Automatic Browsing of Large Pictures on Mobile Devices

Hao Liu[*1], Xing Xie[2], Wei-Ying Ma[2], Hong-Jiang Zhang[2]

Institute of Electronics, Chinese Academy of Science[1]
No. 17, Zhong Guan Cun Road, P. O. Box 2702, Beijing, 100080, P.R.China
hliu@mail.ie.ac.cn

Microsoft Research Asia[2]
5F, Sigma building, No. 49, Zhichun Road, Beijing, 100080, P.R.China
{xingx, wyma, hjzhang}@microsoft.com

ABSTRACT

Pictures have become increasingly common and popular in mobile communications. However, due to the limitation of mobile devices, there is a need to develop new technologies to facilitate the browsing of large pictures on the small screen. In this paper, we propose a novel approach which is able to automate the scrolling and navigation of a large picture with a minimal amount of user interaction on mobile devices. An image attention model is employed to illustrate the information structure within an image. An optimal image browsing path is then calculated based on the image attention model to simulate the human browsing behaviors. Experimental evaluations of the proposed mechanism indicate that our approach is an effective way for viewing large images on small displays.

Categories and Subject Descriptors

I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms

Algorithms, Human Factors

Keywords

Adaptive content delivery, image adaptation, attention model, form factor, browsing path, information foraging

1. INTRODUCTION

Recently, we have experienced an explosive growth of mobile multimedia applications. In these applications, images play a more and more important role in sharing, expressing and exchanging information in our daily lives. People prefer to share special moments with each other using visual contents such as images to stay connected when they are apart. Accompanying with this revolution, mobile handheld devices with diverse capabilities including embedded digital camera are also undergoing a considerable progress because of their portability and mobility. Now people can easily capture and share personal photos on these small-form-factor devices anywhere and anytime.

In order to make people really enjoy the ease of mobile communications, many hurdles still need to be crossed [15]. Among them, major crucial challenges include the limited accessing bandwidth and display sizes of mobile devices. Thanks to the galloping development of both hardware and software, the bandwidth condition is expected to be greatly improved. However, in the foreseeable future, the display, i.e. the form factor, will continue to be the major constraint on small mobile devices such as cell-phones and handheld PCs. A brief overview of the current display capabilities of various mobile devices have been given in [14]. In this paper, we will focus on facilitating image viewing on devices with limited display sizes.

Many efforts have been put on image adaptation and related fields from quite different aspects. For instance, the ROI coding scheme and Spatial/SNR scalability in JPEG 2000 [7] have provided a functionality of progressive encoding and display. It is useful for fast database access as well as for delivering different resolutions to terminals with different capabilities. Foveated coding [2] has been employed to reduce the transmission bandwidth of images by exploiting the space-variant degradation in the resolution of the human eye. Smith J. R. et al. [18][20] presented an image transcoding system based on the classification of image type and purpose. Most of these works only focused on compressing and caching contents in order to reduce the data transmission for fast delivery. Therefore, the results are often not consistent with human perception on small displays because of excessive resolution reduction or quality loss.

We have recently proposed an attention model based image adaptation approach in [3][8]. Instead of treating an image as a whole, we manipulate each region-of-interest in the image separately, which allows delivery of the most important region to the client when the screen size is small.

In [3], an extensible image attention model is introduced based on three attributes (region of interest, attention value, and minimal perceptible size) associated with each attention object and a set of automatic modeling methods are presented to support this approach. One of our modeling methods leverage the work on saliency map [11], which has been proposed as a computational model of focal visual attention. A branch and bound algorithm has been also developed to find the optimal solution efficiently. Though this approach achieved satisfactory results in our user

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

MM'03, November 2-8, 2003, Berkeley, California, USA.

study, much other information which a user cares may be lost due to the space limitation, especially for large images.

In [8], we proposed to employ a widely-used presentation technique, Rapid Serial Visual Presentation (RSVP), in which space is traded for time [1]. An image is decomposed into a set of spatial-temporal information elements which are displayed serially, each for a brief period of time, to aid users' browsing or searching through the whole image, which coincides with an important psychophysiological activity – visual attention shifting [11]. Visual attention can rapidly direct and shift the gaze towards interesting parts of the visual input. We depicted 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 in a way similar to RSVP. The acquirement of the fixated region depends on the algorithm in [3]. 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" [11], which demonstrates that current attention focus will be suppressed while selecting the next focus. We implemented it by removing the attention objects contained in the current display area and applying the same algorithm to the rest objects when selecting the next fixating area. The trace of the saccade is defined as the shortest path between centers of the two fixation areas.

Nevertheless, the method described in [8] is a heuristic approach to generate the RSVP process where the number of iterations, fixation duration, and saccade speed are all predefined. It does not tell us how to calculate an optimal browsing path to maximize the information throughput under limited space and time. In this paper, we address this problem by first extending our previous image attention model [3] to include a time constraint. We also mathematically formulate the optimal browsing path problem based on information foraging theory [6][19] which has been used to analyze the trade-offs in the value of information gained against the costs of performing activity in human-computer interaction tasks.

The novel contributions of this paper include:

- The automatic image browsing problem is formulated into an optimal browsing path selection problem based on the information foraging theory [19].
- The image attention model based on our previous work is extended to model users' attention which considers both spatial and temporal constraints.
- An efficient algorithm to generate the optimal browsing path based on extended image attention model is proposed.

The rest of this paper is organized as follows. Section 2 compares three different image adaptation methods and introduces the system framework of our approach. Section 3 discusses in detail the image attention model and our extensions. Based on this model, we formulate the optimal browsing path problem and present the corresponding algorithms in Section 4. In Section 5, we give the experimental results and demonstrate the usefulness of our proposed scheme. Finally, concluding remarks and discussions are provided in Section 6.

2. OUR SYSTEM FRAMEWORK

2.1 Information Fidelity

Visual attention can be seen as the ability of a portion of an image to attract the user's attention. Studies show that areas with higher visual attention tend to have more information. When it comes to the problem of image adaptation or browsing based on attention model, it is helpful to introduce a concept of *Information fidelity* as the perceptual 'look and feel' of a modified version of content object, a subjective comparison with the original version. Information fidelity can be calculated as the sum of attention values during the adaptation or browsing process. Its value is confined between 0 (lowest, all information lost) and 1 (highest, all information kept).

In the following, we use this concept to compare the existing approaches with our new approach for image adaptation and browsing.

Let us consider an image I as a set of M x N evenly distributed information blocks I_{ij}:

$$I = \{I_{ij}\} = \{(AV_{ij}, r_{ij})\}, \quad 1 \le i \le M, 1 \le j \le N, r_{ij} \in (0,1) \quad (1)$$

where (i, j) corresponds to the location at which the information block I_{ij} is sampled; AV_{ij} is the visual attention value of I_{ij}; r_{ij} is the spatial scale of I_{ij}, representing the minimal spatial resolution to keep I_{ij} perceptible. For example, considering a typical outdoor image, the house is around ten meter range, and the faces are likely in less than three decimeter scale range. Therefore, the house can be scaled down more aggressively than the faces.

In our work, the attention value of each information block in an image is normalized so that their sum is 1.

2.2 Direct Down-sampling

For direct image down-sampling as shown in Figure 1(a), the information fidelity on a specific display size can be described as a function f_D:

$$f_D(I) = \sum_{I_{ij} \in I} AV_{ij} u(r_D - r_{ij}) \quad (2)$$

where $u(x)$ is a step function defined as

$$\mathrm{u}(x) = \begin{cases} 1, & \text{if } x \ge 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

and r_D is the down-sampling scale defined as

$$r_D = \min\left(\frac{Width_{Screen}}{Width_{Image}}, \frac{Height_{Screen}}{Height_{Image}} \right) \quad (4)$$

The bigger the image, the smaller the r_D. When r_D is smaller than most of the spatial scales, the information will be lost substantially.

2.3 Cropping-based Approach

A cropping based image adaptation approach was described in [3]. Its philosophy is to choose the most important region in the original image to accommodate as much information as possible, as shown in Figure 1(b). The information fidelity of this image adaptation strategy can be defined as a function f_C

$$f_C(I) = \sum_{I_{ij} \in I_C} AV_{ij} u(r_C - r_{ij}), I_C \subseteq I \quad (5)$$

where I_C is a subset of the entire image blocks and

$$r_C = \max_{I_{ij} \in I_C} r_{ij} \leq \min\left(\frac{Width_{Screen}}{Width_{I_C}}, \frac{Height_{Screen}}{Height_{I_C}}\right) \quad (6)$$

This strategy assumes that most information is usually confined to a small number of image blocks, which however, is not always true for images of natural scenes.

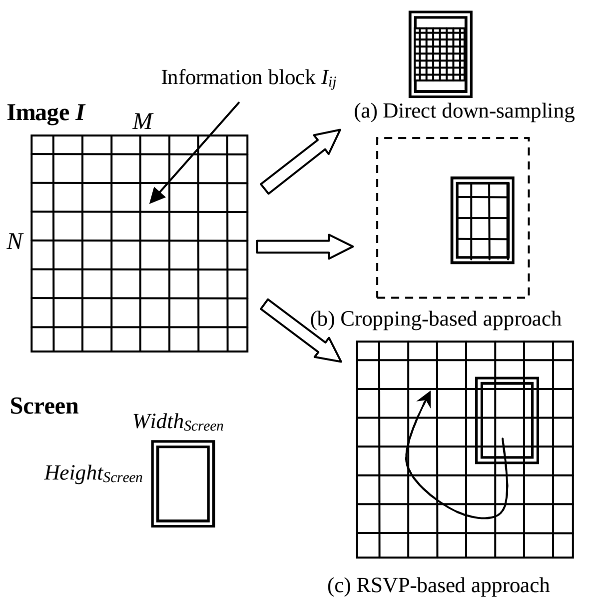


Figure 1. A comparison of three image adaptation strategies.

2.4 Our RSVP-based Approach

Rapid Serial Visual Presentation (RSVP), a representation technique used to support electronic information browsing, was used to improve the image browsing on small devices in [8]. As we discussed in the introduction, the previous method is a heuristic approach and does not tell us how to calculate an optimal browsing path under limited space and time.

The information fidelity of this RSVP-based approach can be described as a function of both space and time:

$$f_{RSVP}(I,T) = \int_0^T \sum_{I_{ij} \in I_{RSVP}(t)} AV_{ij} u(r(t) - r_{ij}) dt, I_{RSVP}(t) \subseteq I \quad (7)$$

where $I_{RSVP}(t)$ is a subset of the information blocks and varies with time and

$$r(t) = \max_{I_{ij} \in I_{RSVP}(t)} r_{ij} \leq \min\left(\frac{Width_{Screen}}{Width_{I_{RSVP}(t)}}, \frac{Height_{Screen}}{Height_{I_{RSVP}(t)}}\right) \quad (8)$$

Figure 1(c) shows an example. In fact, the RSVP-based approach is just a simulation of our real-life experience when browsing large images on a small screen. We manually scroll the window horizontally and vertically to view the different parts of an image, which is equivalent to change $I_{RSVP}(t)$. Viewing the detail of an interesting region by zooming is then equivalent to adjust $r(t)$.

As it is an inconvenient and very time-consuming process to zoom and pan a large image fully manually on a small device, our objective is to detect the optimal path to automatically pan/zoom the window to browse through the different parts of an image.

A complete framework of our approach is shown in Figure 2, which includes different stages from image modeling, pre-processing to the optimal browsing path generation. First of all, the input image is analyzed to extract both top-down and bottom-up features, e.g., human faces, texts or saliency objects. Then different modeling methods are adopted to generate a visual attention model for the image. Afterwards, some heuristic rules are applied to the pre-processing phase in order to generate a set of attention groups which can fit into the display size. Finally, an optimal browsing path is generated according to the user preference. An interactive browsing scheme can be also employed in our approach, in which users can interrupt the automatic process to adjust the browsing path. The details of each step in this framework will be discussed further in following sections.

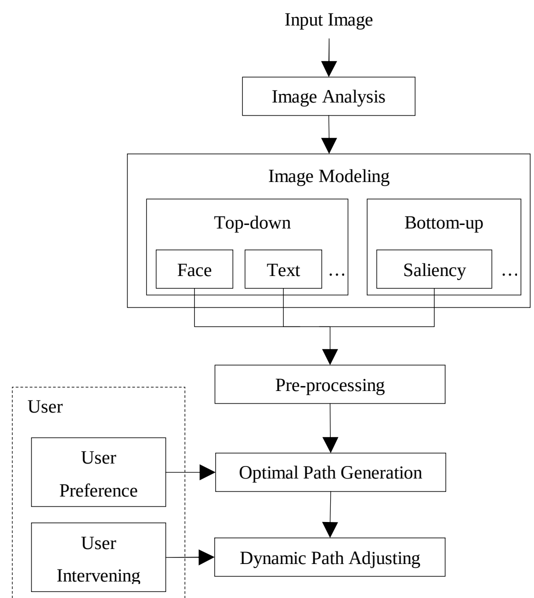


Figure 2. The framework of attention based image browsing.

3. EXTENDING IMAGE ATTENTION MODEL

Computational attention allows us to break down the problem of understanding a content object into a series of computationally less demanding and localized analytical problems. Although there are still debates on the mechanics of visual attention and many models have been proposed to depict the visual attention activities [5][10][17], it is gradually deemed that a complete computational model would be the integration of bottom-up (e.g. saliency map) and top-down (e.g. face and text) cues.

With the new image browsing mechanism based on RSVP, where space can be traded for time, the different parts of an image can be displayed serially, each for a brief period of time to the user. Therefore, we need to extend the existing image attention model to include a time factor. That is, the attention value of an information block in an image should not only be affected by its area of presentation, but also the duration of presentation on the screen.

Definition 1: The visual attention model for an image is defined as a set of attention objects:

$$\{AO_i\} = \{(ROI_i, AV_i, MPS_i, MPT_i)\}, \quad 1 \le i \le N \quad (9)$$

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
MPT_i,	minimal perceptible time of AO_i
N,	total number of attention objects in the image

We assign four attributes to each attention object, which are *Region-Of-Interest (ROI)*, *attention value (AV)*, *minimal perceptible size (MPS)*, and *minimal perceptible time (MPT)*. The notion of '*Region-Of-Interest (ROI)*' is borrowed from JPEG 2000 [7], which is referred as a spatial region within an image that corresponds to an attention object. *Attention value (AV)* is a quantified value indicates the weight of each attention object in contribution to the information contained in the original image. *Minimal perceptible size (MPS)* represents the minimal allowable spatial area of an attention object. It is introduced as a threshold to avoid excessively sub-sampling during the reduction of display size. The values of *AV* and *MPS* can be computed from the type, position and size of each object [3].

We introduce *minimal perceptible time (MPT)* in this paper as a threshold for the fixation duration when browsing an attention object. If an attention object does not stay on the screen longer than *MPT*, it may not be perceptible enough to let users catch the information. For instance, considering a human face, its *MPT* can be defined to be 200ms which is the shortest possible time to show the face without severely degrading its perceptibility.

The image attention model can be manually pre-assigned by authors or publishers, which however could be labor intensive. A more plausible approach is to detect each attention object automatically and build up the entire image attention model. A set of algorithms to generate the attention model has been discussed in [3], and in this paper we will focus on the calculation of *MPT* attribute.

The Computation of *MPT*

In our current implementation, we consider three types of image features, saliency [12][16], face [13] and text [4], to detect the attention objects.

Fixation durations are variable, typically ranging from 100ms to 300ms. In our experiments, we predefine the *MPT* of a saliency region proportional to its attention value. It can be adjusted according to the user preference and the display context such as whether the screen is reflective or backlit. For example, the *MPT* of a saliency region in perusing mode (more information preferred) will be larger than that in skimming mode (less time preferred).

The appearance of dominant faces in images will certainly attract viewers' attention. By employing the face detection algorithm in [13], we obtain the face information including the number of faces, the pose, region, and position of each face. In our implementation, the *MPT* of a face region is defined as 200ms.

Similar to human faces, text regions also attract viewers' attention in many situations. The *MPT* of a text region can be defined as a function of several parameters such as the number of words, which can be estimated by a text detection module. We define an average *MPT* for each word according to the experimental results in [21][22]. In our system, we employ an $MPT_{average}$ of 250ms for a word and the *MPT* of a text region is defined as:

$$MPT_{text} = N_{text} \times MPT_{average} \quad (10)$$

where N_{text} denotes the number of detected words. Since the fixation duration for a text region usually changes with users' education background and age, MPT_{text} can be also adapted to users' preference.

4. AUTOMATIC BROWSING OF IMAGES

According to our observations, when browsing large images on small devices, people often devote considerable efforts in scrolling and zooming to view the content. Our automatic RSVP-based image browsing method would provide users most of the image content in reasonable time, that is, making an effective use of both time and space.

4.1 The Definition of Image Browsing Path

The human browsing behavior can be approximately modeled by two mutually exclusive statuses: the fixation status (e.g., exploiting an interesting region) and the shifting status (e.g., scrolling to the next region). The fixation status corresponds to the static viewing of an attention objects, and the shifting status can be simulated by traveling between different attention objects. The shifting path is the shortest path between centers of the two fixation areas (i.e. attention objects).

Definition 2: The image browsing path P is defined as a collection of successive path segments:

$$P = \{P_i\} = \{(SP_i, EP_i, SR_i, ER_i, T_i)\}, \quad 1 \le i \le N \quad (11)$$

where

P_i,	the i^{th} path segment
SP_i,	starting point of P_i
EP_i,	ending point of P_i
SR_i,	starting resolution of P_i
ER_i,	ending resolution of P_i
T_i,	time cost for scrolling from SP_i to EP_i

Since the path segments are successive, we should have:

$$SP_i = EP_{i-1} \text{ and } SR_i = ER_{i-1}, \text{ for every } 1<i\le N \quad (12)$$

Generally speaking, three types of browsing patterns exist, that is, panning, zooming, and panning with zooming. Suppose there is a virtual intelligent camera which can pan and zoom in the original image, it will be automatically steered to deliver those important regions in an efficient way. A maximal panning velocity *MPV* and a maximal zooming rate *MZR* are defined to avoid hypermetric motion. We also assume that the velocity is uniform while moving. Therefore, the time cost T_i for scrolling from starting point SP_i to ending point EP_i can be calculated as:

$$T_i = \max\left\{\frac{dist(SP_i, EP_i)}{MPV}, \frac{|SR_i - ER_i|}{MZR}\right\} \quad (13)$$

where $dist(SP_i, EP_i)$ is defined as the Euclidean distance between SP_i and EP_i. The values of *MPV* and *MZR* are currently defined by experience.

4.2 Finding the Optimal Path

According to the information foraging theory [19], people will modify their strategies or the structure of the environment to maximize their rate of gaining valuable information. Let *R* denote the rate of gain of valuable information per unit cost, which can be computed as follows.

$$R = \frac{G}{T_B + T_w} \quad (14)$$

where *G* is the total net amount of valuable information gained, which can be seen as the information fidelity defined in Section 2. T_B is the total amount of time spent on shifting between subsequent fixation areas (attention objects). T_W, the exploiting cost, is the total amount of *MPT*s spent on the fixation areas.

The problem of identifying the optimal image browsing path *P* is to maximize *R*, which is equivalent to the following equation:

$$\underset{P}{Max}\left\{\frac{f_P(I, T_P)}{T_P}\right\} \quad (15)$$

where $f_P(I,T_P) = G$ and $T_P = T_B + T_W$. $f_P(I,T_P)$ denotes the information fidelity provided in the image browsing path *P*, and T_P stands for the total amount of time spent for fixation and shifting in *P*. Note that $f_P(I,T_P)$ is a special case of $f_{RSVP}(I,T)$ in Equation (7) because of the additional conditions imposed on the image browsing path (Section 4.1).

From Definition 2, we know that

$$T_P = \sum_{1 \le i \le N} T_i + \sum_{AO_j \in A(EP_N, ER_N)} MPT_j + \sum_{1 \le i \le N} \sum_{AO_j \in A(SP_i, SR_i)} MPT_j \quad (16)$$

where *A(loc, res)* stands for the set of attention objects which are perceptible when the focus location is *loc* and the display resolution is *res*. It can be easily calculated based on the image attention model and the screen size. We refer to [3] for details.

Similarly, we have

$$f_P(I, T_P) = \sum_{AO_j \in A(EP_N, ER_N)} AV_j + \sum_{1 \le i \le N} \sum_{AO_j \in A(SP_i, SR_i)} AV_j \quad (17)$$

The maximization of *R* can be achieved by either maximizing information fidelity or minimizing time cost. Therefore, we propose two image browsing modes here:

- *Perusing mode*: users prefer to spend as little time as possible as long as a certain percentage of the information is presented to him.
- *Skimming mode*: users prefer to read as much information as possible within a limited period of time.

For the skimming mode, the optimization problem becomes:

$$Given \quad T_P \le \lambda_T, \quad \underset{P}{Max}\{f_p(I, T_P)\} \quad (18)$$

where λ_T is the threshold of maximal time cost that the user can afford, which is usually less than the minimal time cost to browsing all the attention objects.

For the perusing mode, the problem becomes:

$$Given \quad f_P(I, T_P) \ge \lambda_{AV}, \quad \underset{P}{Min}\{T_P\} \quad (19)$$

where λ_{AV} stands for the minimal attention value or information percentage that user prefer to obtain, which is usually less than the total attention value in the image.

In order to solve the optimization problem efficiently, we employ a two-step approach here. First we need to do some preprocessing to split large attention objects and group nearby objects to form a set of *attention groups*. Then the optimal browsing path is generated to connect the attention groups.

4.2.1 Pre-processing

The image attention model itself does not ensure that the minimal display area for each attention object will be smaller than all kinds of target screen sizes. A large attention object will not be able to be viewed properly if the screen size is smaller than its *MPS*. This causes problems to the optimal browsing path generation. Therefore, there is a need to split large attention objects before grouping. A big saliency or text attention object will be split evenly according to the screen size and its *AV*, *MPS* and *MPT* values will be evenly distributed to the newly created objects proportional to their area. For a face attention object, since its *MPS* will be usually smaller than the screen size, we will not split it.

After the splitting stage, we will combine nearby attention objects to form a set of attention groups. This helps to reduce the computational complexity of browsing path generation algorithms. The combination is based on the branch and bound algorithm in [3]. After obtaining the first attention group, we remove the objects in this group and apply the same algorithm to the rest attention objects iteratively. We define the *AV* of an attention group as the sum of *AV*s of perceptible attention objects in the group and the *MPT* as the sum of corresponding *MPT*s.

4.2.2 Path Generation under Skimming Mode

In this browsing mode, the objective of optimization is to find a sequence of P_i to maximize the information fidelity within a limited period of time. The general problem is *NP*-hard and no efficient solution exists in terms of computational time or storage space. Here we use a backtracking algorithm to enumerate all the possible paths and then find the best one among them:

1) Arrange the attention groups according to their attention values in a decreasing order. For each group, select it as the starting point and do the step 2 and 3.
2) Use the backtracking algorithm to search among all possible paths from the starting point and calculate the total browsing time and the information fidelity for each path.
3) For each node in the backtracking tree, check the time cost to ensure that it is smaller than the predefined threshold.
4) Finally, select the browsing path with the largest information fidelity as the resulting path.

4.2.3 Path Generation under Perusing Mode

In the perusing mode, the objective of optimization is to find a sequence of P_i to minimize the time cost as long as a certain percentage of the information is presented to him. To solve this problem, we just slightly adjust the previous backtracking algorithm as following:

1) Arrange the attention groups according to their *MPT*s in an increasing order. For each group, select it as the starting point and do the step 2-4.
2) Use the backtracking algorithm to search among all possible paths from the starting point and calculate the total browsing time and the information fidelity for each path.
3) For each node in the backtracking tree, there is a bound on the possible time cost it would spend among all of its sub-trees. The lower bound is just the time that has been spent and the upper bound is the addition of all possible time costs for those unchecked attention groups after current level in the backtracking tree.
4) Whenever the lower bound of a node is larger than the best time cost currently achieved, the whole sub-tree of that node will be truncated. If the information fidelity currently achieved is already greater than the predefined threshold, we will also not go any deeper in the sub-tree.
5) Finally, we select the browsing path with the least time cost as the resulting path.

The computational complexity of this algorithm is exponential. However, by checking both the bound on possible information fidelity value and the time bound of each path, the computation cost is greatly reduced.

If λ_{AV}=1, which may be the most common case indicating that the user want to view all the information, we can transform the problem to a *Traveling Salesman Problem*. Therefore, some approximation algorithms can be applied to get a fast but sub-optimal solution when the number of attention groups is large.

4.3 Interactive Image Browsing

Though the browsing path can be generated automatically for a given image, it is reasonable to allow the user to stop the automated process at any time, choose where to look at interactively, and resume the automatic browsing process afterwards.

The optimal browsing path after user's intervening can be generated based on the following algorithm:

1) When the browsing process is paused, record the remaining set of attention objects S_r.
2) During interaction, record the set of attention objects S_m which has been viewed by the user.
3) Re-generate the optimal path $\{P_i\}$ based on S_r - S_m.
4) Move smoothly from current location to the starting point of P_1, and complete the rest of path.

5. EVALUATION RESULTS

In order to validate the performance of our proposed scheme, we have implemented a prototype image browser on a Compaq iPaq 3670 with 64M memory, 320x240 display and PocketPC 2002 as its operating system.

Figure 3 gives an example of the optimal browsing path generation. In the image, eleven attention objects are detected and marked with green rectangles, including two text objects, seven face objects and two saliency objects. After the pre-processing step, we get four attention groups marked with dashed rectangles. We label these four attention groups from the left most one clockwise as *AB1*, *AB2*, *AB3* and *AB4*. The optimal browsing path is represented by a red line in this example, which is generated by setting $\lambda_{AV}=1$ under perusing mode. During the automatic browsing process, the window will move from *AB1* to *AB2*, then *AB3* and *AB4*. The fixation points are marked by a red circle whose diameter represents the fixation duration. The corresponding attention value accumulation curve is presented at the bottom of Figure 3. In our experiments, the attention value of each object in an image is normalized so that their sum is 1. Figure 4 shows the prototype with a real browsing example.

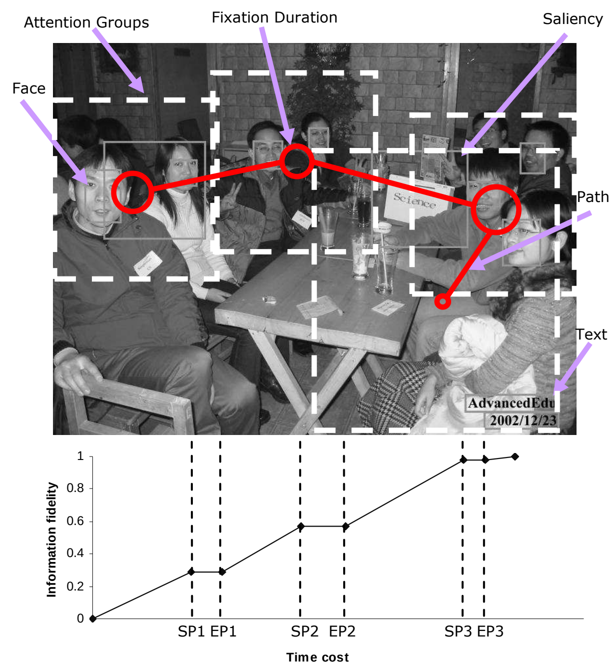
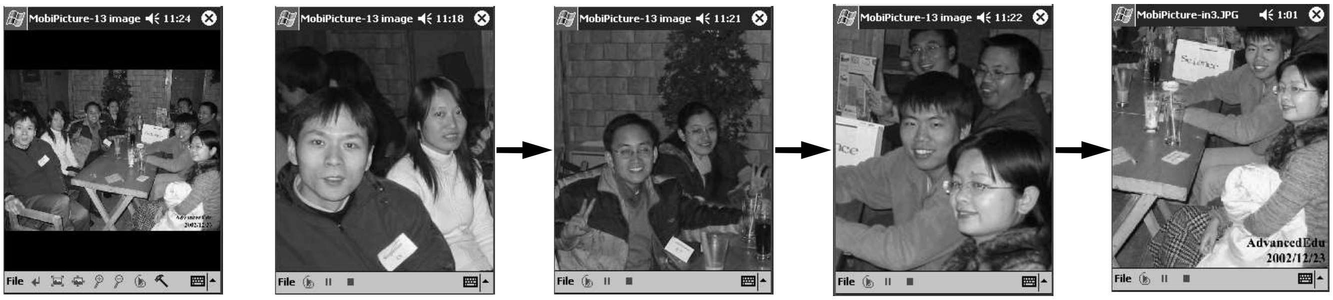


Figure 3. An example of optimal browsing path generation.

During a pilot study of the browsing path generation algorithm, we found that ensuring "locality" is a very important factor of its performance. There are two types of locality: zooming locality and scrolling locality. People usually do not prefer zooming back and forth within a small area in the image, or scrolling left and right for a long distance. For example, in Figure 3, AB1-AB3-AB2-AB4 is not a good browsing path since it contains back and forth scrolling which is quite annoying. According to equation (13), both of the cases will result in a large T_i. Therefore, the segment corresponding to these two cases will very likely be excluded in the optimal browsing path.

Although many researchers have addressed the issue of image adaptation, still there is no objective measure to evaluate the performance of the algorithms. In this paper, we carried out a user study to evaluate the performance of our algorithm.



Fit-to-window Automatic browsing

Figure 4. Automatic browsing of the example image.

We chose 30 test images from various sources. These images vary from 326x450 to 1600x1200 in size. Many of them are obtained from personal albums and popular Web sites. They are divided into five classes: group photos, news pictures, personal indoor or outdoor photos and scenery images, each class with 6 images.

Sixteen volunteers, including eleven males and five females, were invited to give their subjective scores on the browsing performance by our approach while comparing with results from traditional image browsing method. They are familiar with computers and are without any knowledge of the work in this paper. All the subjects were asked to give their judgments at the following three questions:

1. *Is our approach better than traditional image viewer?*
2. *Do the fixated regions really represent interesting areas?*
3. *Is the generated image browsing path reasonable?*

The evaluation results are listed in Table 1, Table 2 and Table 3, respectively.

Table 1. Evaluation results for the question #1.

Image class	Good	No difference	Bad
Group	78.20%	15.38%	6.41%
Indoor	67.78%	22.22%	10.00%
Outdoor	70.00%	23.33%	6.67%
Scenery	64.44%	27.78%	7.78%
News	62.22%	27.78%	10.00%
Average	**68.53%**	**23.30%**	**8.17%**

Table 2. Evaluation results for the question #2.

Image class	Interesting	Uninteresting
Group	76.92%	23.08%
Indoor	67.78%	32.22%
Outdoor	74.44%	25.56%
Scenery	57.78%	42.22%
News	73.33%	26.67%
Average	**70.05%**	**29.95%**

Table 3. Evaluations results for the question #3.

Image class	Reasonable	Unreasonable
Group	82.05%	17.95%
Indoor	68.75%	31.25%
Outdoor	68.89%	31.11%
Scenery	72.22%	27.78%
News	62.22%	37.78%
Average	**71.49%**	**28.51%**

As can be seen, for the first question, more than 68% of the subjects consider our solution better than the conventional method and only 8% of them consider worse. This means that most users prefer the new functions based on our technique. According to the results, our browsing technique is especially suitable for group photos, in which face objects, very important semantic factors guiding the visual attention, are distributed widely on the entire image. The viewer can take advantage of the automatic browsing path to quickly and conveniently catch the content of the image.

As to the second question, the experimental results show that our extended attention model is effective, especially for the group photos, for locating the attention-getting regions in the image. The average result of the third question on browsing path quality is also satisfactory. More than 71% of the subjects think that the optimal browsing path we generated is reasonable and helpful for browsing large images on the small display. For this question, we use the perusing mode and set $\lambda_{AV}=1$. The skimming mode has not been tested since its performance heavily depends on the setting of time bound. Different users usually choose different time bounds when browsing the images.

We also conducted an experiment on the efficiency of our algorithm. The time cost includes the pre-processing and optimal path generation procedure. We got an average time cost at 230 microseconds per image, with variation from 150 to 350 microseconds. Without code optimization, our technique is already fast enough to be employed on mobile devices for real-time image browsing applications.

6. CONCLUDING REMARKS

Currently, the predominant methods for accessing large images on small devices are down-sampling or manual browsing by zooming and scrolling. Image down-sampling or thumbnail view results in significant information loss due to the excessive resolution reduction. Manual browsing can avoid information loss but it is often time-consuming for users to catch the most crucial information of an image.

In this paper, we propose a novel image browsing strategy based on image attention model to facilitate scrolling and navigation of large pictures on devices with small displays. Experimental evaluations indicated that our approach has significantly improved the user's browsing experiences on a variety of images.

We are currently considering improving those adaptation parameters in our approach by learning user feedbacks. With the satisfactory results from our experiments, we plan to extend our work to other media types, such as videos [9] and Web pages. We will continue to investigate these directions in our future work.

7. ACKNOWLEDGEMENTS

We would like to express our special appreciation to Liqun Chen, Xin Fan and Yusuo Hu for their insightful suggestions and the Media Computing Group of Microsoft Research Asia for their generous help in building some of the image analysis modules. We also thank all the voluntary participants in our user study experiments.

8. REFERENCES

[1] O. Bruijn and R. Spence, Rapid serial visual presentation: a space-time trade-off in information presentation, Proc. of Advanced Visual Interfaces, pp189-192, 2000.

[2] E.C. Chang, S. Mallat, and C. Yap, Wavelet foveation, Journal of Applied and Computational Harmonic Analysis, Vol. 9, No. 3, pp312-335, Oct. 2000.

[3] L.Q. Chen, X. Xie, X. Fan, W.Y. Ma, H.J. Zhang, and H.Q. Zhou, A visual attention model for adapting images on small displays, ACM Multimedia Systems Journal, to appear.

[4] X.R. Chen and H.J. Zhang, Text area detection from video frames, Proc. of 2nd IEEE Pacific-Rim Conf. on Multimedia, pp222-228, Beijing, China, Oct. 2001.

[5] D.A. Chernyak and L.W. Stark, Top-down guided eye movement, IEEE Trans. on Systems, Man and Cybernetics, Vol. 31, pp514-522, Aug. 2001.

[6] E.H. Chi, P. Pirolli, K. Chen, and J. Pitkow, Using information scent to model user information needs and actions on the Web, ACM CHI 2001, Seattle, Washington, Mar. 2001.

[7] C. Christopoulos, A. Skodras, and T. Ebrahimi, The JPEG2000 still image coding system: an overview, IEEE Trans. on Consumer Electronics, Vol. 46, No. 4, pp1103-1127, 2000.

[8] X. Fan, X. Xie, W.Y. Ma, H.J. Zhang, and H.Q. Zhou, Visual attention based image browsing on mobile devices, Proc. of ICME 2003, Vol. I, pp53-56, Baltimore, USA, Jul. 2003.

[9] X. Fan, X. Xie, H.Q. Zhou, and W.Y. Ma, Looking into video frames on small displays, ACM Multimedia 2003, Berkeley, CA, USA, to appear.

[10] L. Itti and C. Koch, A comparison of feature combination strategies for saliency-based visual attention system, Proc. of SPIE: Human Vision and Electronic Imaging IV, Vol. 3644, pp473-482, 1999.

[11] L. Itti and C. Koch, Computational modeling of visual attention, Nature Reviews Neuroscience, Vol. 2, No. 3, pp194-203, Mar. 2001.

[12] L. Itti, C. Koch, and E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 20, No. 11, pp1254-1259, 1998.

[13] S.Z. Li, L. Zhu, Z.Q. Zhang, A. Blake, H.J. Zhang, and H. Shum, Statistical learning of multi-view face detection, Proc. of 7th European Conference on Computer Vision, Vol. 4, pp67-81, Copenhagen, Denmark, May 2002.

[14] J. Luo, A. Singhal, G. Braun, R.T.Gray, O. Seignol, and N. Touchard, Displaying images on mobile devices: capabilities, issues, and solutions, Proc. of International Conference on Image Processing 2002, Vol. 1, pp13 -16, Rochester, New York, 2002.

[15] W.Y. Ma, I. Bedner, G. Chang, A. Kuchinsky, and H.J. Zhang, A framework for adaptive content delivery in heterogeneous network environments, Proc. of Multimedia Computing and Networking 2000, SPIE Vol. 3969, pp86-100, San Jose, USA, 2000.

[16] Y.F. Ma and H.J. Zhang, Contrast-based image attention analysis by using fuzzy growing, ACM Multimedia 2003, Berkeley, CA, USA, to appear.

[17] R. Milanese, S. Gil, and T. Pun, Attentive mechanisms for dynamic and static scene analysis, Optical Engineering, Vol. 34, No. 8, pp2428-2434, 1995.

[18] R. Mohan, J.R. Smith, and C.S. Li, Adapting multimedia Internet content for Universal Access, IEEE Trans. on Multimedia, Vol. 1, No. 1, pp.104-114, 1999.

[19] P. Pirolli and S.K. Card, Information foraging, Psychological Review, Vol. 106, No. 4, pp643-675, 1999.

[20] J.R. Smith, R. Mohan, and C.S. Li, Content-based transcoding of images in the Internet, Proc. of Int. Conf. on Image Processing 1998, Vol. 3, pp7-11, Chicago, USA, Oct. 1998.

[21] F.C. Sun and L.Y. Chen, Identification of fixations in reading eye movements by a multi-layer neural network**,** *Proc. of* IEEE International Conference on Neural Networks 1995, Vol. 5, pp2309-2313, Perth, Australia, Nov. 1995.

[22] F. Vitu, G.W. McConkie , P. Kerr, and J. K. O'Regan, Fixation location effects on fixation durations during reading: an inverted optimal viewing position effect, Vision Research, Vol. 41, No. 25-26, pp3513–3533, 2001.