Photo2Video

Xian-Sheng HUA, Lie LU, Hong-Jiang ZHANG
Microsoft Research Asia
5F, Beijing Sigma Center, No.49 Zhichun Road, Beijing 100080, P.R.China
{xshua;llu;hjzhang}@microsoft.com

ABSTRACT
To exploit rich content embedded in a single photograph, a system named *Photo2Video* was developed to automatically convert a photographic series into a video by simulating camera motions, set to incidental music of the user's choice. For a chosen photographic series, an appropriate camera *motion pattern* is selected for each photograph to generate a corresponding *motion photograph clip*. Then, the final output video is rendered by connecting a series of motion photograph clips with specific transitions, and aligning with the selected incidental music. *Photo2Video* provides a novel way to browse a series of images and can be regarded as a system exploring the new medium between photograph and video.

Categories and Subject Descriptors
H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—*animations*; I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—*representations*.

General Terms
Algorithms, Experimentation.

Keywords
Image content analysis, face detection, attention detection, image clustering, audio segmentation, optimization, integer programming.

1. INTRODUCTION
Though static and two dimensional, a single photograph contains extremely rich content. When we view a photograph, we often look at it with more attention to specific objects or areas of interest after our initial glance at the overall image. In other words, viewing photographs is a temporal process which brings enjoyment from inciting memory or from rediscovery. That is, a single photograph may be converted into a *motion photograph clip* by simulating temporal variation of viewer's attention using simulated camera motions. For example, zooming simulates the viewer looking into the details of a certain area of an image, while panning simulates scanning through several important areas of the photograph. Connecting the motion photograph clips following certain editing rules forms a slide show in this style, a video which is much more compelling than the original images. Such a video composed from motion photograph clips is a new medium that captures the story-telling of image viewing/sharing and enhances the enjoyment of the photograph viewing process. In this demonstration, we present a system named *Photo2Video* developed to automatically convert photographs into video by simulating temporal variation of people's study of photographic images using simulated camera motions.

MM'03, November 2-8, 2003, Berkeley, California, USA.
ACM 1-58113-722-2/03/0011.

2. SYSTEM OVERVIEW
As illustrated in Figure 1, the *Photo2Video* system consists of three main components: motion photograph clip generation, music segmentation, and final rendering. The primary component is motion photograph clip generation, which includes focus detection, motion pattern determination, and motion generation.

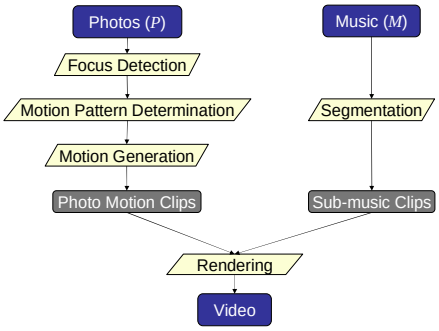

Figure 1. Flow chart of Photo2Video.

Focuses are areas in a photograph that most likely will attract a viewer's attention or focus, which will be used to determine the simulated camera motions. In *Photo2Video*, the focuses in a photograph include dominated faces detected using a robust face detection algorithm [2], and visual attention areas detected by using a contrast-based attention model [3].

In motion generation, motion trajectory and duration for showing a picture is determined based on the detected focuses in the image. Based on the trajectory and duration, a series of *motion rectangles* for the original photograph are generated along the trajectory. Video frames of the motion photograph clip are then constructed by resizing the sub-photographs defined by these rectangles.

The incidental music is segmented into *sub-clips* by detecting strong beats. Each music sub-clip is used for one photograph. That is, the display duration of one motion photograph clip is equal to the duration of the corresponding music sub-clip. Also, transitions between two motion photograph clips occur at the music beats such that motion photograph clip boundaries and music sub-clip boundaries are aligned in the final rendering process. Additionally, the transition effect between two motion photograph clips is determined using the content of the two photographs [1].

To give the output video structural information, photographs are clustered into several groups based on timestamp and/or visual similarity. The first photograph of each cluster is selected to represent the corresponding cluster. A so-called *storyboard*

photograph is generated by patching the thumbnails of these selected photographs into one image. We insert this storyboard photograph before the first photograph of each cluster. The motion pattern of this storyboard photograph in each cluster is fixed as "zooming into the corresponding thumbnail to be shown"

The reminder of this paper describes, in more detail, the definition of a number of typical motion patterns, motion pattern selection schemes for a particular photograph and the entire photographic series, and motion generation.

3. MOTION PATTERN DETERMINATION

In *Photo2Video*, motions through a particular image are generated according to the detected focuses. By observing professionally edited videos, such as scenery programs, it shows that there are two primary motion patterns (MPs): panning and zooming, which are also often used in combination to generate more expressive motion patterns. Based on this observation, eight basic motion patterns are defined, including *Still*, *Light Zooming-in*, *Light-Zooming-out*, *Panning*, *Zooming-in*, *Zooming-out*, *Panning with Zooming-in*, *and Panning with Zooming-out*. Although reasonable combinations of the eight basic patterns, such as "zooming in – panning" and "panning – still – zooming out", are also feasible as compound motion patterns, in this demonstration, we only consider the eight basic motion patterns.

Selecting appropriate *motion patterns* to generate a motion photograph clip for an image with given focus areas is a non-trial task. A particular photograph may have several acceptable motion patterns, such as zooming into a focus, or panning from one focus to another. However, if most of the photographs in a series have the same or similar motion patterns, the final video will be too monotonous.

In general, the manner in which a user studies a picture is strongly related to the focuses, and in particular, the number of focuses vying for the user's attention. For example, a photograph with a single focus most likely will lead the viewer to concentrate on the details of that focus after the first glance of the entire photograph. Additionally, some other semantic features as whether the photograph is taken indoors or outdoors, and whether the photograph is landscape or cityscape, also provide us cues to select appropriate motion pattern for a particular photograph. For example, outdoor and landscape photographs have more types of acceptable motion patterns than indoor or cityscape photographs. Accordingly, in order to choose appropriate motion patterns for a particular photograph and the entire photographic series, an *MP suitability matrix* $M = (m_{ij})_{8\times 8}$ is defined according to the number of detected focuses in a photograph and some semantic features.

Suppose θ is a solution of MP selection for all photographs, then the set of all feasible solutions is denoted by Θ. For a series of photographs, a straightforward scheme to assign motion patterns is maximizing the overall motion suitability ($S(\theta)$) on all photographs, i.e., to determine

$$\theta^* = \arg\max_{\theta\in\Theta} S(\theta) \tag{1}$$

However, in order to avoid monotony, another two constrains should also be considered. One is distribution uniformly of every motion pattern (H_j), and the matching degree of actual appearing rates and desired appearing rates of the motion pattern (R).

Consequently, motion pattern selection is formulated as the constrained integer programming program shown in Equation 2.

$$\theta^* = \arg\max_{\theta\in\Theta} S(\theta) \quad \text{subject to } H_j(\theta) \ge h_j, R(\theta) \ge r_0 \tag{2}$$

In the above equation r_0, and h_j, $0 \le j < N$, are predefined target thresholds. Genetic Algorithm is applied to find solutions approaching the global optimum.

4. MOTION GENERATION

Motion patterns involving panning require defining the trajectories. It is assumed that the motion trajectory of a simulated camera is a curve that connects all the targeted focuses. For example, if a motion pattern is a panning from focus F_1 to F_0, the center trajectory may be the straight line segment $\overline{F_1F_0}$. In addition, panning speed and zooming speed along a trajectory are also necessary to completely define a motion pattern.

In *Photo2Video*, if there are more than two focuses, the panning trajectory is defined by an open interpolating curve across the centers of all the focuses. However, there are many possible solutions for such a curve. Generally, when shooting with real video camera, the trajectory will not vary quickly, but move smoothly and steadily. In *Photo2Video*, the *cubic interpolating spline* with the smallest *maximal curvature* (maximal value of the curvature along the trajectory) is used as the panning trajectory.

Speed control is applied to determine the panning/zooming speed along the trajectory. Typically real camera motion will be slower at the beginning and the end, called ease-in/ease-out. In *Photo2Video*, a uniquely predefined *speed control function* is applied to set the speed at any position. Panning/zooming speed is set proportional to the speed control function.

5. DISCUSSION AND FUTURE WORK

Photo2Video also can be regarded as a system exploring the new medium of motion picture style slide shows. In fact, all motion patterns determined for each photograph in a series can be recorded in a script file, which can then be used to render a video with suitable incidental music at the user's demand. As *Photo2Video* generates motion photographs in a fully automatic manner, it is convenient to adopt this system for many applications, such as creating automatic walkthroughs of photograph galleries, motion photographs on website, electronic greeting cards and personalized Karaoke. Our next step work will also explore how to better automatically or semi-automatically expose the semantics/stories behind the photographic series based on the basic idea of *Photo2Video*. A work moving one-step ahead would be integrating face recognition or face annotation into *Photo2Video*, thus the video will have a "leading actor".

6. REFERENCES

[1] Hua, X.S., Lu, L. and Zhang, H.J. Content-Based Photograph Slide Show with Incidental Music. *Proc. of ISCAS 2003. Vol. II, pp. 648 -651, 2003*

[2] Li, S.Z. et al. Statistical Learning of Multi-View Face Detection. *Proc. of ECCV 2002*.

[3] Y.F., and Zhang, H.J. Contrast-based Image Attention Analysis by Using Fuzzy Growing. *Proc. of ACM Multimedia 2003.*