we exploit dominant color and color layout descriptors defined in MPEG-7 as features. Dominant color represents the statistical color characteristics of an image. Color layout represents a spatial distribution of colors and roughly describes the structure of whole image. The average of normalized dominant color and color layout distance is used to measure the similarity between photos.

In clustering process, we conceptually prefer that photos in the same cluster should be similar to each other, and photos in different clusters should be distinct as much as possible. Given a set of photos, we try different clustering cases and evaluate the corresponding goodness, which is given by the ratio of inter-cluster distance over intra-cluster distance. The clustering case with the largest goodness value is adopted. We have shown the effectiveness of this clustering method in [8], and we provide more evaluation in this paper.

2.2.3 Region of Interest Determination

Because the frame space is smaller than that of multiple photos, it is inevitable to shrink photos 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 less as possible, we prefer cropping a region that retains the most important and attractive part from the original photo.

In this work, region of interest (ROI) is detected based on user attention model [9], which is used to evaluate the attentive values of visual data. According to whether human faces exist in photos, user attention is modeled by top-down or bottom-up approaches. Then the detected attentive region will be the unit for cropping and resizing.

3. 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 detect music beats based on the algorithm proposed in [10]. This method analyzes music signals in different frequency bands and estimates beats information therein. Beat information serves as the timer for photo presentation, and the timing for frame switching and photo displaying are determined as follows.

- Timing for frame switching

In addition to music beats, we also consider sound energy differences between adjacent audio frames in frame switching. In the example of Figure 3, 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 3. In our implementation, r_1 and r_2 can be adjusted to control the displaying speed and meet different people's preferences.

- Timing for photo display at a frame

At each frame, we have to determine the occurrence timestamp of each photo. 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 (e.g. from t_1 to t_4) 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.

4. Tiling Slideshow Composition

After the processes described above, we try to put photos in the same cluster at the same frame to compose descriptive presentation. At the composition stage, we have to face several challenging problems.

Challenge 1: Given a time-limited music clip, we often have to select a subset of photo clusters for displaying. If a frame lasts for 4~6 seconds (assigned by users), 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.

Challenge 2: Given a cluster of photos, we hope to reasonably manipulate them so that more important or attractive photos occupy larger space, and photos having similar characteristics are located closely. Based on predefined templates, we have to devise a method to select the most appropriate one to be the display platform.

Challenge 3: Once the matching between photos and tiles are determined, we have to elaborately resize or crop the original photos to fit in with the limited region.

4.1 Cluster Selection (for Challenge 1)

Given a user-selected music clip, it is divided into smaller segments according to the information described in Section 3. A cluster of photos are expected to be displayed within each music segment. 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) can only be divided to tens of segments. To solve this problem, photo clusters are sorted based on the cluster-based importance in descending order, and the first N clusters are picked for presentation if only N music segments are available.

We estimate the importance of each cluster by two features: *PPM* (photos per minute) and *PC* (photo conformance). PPM denotes the shooting frequency of photos in a cluster, while PC denotes the content-based similarities between photos in the same cluster. These two features are fused together to describe the importance of a given cluster. Details of cluster-based importance measurement please refer to [8].

4.2 Template Determination (for Challenge 2)

4.2.1 Template Design

Once the clusters to be displayed are determined, the problem now is how to select appropriate layouts for presentation. According to the guidelines of publication layout [11], we design several templates for showing different numbers of photos.

- *Show limited content in a limited space*: presenting too many photos at the same frame confuses viewers and obscures the subject of presentation. Therefore, the number of photos in a frame is limited to no larger than twelve in this work. A region that displays one photo is called a *cell*, and we design various displaying templates containing from one cell to twelve cells [8].

- *Enlarge important photos to drive visual perception*: photos at the same frame are elaborately scaled into different sizes to show their relative importance. Therefore, the areas of different cells in a template should be differentiated to show variations.

- *Designing layouts by adjusting uniform subunits*: this way not only enriches the spatial arrangement but also maintains the regularity of presentation. In our implementation, we divide a frame into twelve equal-sized basic units. To construct a template that consists of four cells, for example, one construction method is to merge the left nine regions into a large cell and leaves the remaining three regions as three small cells. Figure 4 shows some templates for showing three, four, and five photos.

4.2.2 Template Determination

Intuitively, if the number of photos in a selected cluster is four, we just take the templates with four cells for presentation. As we describe above, we design several templates with four cells to enrich displaying layouts. We have to determine which 4-cell template is appropriate for showing the given photo cluster, and determine which photo should be put into which cell.

Conceptually, we want to find the “best match” between templates and the given photo cluster. More “representative” photos or more “important” photos should be allocated larger space. Therefore, we first define template-based and photo-based importance values to be the metrics for template determination.

- Template-based importance: as shown in Figure 4, each tiling template consists of at least one topic cell and several supportive cells. Based on the ratio of the area of a cell over the whole frame, we calculate each cell’s importance. Importance values of the cells in the same template are then sorted in descending order.

- Photo-based importance: the photo-based importance is calculated based on "face region" and "attention value." A linear weighting method is applied to combine these two features and derive the photo-based importance [8]. The calculated photo-based importance is also sorted in descending order.

On the basis of these importance values, the "best match" between templates and photos can be determined by finding the pair that has the most similar importance distribution. In this system, we try to accomplish this work from two perspectives: vector angle and relative entropy.

In the vector angle approach, we respectively pack template-based and photo-based importance values into vectors. Given a set of k-cell templates, $\Gamma=\{T_{1,}\ T_2,\ldots,\ T_s\}$, a cluster that contains k photos should be mapped to one of these s templates. The included angle between the photo-based importance vector PV and a template importance vector TV_i is defined as the metric of template determination:

$$i^* = \arg\min_{i=1,2,\ldots,s} \operatorname{acos}\left(\frac{PV \cdot TV_i}{\| PV \| \| TV_i \|}\right), \tag{1}$$

where TV_i is the corresponding template-based importance vector of the template T_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. Through this process, more important photos would be put into larger cells.

On the other hand, we can construct probability mass functions to describe the distribution of template importance (P) and photo-based importance (Q). From this viewpoint, the symmetric Kullback-Leibler (KL) distance between photo-based importance distribution and template importance distribution can also be defined as the metric of template determination:

$$i^* = \arg\min_{i=1,2,\ldots,s} \left(D\left(P_i \| Q\right) + D\left(Q \| P_i\right)\right), \tag{2}$$

where P_i is the distribution of the ith template, and $D(\cdot\|\cdot)$ is the KL distance.

These two approaches both determine templates based on the concept of importance distributions, and the final results don't pose significantly difference in human's sense. Therefore, we mainly use vector angle approach in current implementation.

4.3 Spatial Composition (for Challenge 3)

The final task to generate a tiling frame is to put photos into the designated cells. However, the aspect ratio of the targeted cell is often different from that of the original 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). Therefore, it's unavoidable that we should resize and/or crop photos to fit in with the template.

In order not to largely distort the content of each photo, we want to find a region that has the same aspect ratio as the targeted cell and possesses the largest attractive content. As described in Section 2.2.3, we can find region of interest through top-down or bottom-up approaches. 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. The selected region is then resized to stick on the targeted cell. Likewise, the only difference in bottom-up cases is that we start expansion from the centroid of salience-based ROI.

Through the processes of cluster selection, template determination, and spatial composition, we can construct photographic paragraphs. After determining the timing for displaying a photo or switching frames, photographic paragraphs are concatenated and a tiling slideshow is finally generated. To facilitate more gorgeous presentation, we also include transition effects such as fade-in and fade-out in displaying.

We released the executable program of this system, and provide some examples at http://www.cmlab.csie.ntu.edu.tw/~wtchu/TilingSlideshow.

5. EVALUATION

We evaluate the proposed approaches from both objective and subjective perspectives. Five photo sets taken by different amateurs are used for evaluation, as listed in Table 1. Two of them were taken in progress of travel, two were taken in special events such as wedding and graduate ceremonies, and one of them contains totally landscapes. Table 1 also shows the length of the incidental music for the final presentation. These music clips are all pop music. We use the same music clip for photo sets 2, 4, and 5 to show that different results would be made due to visual diversity.

5.1 Clustering Performance

In this experiment, we invite the owners of the evaluated photos and ask them to judge whether the photos at the same frame belong to the same event or scenic spot. After the judgment of the content owners, very few frames consist of ill-clustered photos, as shown in Table 2. The result of the fifth photo set is slightly worse than others because of significantly diverse luminance conditions. In this case, the content-based clustering method may erroneously separate photos that should be clustered together. Even so, the "ill-clustered photos" are often placed in supportive cells rather than the conspicuous topic cell. We also show that due to different taking habits, the average number of photos at the same frame (cluster) varies for different data sets.

5.2 Cropping Performance

To evaluate cropping performance, we check each photo in frames and judge whether it's ill-cropped or not. The experimental results are shown in Table 3, where we can see very few photos are ill-cropped. We also check whether ill-cropped photos are placed in the topic cell, because ill-cropping in it brings viewers more negative effects. Generally, people are sensitive to erroneous cropping on human faces. Therefore, there is higher probability to sense a cropping error in photo sets 3 and 4, in which people are the major targets.

Two ill-cropped examples are shown in Figure 5. The reason for ill-cropping may be from non-robust face detection, like the side-view face in Figure 5(a). Furthermore, significantly different aspect ratio between the targeted cell and the important object may cause crude cropping to the original photo, like the case in Figure 5(b).

5.3 User Study

For subjective evaluation, we compare user satisfaction of the slideshows generated by the proposed system, ACDSee, and Photo Story. ACDSee generates conventional slideshow by sequentially displaying photos one-by-one. 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. They are asked to judge the satisfaction of different results according to their subjective perception. Generally, the titling slideshow has significantly better satisfactions than others. People feel that tiling slideshow provides more impressive presentation and easily helps them experience the content conveyed by photos. Detail experimental settings and results please refer to [8].

6. Discussion

On the basis of the proposed idea, various issues and extensions can be investigated. We provide some discussions from different perspectives.

6. 1 Influence of User Intervention

One of the major factors to affect subjective satisfaction is user preference. One may prefer to always put a specific person’s photo on the topic cell. In addition, some photos may be significantly valuable to someone even if they are blurred. These kinds of subjective preference or specific consideration are not taken in the fully-automatic process. We currently provide different profiles so that users can select whether to perform photo filtering, select the granularity of photo organization, and control the pace of frame switching. This flexibility somewhat provides a method for users to generate personalized tiling slideshows.

6.2 Photo Organization

On the basis of photo characteristics, we may have different clustering requirements at the organization stage. For the photos that have strict temporal order, such as travel, wedding, and graduate ceremony, we suggest that time-based clustering should be applied first to find social event boundaries. Content-based clustering is then applied to each time-based cluster for finer organization. In this way, the final presentation would follow the timeline and reproduce the progress of the targeted photo set. On the other hand, if the targeted data are photos in daily life or are simply landscapes, we can apply content-based clustering first to collect photos with similar appearance together.

6.3 Impacts from High-Level Concept Detection

Some erroneous results, such as ill-cropped or ill-clustered cases, would derive from the semantic gap between low-level features and high-level concepts. Because we only exploit content-based features in organization and manipulation, main objects or side-view faces would be crudely cropped (c.f. Figure 5), and photos that should be clustered together in human sense would be separated. To generate more elaborate results, we can appeal to semantic concept detection or image annotations that widely used in photo sharing community.

High-level concept detection also impacts on the arrangement of a frame. For example, a photo with "sky" on the top would be favorably put upper than the one without it. A photo with apparently vertical structure, such as a rise high building, should be put in a vertical-bar cell rather than horizontal-bar cell. Taking account of these considerations would enhance visual coordination.

6.4 Extension

Tiling slideshow can be adopted to build many interesting applications. For example, given a traveling schedule and its corresponding photos, a photo-based tour can be made. Similar ideas can also be applied to generate customized auto-guidance or electronic lecturing. Moreover, there are still many research issues if we take textual information into account. We can allocate some space for text and coordinate more diverse media (text, photo, music). Finally, although tiling slideshow is currently presented as a video clip, different kinds of visualization can be made in the future. We can just output a text-based script that describes the involved photos and their occurrence timestamps and locations. This script is read by a multimedia player such as Flash and is then visualized.

7. CONCLUSION

The proposed system automatically generates an audiovisual presentation, which provides a new browsing experience for consumer photos. Photos are first examined by quality estimation, and that with defects are filtered out. The remaining photos are then organized according to temporal and spatial contexts. Photos in the same cluster are usually taken at the same scene or represent the same event, and will be presented at the same frame. For a cluster of photos, the cluster-based importance and photo-based importance are estimated, which are the metrics for cluster selection and smart photo manipulation. For the user-indicated music, we detect beat information and use it as the cues to determine the timing of photo presentation. Through elaborate composition process, photos are smartly manipulated to construct vivid presentation. The objective evaluation demonstrates high accuracy of clustering and cropping, while the user study shows that tiling slideshows have superior acceptance over conventional slideshows.

We can enhance the work from different viewpoints, such as enhancing content-based organization from content analysis perspective or elaborating spatial arrangement and audio-video synchronization from media aesthetics. Furthermore, user's preference is always an important factor to construct an attractive presentation. We would put more efforts on how to easily and effectively cooperate the proposed techniques with user preference in the near future.

8. REFERENCES

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

[2] 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.

[3] Kustanowitz, J., and Shneiderman, B. Hierarchical layouts for photo libraries. *IEEE Multimedia*, vol. 13, no. 4, pp. 62-72, 2006.

[4] 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.

[5] 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.

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

[7] 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.

[8] Chen, J.-C., Chu, W.-T., Kuo, J.-H., Weng, C.-Y., and Wu, J.-L. Tiling Slideshow. *Proceedings of ACM Multimedia Conference*, pp. 25-35, 2006.

[9] 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.

[10] 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.

[11] White, J.V. Editing by Design: A Guide to Effective Word and Picture Communication for Editors and Designers. R. R. Bowker Co., 1982.

Authors

Wei-Ta Chu received Ph.D. degree in computer science from National Taiwan University (NTU), 2006. He is now a postdoctoral research fellow in NTU. His research interests include content analysis and multimedia indexing. He was awarded as the best technical full paper in ACM Multimedia Conference 2006, and received the best Ph.D. thesis award from Institute of Information & Computing Machinery.

Jun-Cheng Chen received his master degree in computer science from National Taiwan University, 2006. His research interests include video processing and multimedia systems. He was awarded as the best technical full paper in ACM Multimedia Conference 2006.

Ja-Ling Wu is the head of Graduate Institute of Networking and Multimedia in NTU**.** His research interests include DSP, image/video compression, content analysis, digital watermarking, and DRM systems. Prof. Wu was the recipient of many outstanding awards in DSP algorithm designs, digital watermarking, and content analysis. He was elected to be the lifetime distinguished professor NTU, Oct. 2006.

Readers may contact authors at wtchu@cmlab.csie.ntu.edu.tw, pullpull@cmlab.csie.ntu.edu.tw, and wjl@cmlab.csie.ntu.edu.tw, respectively.

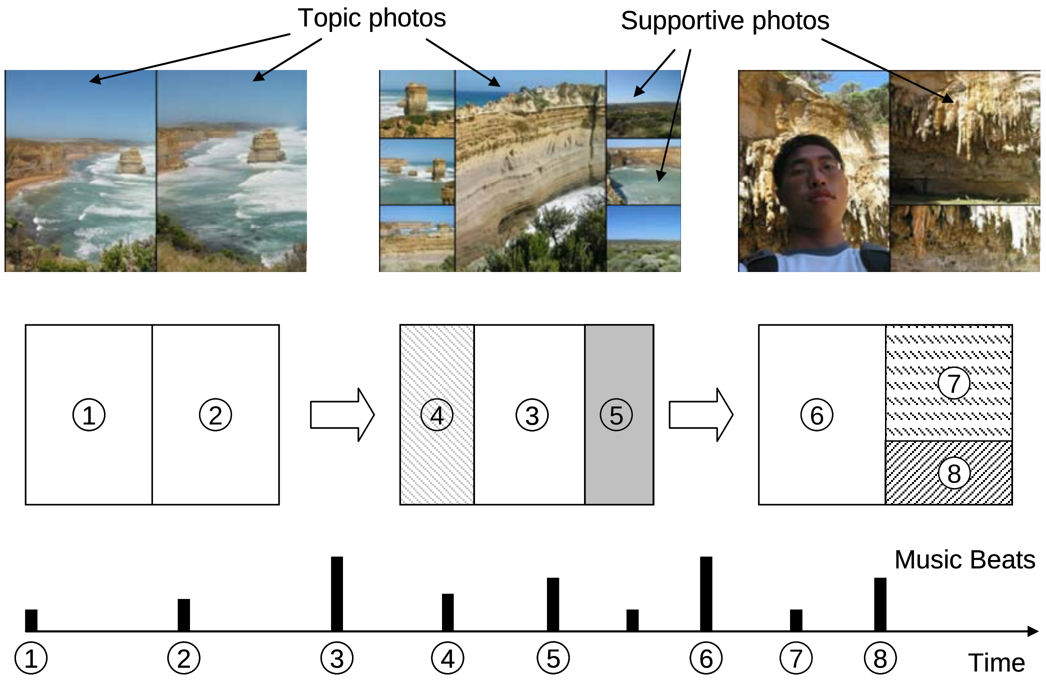

Figure 1. An example of tiling slideshow.

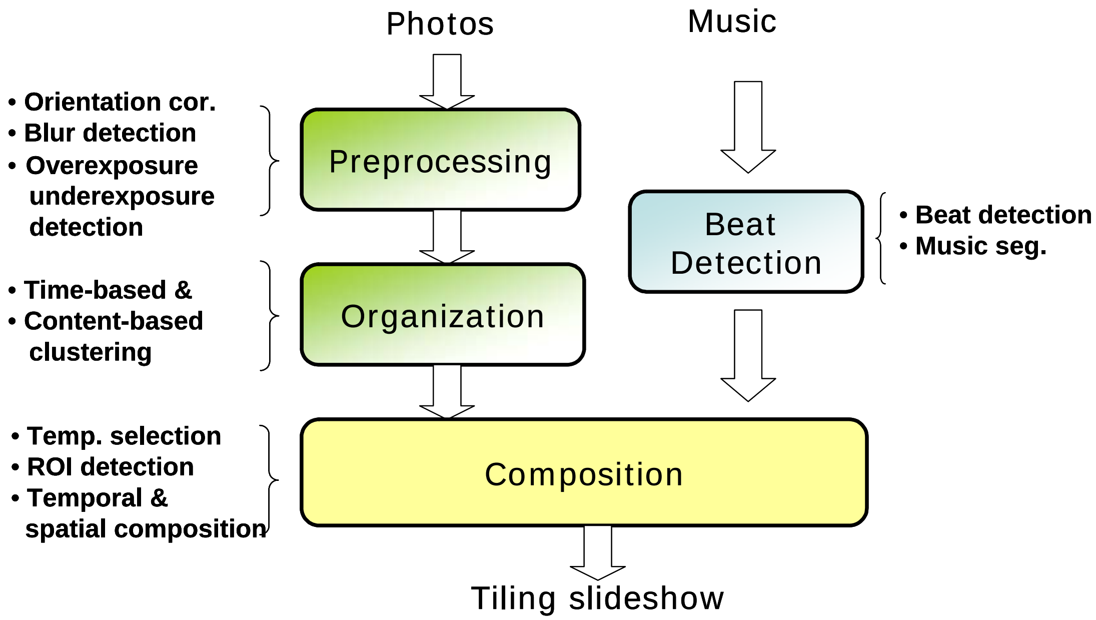

Figure 2. System flowchart of the proposed titling slideshow.

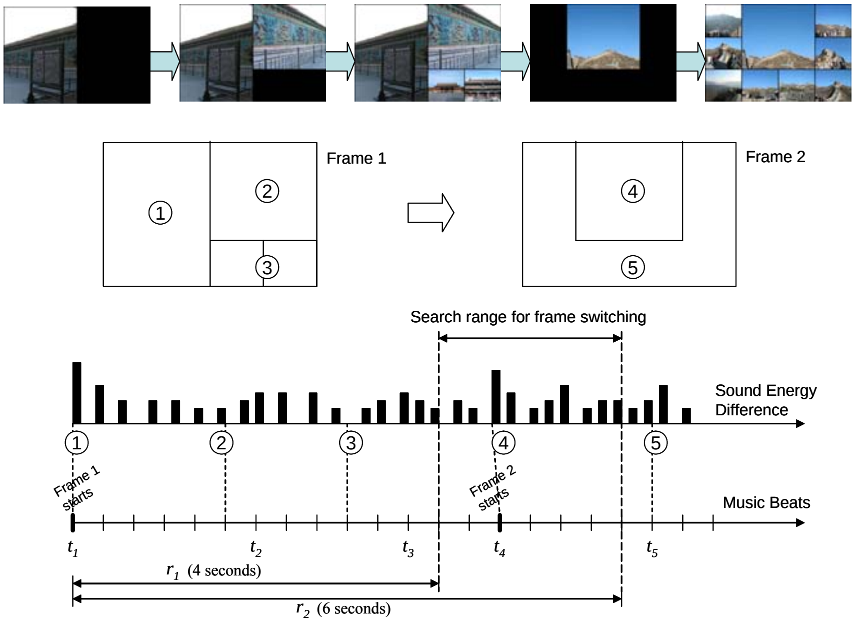

Figure 3. An example of determining the timing for frame switching and photo display.

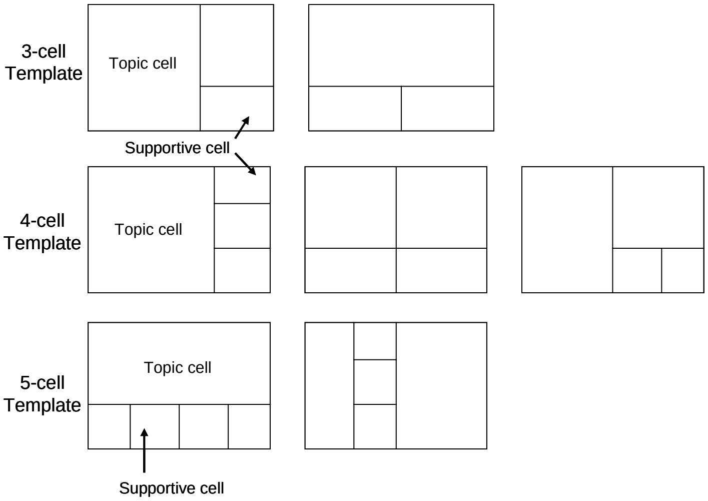

Figure 4. Examples of different kinds of templates.

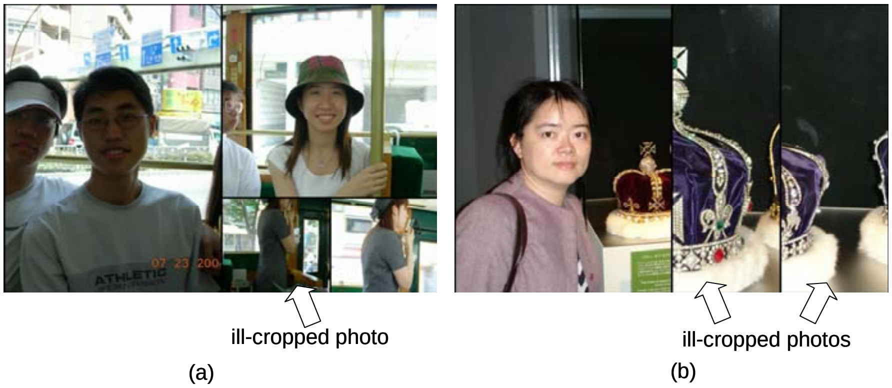

Figure 5. Examples of ill-cropped photos.

Table 1. Information of the evaluation photo sets.

	Type	# photos	Descriptions	Length of incidental music
Set 1	Travel	780	Traveling in Japan. Including people, cityscape, and landscape.	3’31’’
Set 2	Travel	522	Traveling in Australia. Including people, cityscape, and landscape.	4’38’’
Set 3	Wedding	388	A Chinese wedding ceremony. People are the main targets.	3’49’’
Set 4	Graduate ceremony	227	A graduate ceremony. People are the main targets.	4’38’’
Set 5	Landscape	133	Pure landscape in Taiwan, including mountain, river, greenwood.	4’38’’

Table 2. Clustering performance evaluation.

	#frames	# photos	# frame with clustering error	Avg. number of photos in a frame
Slideshow 1	37	127	1	3.43
Slideshow 2	48	172	1	3.58
Slideshow 3	38	155	0	4.08
Slideshow 4	48	212	0	4.41
Slideshow 5	48	131	8	2.73

Table 3. Cropping performance evaluation.

	# photos	# ill-cropped photos	# ill-cropped photos in topic cell
Slideshow 1	127	5	1
Slideshow 2	172	5	0
Slideshow 3	155	8	6
Slideshow 4	212	9	5
Slideshow 5	131	2	0