Attention-Oriented Photo Slideshow

Hsu Hsu[*] Sheng-Jie Luo[*] Bing-Yu Chen[†]

National Taiwan University

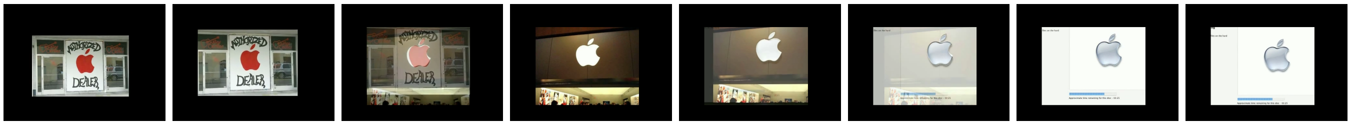

Figure 1: *An example frame transition of our slideshow system (from left to right).*

1 Introduction

Photo slideshow is a popular method to demonstrate photo sets to a large number of people. It often covers a main topic, and contains the related characters or objects to tell people the story. For example, a wedding slideshow contains the photos of the groom and the bride. Audiences' attentions would be attracted by these objects. However, a photo may contains rich information, audiences' attentions could be distracted when watching a photo slideshow. Therefore, to produce a slideshow which specifically attracts audiences' attention toward the main characters or objects is challenging. In this extended abstract, we present the attention-oriented photo slideshow to address this challenge. Based on the observation that by smoothly connecting similar objects or characters (i.e., common object) in a photo slideshow, audiences' attentions could be effectively attracted toward them, our idea is to emphasize these objects or characters by smoothly connecting and transiting them between images. With our slideshow, audiences are more attracted to common objects that repetitively appear in the photo sets.

2 Attention-Oriented Photo Slideshow

Psychological studies about human perception have shown that visual signals contrast such as motion, color, and object size are likely to attract people's visual attentions [Wolfe and Horowitz 2004]. According to the study, our system adjusts the main objects' sizes and transition paths smoothly to attract the users' attention. The system consists of three main stages: *common object extraction*, *object path planning*, and *frame rendering*.

Common object extraction. Given a set of photos $\mathbf{P} = \{P_i\}_{i=1\ldots N}$ in an user specified order and an example object image I, our system first exploits the local self-similarity descriptors [Shechtman and Irani 2007] to extract the common objects $\mathbf{C} = \{C_i\}_{i=1\ldots N}$ from $\mathbf{P}$ that have similar shapes with the example object, where $C_i = (\mathbf{p}_{c_i}, s_{c_i})$ is the common object of a photo P_i, and $\mathbf{p}_{c_i}$ stands for the location of the object and s_{c_i} is its relative scale to the example object in I. The object scale is calculated as $s_{c_i} = \frac{1}{|\mathbf{D}|}\sum_{d\in\mathbf{D}} |\mathbf{p}_d - \mathbf{p}_{c_i}|/|\mathbf{p}_{d'} - \mathbf{p}_{c_i}|$, where $\mathbf{D}$ refers to the matched descriptor set, d is a descriptor in $\mathbf{D}$, d' is the matched descriptor of d in I, $\mathbf{p}_d$ and $\mathbf{p}_{d'}$ stands for d's and d''s locations in image respectively.

Common object path planning. The transitions of these extracted common objects are planned so that the slideshow provides smooth watching experience. The input photos $\mathbf{P}$ are regarded as keyframes of the slideshow. Then we synthesize and interpolate a set of frames between neighboring keyframes, and obtain a series of frames $\mathbf{F} = \{F_i\}_{i=1\ldots k}$. In this step, the positions and scales of all the frames are calculated such that the defined energy function is minimized.

Data term. This term is designed to encourage keyframes to not exceed the screen and also to maintain their original sizes. Denote by $\{\mathbf{v}_j\}_{j=0,1,2,3}$ the four corner points of a keyframe, we encourage that $\tilde{\mathbf{v}}_j = \bar{\mathbf{v}}_j$, where $\{\tilde{\mathbf{v}}_j\}_{j=0,1,2,3}$ are the optimal locations of the four corner points to be solved, and $\{\bar{\mathbf{v}}_j\}_{j=0,1,2,3}$ are the corresponding locations when placing the keyframe at the center. The corner points of each scaled keyframe P_i can be expressed as $\tilde{\mathbf{v}}^i_j = \mathbf{p}_{c_i} + \tilde{s}_i \Delta\mathbf{v}^i_j$, where $\tilde{s}_i$ is the scale factor of the keyframe. Therefore, the data term is defined as

$$E_d = \frac{1}{|\mathbf{P}|} \sum_{P_i \in \mathbf{P}} \sum_{j=0}^{3} \left\| \tilde{\mathbf{p}}_{c_i} + \tilde{s}_i \Delta\mathbf{v}^i_j - \bar{\mathbf{v}}^i_j \right\|^2 . \quad (1)$$

Smoothness term. The purpose of smoothness term is to control the path smoothness of the common objects; i.e., the location and scale of the common objects should change smoothly. In order to achieve this, we make the first order differentiation of the transition path equals to zero, and obtain

$$E_s = \frac{1}{|\mathbf{F}|-2} \sum_{t=2}^{|\mathbf{F}|-1} \left(\begin{array}{c} \left\| \tilde{\mathbf{p}}^{F_{t-1}}_c - 2\tilde{\mathbf{p}}^{F_t}_c + \tilde{\mathbf{p}}^{F_{t+1}}_c \right\|^2 + \\ \left\| \tilde{s}^{F_{t-1}}_c - 2\tilde{s}^{F_t}_c + \tilde{s}^{F_{t+1}}_c \right\|^2 \end{array} \right), \quad (2)$$

where $\tilde{\mathbf{p}}^{F_t}_c$ and $\tilde{s}^{F_t}_c$ are the location and relative scale of the common object on the screen in the t-th frame. The scale is computed as $\tilde{s}^{F_t}_c = s^{F_t}_c \times \tilde{s}_t$, where $s^{F_t}_c$ denotes the relative scale of the common object in the original photo, and $\tilde{s}_t$ is the scale factor of F_t.

The total energy function is then defined as $E = \lambda_d E_d + \lambda_s E_s$, where $\lambda_d = 1$ and $\lambda_s = 500$. We solve for the the unknown locations of the common objects in frames and the scale factors of all frames to minimize the energy function.

Frame rendering. The final step is to render all frames into the slideshow video. The interpolated frames are generated by smoothly blending the neighboring keyframes. Formally, an interpolated frame F_t between two keyframes P_i and P_{i+1} are obtained by $F_t = w\mathbf{T}_{F_t}(P_i) + (1-w)\mathbf{T}_{F_t}(P_{i+1})$, where w is the linear blending weight, and $\mathbf{T}_F(P)$ denotes the transformation operator that transforms the common object of P to match that of F.

3 Conclusion and Future Work

Photo slideshow has become a popular way to demonstrate a set of photos to a large number of people. When using our method to convey sequences of photos, users tend to move their attention from the whole photos to the main characters and objects. In the future, we would improve the usability of the presentation method and extend the system to handle multiple common objects.

References

Shechtman, E., and Irani, M. 2007. Matching local self-similarities across images and videos. *IEEE CVPR*.

Wolfe, J. M., and Horowitz, T. S. 2004. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience 5*, 6, 495–501.

[*] e-mail:{r98922001,forestking}@cmlab.csie.ntu.edu.tw

[†] e-mail:robin@ntu.edu.tw