

# VISUAL ATTENTION BASED IMAGE BROWSING ON MOBILE DEVICES

Xin Fan<sup>1\*</sup>, Xing Xie<sup>2</sup>, Wei-Ying Ma<sup>2</sup>, Hong-Jiang Zhang<sup>2</sup>, He-Qin Zhou<sup>1</sup>

<sup>2</sup>Microsoft Research Asia, 3/f Sigma Center, No. 49, Zhichun Road, Beijing, 100080, P.R. China

<sup>1</sup>Dept. of Automation, Univ. of Sci. and Tech. of China, Hefei, 230027, P.R. China

## ABSTRACT

Images have become more and more common in mobile communications. People now can easily take and exchange pictures on the move using their mobile devices and digital cameras. However, a crucial challenge is to provide a better user experience for browsing large images on limited and heterogeneous screen sizes of mobile devices. In this paper, we propose a novel image viewing technique based on an adaptive attention shifting model. A presentation technique named *Rapid Serial Visual Presentation (RSVP)*, borrowed from the UI community, is used to simulate the attention shifting process. We show a prototype image viewer developed for Pocket PC and conduct some evaluations to demonstrate the effectiveness of our approach.

## 1. INTRODUCTION

Mobile devices are undergoing considerable progress during recent years. People now can stay connected with their family and friends while they are apart, sharing and expressing their lives and personal experiences. In these applications, images are playing a more and more important role. However, to make people really enjoy the ease of mobile communications, many hurdles still need to be crossed. Among them, major crucial challenges include the limited accessing bandwidth and display sizes of mobile devices [1]. Thanks to the galloping development of both hardware and software, the bandwidth condition is expected to be greatly improved in the near future. However, in the foreseeable future, the display, i.e. the form factor, will continue to be the major constraint on the small mobile devices such as cell-phones and handheld PCs. In this paper, we will focus on facilitating image viewing on devices with limited display sizes.

Image adaptation is not a new problem. Image transcoding based on the classification of image types and purposes has been introduced in many previous works [1] [2]. From the system side, a transcoding proxy [3] was presented for determining when/whether/how to adapt

images according to bandwidth, file size, etc. Although there have been already many approaches for adapting images, most of them focus on compressing and caching contents in order to reduce data transmission, speed up delivery or minimize energy consumption [4]. User perception and the characteristic of human visual system have received little attention in these studies.

It has become clear that not all but only a small part of incoming visual information can reach short-term human memory for further processing, i.e., *the Attention as Filter Metaphor* [5]. Attentional selection allows only attention-getting parts be presented to the user without affecting much user experience. In our prior work [6], the most important part of a large image is cropped to fit the limited display size. However, much other information which a user cares may be lost by this approach. Based on this consideration, when viewing images on small displays, we propose to employ a widely-used presentation technique, Rapid Serial Visual Presentation (RSVP), in which space is traded for time [7]. In RSVP, amounts of contents are displayed serially, each for a brief period of time, to aid users' browsing or searching through the whole contents.

RSVP of text has been studied extensively but that of image is deficiently attended. We notice that there is an important psychophysiological activity – visual attention shifting. Visual attention can rapidly direct and shift the gaze towards interesting parts of the visual input by combining bottom-up, image-based cues and top-down, task-dependent cues [8]. Image browsing on small devices can be improved by simulating the fixation and shifting process in a way similar to RSVP. Based on our previous algorithm [6], we adapt the images by a selective attention model counting in both the bottom-up and top-down cues. A RSVP simulation is designed according to the attention shifting process as well as other visual characteristics. An image viewer is implemented on a Pocket PC to validate the performance of our schemes.

The rest of this paper is organized as follows: Section 2 introduces the attention model for adapting images to the limited display area. The simulation of the attention shifting process is presented in Section 3. Section 4 shows the implementation of our schemes and its evaluation results. Finally, Section 5 gives our conclusions and discussions on future work.

---

\* This work was conducted while the first author was a visiting student at Microsoft Research Asia.

## 2. SELECTIVE ATTENTION MODEL FOR IMAGE ADAPTATION

There are still a lot of debates on the mechanics of visual attention. Many models (bottom-up vs. top-down, space-based vs. object-based) have been proposed to depict the visual attention activities. Majority of them, such as [9], only depend on bottom-up algorithms while top-down biasing can also be used in the formation of low-level features [10]. In other works, the attention is almost exclusively under top-down control [11]. It is gradually deemed that a complete computational model would be the integration of these twofold cues. Furthermore, biologically-plausible study describes that the saliency-based attention is only deployed during the first few decades of milliseconds after the presentation of a new scene while volitional deployment of attention takes several hundreds [8]. Noteworthy, it has been validated for a long time that the attention commonly fixates on the most informative regions [12].

### 2.1. Selective attention model

In this section, we define a selective attention model to reveal the informative regions in an image. A block-assembly and partly object-driven solution is deployed to capture the information in a presented scene. On one hand, to contain more information, the image is sub-sampled by reducing the spatial size for specified display area. On the other hand, the sub-sampling is conditioned on avoiding the loss of information resulting from excessive reduction.

As shown in Definition 1, in our model, a set of information carriers – *attention objects* (*AOs*) are defined, which often represent semantic objects, such as a human face, a flower, a text notation, etc. However, as mentioned above, the scene perceiving usually involves more than a simple collection of semantic objects, purely bottom-up saliency regions are also taken into account.

**Definition 1:**  $\{AO_i\} = \{(ROI_i, AV_i, MPS_i)\}$ ,  $1 \leq i \leq N$  (1)

where

- $AO_i$ , the  $i$ th *attention object* within the image
- $ROI_i$ , Region-Of-Interest of  $AO_i$
- $AV_i$ , attention value of  $AO_i$
- $MPS_i$ , minimal perceptible size of  $AO_i$
- $N$ , total number of attention objects in the image

The notion of Region-Of-Interest is borrowed from JPEG2000. It is referred in our model as a spatial region corresponding to an *AO*. *ROIs* can be arbitrary shapes and allowed to overlap.

Referring to human cognitive systems, the *AOs* carry different amounts of information, i.e., are of different importance. The attention value indicates the weight of

each *AO* in contribution to the information contained in the original image.

As to semantic objects (such as face or text), the information carried is significantly relies on the area of presentation. Therefore, the *minimal perceptible size* (*MPS*) is introduced as a threshold to avoid excessively sub-sampling during reduction of display size.

### 2.2. Automatic attention modeling

Three types of attention objects are taken into account in our model: saliency, face and text.

Early stages of attention processing are deployed by ensemble of low-level features. A two-dimensional topographical “saliency map” [9] is used to determine the attention-getting region within the original image. Similar to [13], we binarize the saliency map to find the *ROIs*. The *AV* can be calculated as:

$$AV_{saliency} = \sum_{(i,j \in R)} B_{i,j} \cdot W_{saliency}^{i,j} \quad (2)$$

where  $B_{i,j}$  denotes the value of pixel  $(i,j)$  in the saliency map. Since people often pay more attention to the region near the center of an image, a normalized Gaussian template centered at the image is used to assign the position weight  $W_{saliency}^{i,j}$ . We use some heuristic rules to calculate the *MPS* of the salient region. For example, bigger regions can be scaled down more aggressively than smaller ones.

Human face and text are important semantic factors which will guide the attention and can be precisely detected now. We acquire the face information including the pose, region and position within an image. The importance of a face is usually reflected by its region size and position:

$$AV_{face} = Area_{face} \times W_{face}^{pos} \quad (3)$$

where  $Area_{face}$  denotes the area of the detected face region and  $W_{face}^{pos}$  refers to the weight of position in [13]. The *MPS* of the face is experientially predefined.

The informative text region can be found by applying a text detection module. The region size and the aspect ratio are involved to evaluate the importance due to the estimation that text headers or titles are often in an isolated single line with larger aspect ratios than text paragraph blocks. Its *AV* is defined as:

$$AV_{text} = Area_{text} \times W_{text}^{pos} \times W_{AspectRatio} \quad (4)$$

The *MPS* of a text region is also experientially predefined.

In order to combine above components in an effective and simple way, the *AV* of each *AO* is normalized to (0,1) and the final *AV* is computed as:

$$AV_i = w_k \cdot \overline{AV_i^k} \quad (5)$$

where  $\overline{AV_i^k}$  represents the normalized AV of  $AO_i$  and  $w_k$  is the weight of the component, which manifests the contribution during the attention-guiding function. Here we preliminarily deem that semantic objects play a more important role than “salient” regions.

Moreover, for images of different representing purposes, the importance of each component may be various. Therefore, different sets of  $w_k$  should be bestowed according to different scenes, for example, news pictures, home photos, sports shots or scenery pictures.

The automatic modeling results are structured in a form of XML descriptions and saved as the metadata within original images for reuse.

### 3. SIMULATION OF ATTENTION SHIFTING

To give a better user experience and seize the whole information, we employ RSVP to aid image browsing based on the study of visual attention shifting. Since it is very difficult to measure and describe the movement of human attention, only some indirect methods such as eye movements have been used for estimating the shifting of attention. Referring to the successful attentional robotics control, in which attention control is often described as a sequence of fixations and smooth pursuits, we approximatively depict the attention movement as two statuses: the fixation status and the saccade status. The iterations of these two statuses compose the whole simulation of the shifting process.

#### 3.1. Fixation status

The acquirement of the fixated region i.e., the most informative region, depends on a branch-and-bound algorithm in our prior work [6]. As shown in Figure 1, the output area can be of different a specified size which contains as much information as possible and ensures the perceptibility.

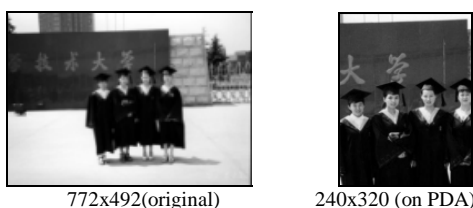


Figure 1. Example of fixated region selection and adaptation

We evaluate the duration of fixation status referring to the experimental results on RSVP rates in [7]. In consideration of the effect of different potential user tasks, such as search or comprehension, on the selective attention process, the duration is endowed with an experiential task-dependent value, for example, 0.8 sec for search and 1.4 sec for comprehension.

#### 3.2. Saccade status

The saccade status can be described as a shifting process from the most informative region to the second one, then the third and so on. A motivation of this process comes from a psychophysical phenomenon called “*inhibition-of-return*” [8], which demonstrates that the current attention focus will be suppressed while selecting the next focus. We implement it by removing the  $AOs$  contained in the current display area and applying the same algorithm to the rest  $AOs$  when selecting the next fixating area. The trace of the saccade is defined as the shortest path between centers of the two fixation areas. The whole status is semblable to equal-rate pan and zoom of the camera, as shown in Figure 2. Similar to the fixation status, the user task is considered to evaluate the duration of the shift process, which is also affected by the distance of the trajectory. Some improvements can also be made in the saccade status. For example, the shifting trace can be nonlinear in order to present more information and the moving speed can be adaptive to the content passed by.



Figure 2. Example of saccade status

The amount of the iterations lies on the complexity of the scene which can be estimated by the number of the  $AOs$  contained and the image resolution compared to the given display area. For most personal images, two iterations will cover all the  $AOs$ .

## 4. IMPLEMENTATION AND EXPERIMENTS

We have implemented an image viewer with two specific browsing functions, named animated view and thumbnail view, on a Compaq iPaq 3670 with Pocket PC 2002 as its OS. A user study is carried out on the prototype viewer to evaluate the performance of our solution.

#### 4.1. Prototype implementation

In the animated view mode, firstly a fit-to-window thumbnail of the image is presented to give the user a whole impression. Then the display will smoothly transit to the most informative region, and analogously the second (if exists) and return to the thumbnail again. Functionally, two sets of task-dependent duration of each

status, e.g., for searching or for comprehending, are available.

When the user looks through a photo album, a thumbnail view can facilitate the browsing while it is often too small to grasp the contents. In our prototype, as shown in Figure 3, the first fixated region mentioned above can be presented as the thumbnail with a customized size.



(a) Original thumbnails (b) Adaptive thumbnails  
Figure 3. Example of the visual attention based thumbnail view.

#### 4.2. Evaluation results

We chose 30 test images from some personal albums and popular websites, with image sizes varying from 614x461 to 800x600 pixels. They are of various classes, e.g., personal indoor or outdoor images, news pictures and scenery photos. Due to the deficient computational power of PDA, the image analysis is conducted on a desktop PC and the results are saved into the image files. 18 volunteers were invited to give their subjective scores at two aspects:

1. Do the fixated regions really present interesting areas?
2. Does our approach facilitate image browsing on a small screen?

The evaluation results are listed in Table 1 and Table 2, respectively.

Table 1. Evaluations for the quality of fixated regions

Class	Interesting	Uninteresting
Indoor	71.11%	28.89%
Outdoor	66.05%	33.95%
News	76.98%	23.02%
Scenery	65.43%	34.58%
<b>Average</b>	<b>69.89%</b>	<b>30.11%</b>

Table 2. Evaluations for the browsing improvement

Class	Better	No difference	Worse
Indoor	76.67%	21.11%	2.22%
Outdoor	74.69%	16.05%	9.26%
News	84.13%	11.11%	4.76%
Scenery	66.67%	24.07%	9.26%
<b>Average</b>	<b>75.54%</b>	<b>18.09%</b>	<b>6.38%</b>

As can be seen, most users prefer the new functions based on our technique. Our selective attention model is

effective especially to the Web news pictures and personal photos. Some of the subjects think our approach helpful even when the fixated regions are not attention-getting. This indicates that our approach is useful even when the attention modeling results are not accurate enough.

#### 5. CONCLUSIONS

In this paper, we proposed a novel solution for browsing image contents on small-form-factor devices based on visual attention. Most of existing work on image adaptation focused on saving file size while human visual characteristics are of deficient consideration. Our approach integrates both the bottom-up and top-down cues of the image. With the satisfactory results from our experiments, we plan to extend our work to other types of media such as video clips.

#### 7. REFERENCES

- [1] Ma W.Y., Bedner I., Chang G., Kuchinsky A., Zhang H.J., "A Framework for Adaptive Content Delivery in Heterogeneous Network Environments," Proc. SPIE Multimedia Computing and Networking 2000 (MMCN'00), vol. 3969, pp 86-100, 2000.
- [2] Mohan R., Smith J.R., Li C.S., "Adapting Multimedia Internet Content for Universal Access," *IEEE Trans. on Multimedia*, vol. 1, no. 1, pp.104-114, 1999.
- [3] Han R., Bhagwat P. et al., "Dynamic Adaptation in an Image Transcoding Proxy for Mobile Web Access," *IEEE Personal Communications*, vol. 5, no. 6, pp.8-17, 1998.
- [4] Taylor C.N. & Dey S., "Adaptive image compression for wireless multimedia communication," *IEEE ICC 2001*, vol. 6, pp.1925-1929, 2001.
- [5] Desimone R. & Duncan J., "Neural mechanisms of selective visual attention," *Annual Review of Neuroscience*, vol. 18, pp.193-222, 1995.
- [6] Chen L.Q., Xie X. et al. "Image adaptation based on attention model for small form factor devices", to appear in 9th International Conference on Multi-Media Modeling, Jan. 2003.
- [7] Bruijn O. & Spence R., "Rapid Serial Visual Presentation: A space-time trade-off in information presentation," *Proceedings of Advanced Visual Interfaces (AVI'2000)*, pp.189-192, 2000.
- [8] Itti L., Koch C., "Computational modeling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194-203, Mar 2001.
- [9] Itti L., Koch C., Niebur E., "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on PAMI*, vol. 20, no. 11, pp.1254-1259, 1998.
- [10] Milanese R., Gil S., Pun T., "Attentive mechanisms for dynamic and static scene analysis," *Optical Engineering*, vol. 34 no. 8, pp.2428-2434, 1995.
- [11] Chernyak D.A., Stark L.W., "Top-Down Guided Eye Movement," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 31, pp. 514-522, Aug. 2001.
- [12] Mackworth N.H., Morandi A.J., "The gaze selects informative detail within pictures," *Perception and Psychophysics*, vol. 2, pp.547-552, 1967.
- [13] Ma Y.F., Lu L., Zhang H.J., Li M.J., "An Attention Model for Video Summarization", *ACM Multimedia 2002*, Dec. 2002.