

**A visual attention model for adapting
images on small displays**

Liqun Chen
Xing Xie
Xin Fan
Wei-Ying Ma
Hong-Jiang Zhang
Heqin Zhou

2002-11-4

Technical Report
MSR-TR-2002-125

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

A visual attention model for adapting images on small displays

LI-QUN CHEN^{1*}, XING XIE²(✉), XIN FAN^{1*}, WEI-YING MA², HONG-JIANG ZHANG², HE-QIN ZHOU¹

¹*Dept. of Automation, University of Science and Technology of China, Hefei, 230027, P.R.China*

E-mail: {lqchen80, van}@mail.ustc.edu.cn, hqzhou@ustc.edu.cn

²*Microsoft Research Asia, 3/F Sigma Center, No. 49 Zhichun Road, Beijing, 100080, P.R.China*

E-mail: {i-xingx, wyma, hjzhang}@microsoft.com

Tel: +86-10-62617711 Ext. 3211

Fax: +86-10-88097306

Part of this work was originally presented in the 9th International Conference on Multi-Media Modeling (MMM'03).

*This work was conducted while the first and third authors were visiting students at Microsoft Research Asia.

Abstract. Image adaptation, one of the essential problems in adaptive content delivery for universal access, has been actively explored for some time. Most existing approaches have focused on generic adaptation towards saving file size under constraints in client environment and hardly paid attention to user's perception on the adapted result. Meanwhile, the major limitation on the user's delivery context is moving from data volume (or time-to-wait) to screen size because of the galloping development of hardware technologies. In this paper, we propose a novel method for adapting images based on user attention. A generic and 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. A branch and bound algorithm is also developed to find the optimal adaptation efficiently. Experimental results demonstrate the usefulness of the proposed scheme and its potential application in the future.

Key words: *Image adaptation -- Attention model -- Region-Of-Interest -- Attention value -- Minimal perceptible size -- Information fidelity*

Abbreviations: AO, attention object; ROI, Region-Of-Interest; AV, attention value; MPS, minimal perceptible size; IF, information fidelity

1. Introduction

As the Internet contents, client devices and user preferences continue to diversify, it is widely acknowledged that adaptive and customized information services are critical for Internet content providers to improve their quality of services by accommodating the large increasing variety of clients so as to attract customers. On the other hand, more and more client users in a heterogeneous environment wish all the information contents to be suitable for universal access [15, 20], i.e. one can access any information over any network from anywhere through any type of client device.

Since a great deal of information on the Internet today is presented by visual contents, it is essential to make image adaptive to the various contexts of clients. At the same time, thanks to the galloping development of information technologies in both hardware and software, more and more new small devices with diverse capabilities, such as Handheld PC, Pocket PC, and Smartphone, are making a population boom on the Internet mobile clients other than the original Desktop PCs because of their portability and mobility. Although these client devices are becoming more and more powerful in both numerical computing and data caching, nevertheless, low bandwidth connections and small displays, the two crucial limitations on accessing the current Internet, are still a great obstruction to their extensive prevalence. The bandwidth condition is expected to be greatly improved with the development of 2.5G and 3G wireless networks, while the display size is more likely to remain unchanged due to the mobility requirement of these devices. In this paper, we'd like to focus on the latter, that is, adapting images for devices with limited screen size.

Many efforts have been put on visual content adaptation and related fields from quite different aspects including the JPEG and MPEG standards. The ROI coding scheme and Spatial/SNR scalability in JPEG 2000 [4] has 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 in terms of display and bandwidth. In the MPEG-7 Multimedia Description Schemes [7, 8], Media Profile Description was proposed to refer to the different variations that can be produced from an original or master media depending on the values chosen for the coding, storage format, etc. Two components, MediaTranscodingHints D and MediaQuality D, of the Media Profile D are designed to provide information for content adaptation and reduce its computational complexity by specifying transcoding hints of the media being described and representing both subjective quality ratings and objective quality ratings, respectively. Currently, MPEG-21 starts to work on defining an adaptation framework named Digital Item Adaptation [9] for multimedia content including images.

The image adaptation problem has also been studied by researchers for some time. A proxy-based architecture to perform on-demand datatype-specific content adaptation was proposed in [5]. In particular, the authors have made image distillation (i.e. compression) to illustrate that adaptation is beneficial in saving data transmission time. They classified three areas of client variation: network, hardware and software; they also gave three sets of corresponding image distillation functions: file size reduction, color reduction, and format conversion. Smith J. R. et al. [17, 19] present an image transcoding system based on the content classification of image type and

image purpose. They first classify the images into image type and image purpose classes by analyzing the image characteristics, the related text and Web document context, and then based upon the analysis results, the transcoding system chooses among the transcoding functions that modify the images along the dimensions of spatial size, fidelity, and color, and that substitute the images with text. The authors of [6] proposed a framework for determining when/whether/how to transcode images in a HTTP proxy while focusing their research on saving response time by JPEG/GIF compression, which is determined by bandwidth, file size and transcoding delay. They discussed their practical policies in transcoding based on experience and simple rules. An analysis of the nature of typical Internet images and their transcoding characteristics was presented in [1] while focusing on file size savings. This work provided useful information to developers of a transcoding proxy server to choose the appropriate transcoding techniques when performing image adaptation. Recently, a new approach of ROI-based adaptation [13] has been investigated. Instead of treating an image as a whole, they manipulate each region-of-interest in the image separately, which allows delivery of the important region to the client when the screen size is small. This is the only work we have seen that has taken user perception into consideration.

Although there have been so many approaches for adapting images, most of them only focused on compressing and caching contents in the Internet in order to reduce the data transmission and speed-up delivery. Hence, the results are often not consistent with human perception because of excessive resolution reduction or quality compression. Furthermore, it is worth to point out that the algorithms involving large number of adaptation rules or over-intensive computation makes them impracticable in the on-the-fly systems of adaptive content delivery.

Aiming at solving the limitation in current algorithms while avoiding semantic analysis, in this paper, we present a generic image adaptation framework based on viewer's attention model. Attention is a neurobiological conception. It means the concentration of mental powers upon an object; a close or careful observing or listening, which is the ability or power to concentrate mentally. 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. Thus, it is powerful to the content analysis in adaptive content delivery by providing the exact information to facilitate the decision making of content adaptation.

The computational attention methodologies have been studied by some researchers. The authors of [12] reviewed recent works on computational models of focal visual attention, and presented a bottom-up, saliency- or image-based visual attention system. By combining multiple image features into a single topographical saliency map, the attended locations are detected in the order of decreasing saliency by a dynamical neural network [10, 11]. A selective attention-based method for visual pattern recognition was presented in [18], together with promising results when applying this method in handwritten digit recognition and face recognition. Recently, Ma Y. F. et al. [16] presented a generic video attention model by integrating a set of attention models in video and applied such modeling in video summarization.

In this paper, we propose a novel solution to generic image adaptation that dynamically modifies the image contents to optimally match the various screen sizes of client devices based on modeling of viewer's attention. The main contributions of our work are twofold: a new scheme to

model user attention in viewing images, and an efficient algorithm to apply such modeling in image adaptation and browsing. From the promising experimental results we obtained, we demonstrate the feasibility and efficiency of our approach.

The rest of this paper is organized as follows: Section 2 introduces the framework of image attention model. In Section 3, several methods for automatic modeling user attention on visual features are presented. Section 4 describes in detail a new approach of image adaptation based on the attention model and the performance of adaptation is evaluated in a user study experiment with the results reported in Section 5. Finally, Section 6 gives the concluding remarks and discussions on future work.

2. Attention model for image

In this section, we give the definition of our visual attention model for image. This is the basis of our image adaptation algorithm.

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

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

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

An *attention object (AO)* is an information carrier that delivers the author’s intention and catches part of the user’s attention as a whole. An attention object often represents a semantic object, such as a human face, a flower, a mobile car, a text sentence, etc. Generally, most perceptible information of an image can be located inside a handful of attention objects and at the same time these attention objects catch the most attentions of a user. Therefore, the image adaptation problem can be treated as manipulating attention objects to provide as much information as possible under resource constraints. We assign three attributes to each attention object, which are *Region-Of-Interest (ROI)*, *attention value (AV)*, and *minimal perceptible size (MPS)*. Each of them is introduced in detail in the following.

2.1. Region-Of-Interest

We borrow the notion of ‘*Region-Of-Interest (ROI)*’ from JPEG 2000 [4], which is referred in our model as a spatial region or segment within an image that corresponds to an attention object. As shown in Figure 1, *ROIs* can be in arbitrary shapes. The *ROIs* of different attention objects are also allowed to overlap. Generally, a *ROI* can be represented by a set of pixels in the original image. However, regular shaped *ROIs* can be denoted by their geometrical parameters instead of pixel sets

for simplicity. For example, a rectangular *ROI* can be defined as $\{Left, Top, Right, Bottom\}$ or $\{Left, Top, Width, Height\}$, while a circular *ROI* can be defined as $\{Center_x, Center_y, Radius\}$.

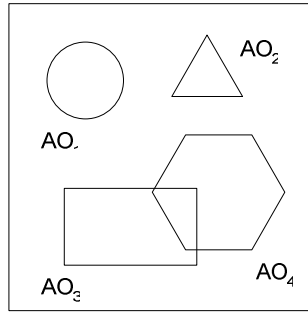


Fig. 1. Different attention objects in an image

2.2. Attention value

Since different attention objects carry different amount of information, they are of different importance. Therefore, we introduce *attention value (AV)*, a quantified value of user's attention on an attention object, as an indicator of the weight of each attention object in contribution to the information contained in the original image.

2.3. Minimal perceptible size

For image adaptation, we can apply resolution scaling, spatial cropping, quality compression, color reduction, and even text substitution to accommodate diverse client constraints. When fitting towards the small screen size, a natural and simple approach is directly down-sampling images to reduce their spatial sizes, but much information will be lost due to the resolution reduction.

Obviously, the information of an attention object is significantly relying on its area of presentation. If an attention object is scaled down too much, it may not be perceptible enough to let users catch the information that authors intend to deliver. Therefore, we introduce the *minimal perceptible size (MPS)* to represent the minimal allowable spatial area of an attention object. The *MPS* is used as a threshold to determine whether an attention object should be sub-sampled or cropped during the adaptation.

Suppose an image contains N number of attention objects, $\{AO_i\}$, $i = 1, 2 \dots, N$, where AO_i denotes the i th attention object within the image. The *MPS* of AO_i indicates the minimal perceptible size of AO_i , which can be presented by the area of scaled-down region. For instance, consider an attention object containing a human face whose original resolution is 75×90 pixels. The author or publisher may define its *MPS* to be 25×30 pixels which is the smallest resolution to show the face region without severely degrading its perceptibility.

2.4. Attention model description

Following the MediaProfile DS in MPEG-7 [8], the attention model description scheme syntax is described in Figure 2, which is extensible to various region shapes and prospective attributes in other models.

```
<complexType name="ImageAttentionModelType">
  <sequence >
    <element name="AO" type="AOType" minOccurs="0" maxOccurs="unbounded"/>
  </sequence >
  <attribute name="id" type="ID" use="optional"/>
</complexType>

<complexType name="AOType">
  <choice>
    <element name="RectangleROI" type="RectangleRegionType"/>
    <element name="CircleROI" type="CircleRegionType"/>
    <element name="PolygonROI" type="PolygonRegionType"/>
  </choice>
  <attribute name="AV" type="float"/>
  <attribute name="MPS" type="float"/>
  <attribute name="id" type="ID" use="optional"/>
  <attribute name="note" type="string" use="optional"/>
</complexType>

<complexType name="RectangleRegionType">
  <attribute name="RectLeft" type="nonNegativeInteger"/>
  <attribute name="RectTop" type="nonNegativeInteger"/>
  <attribute name="RectRight" type="nonNegativeInteger"/>
  <attribute name="RectBottom" type="nonNegativeInteger"/>
</complexType>

<complexType name="CircleRegionType">
  <attribute name="CircleCenterX" type="nonNegativeInteger"/>
  <attribute name="CircleCenterY" type="nonNegativeInteger"/>
  <attribute name="CircleRadius" type="nonNegativeInteger"/>
</complexType>

<complexType name="PolygonRegionType">
  <sequence >
    <element name="PolyV" type=" PolygonVertexType" minOccurs="3" maxOccurs="unbounded"/>
  </sequence >
</complexType>

<complexType name="PolygonVertexType">
  <attribute name="PointX" type="nonNegativeInteger"/>
  <attribute name="PointY" type="nonNegativeInteger"/>
  <attribute name="id" type="ID" use="optional"/>
</complexType>
```

Fig. 2. Attention model description scheme syntax

The description in Figure 3 presents the attention model of the example image shown in Figure 1 which contains four attention objects in different shapes, i.e. rectangle, circle, triangle, and hexagon, respectively.

```

<Model id="ExampleImage">
  <AO id="AO1" AV="0.3" MPS="201">
    <CircleROI CircleCenterX="39" CircleCenterY="39" CircleRadius="20"/>
  </AO>
  <AO id="AO2" AV="0.2" MPS="46" note="triangle">
    <PolygonROI>
      <PolyV id="Point1" PointX="98" PointY="17"/>
      <PolyV id="Point2" PointX="81" PointY="46"/>
      <PolyV id="Point3" PointX="116" PointY="46"/>
    </PolygonROI>
  </AO>
  <AO id="AO3" AV="0.1" MPS="1178" note="rectangle">
    <RectangleROI RectLeft="28" RectTop="92" RectRight="93" RectBottom="129"/>
  </AO>
  <AO id="AO4" AV="0.4" MPS="968" note="hexagon">
    <PolygonROI>
      <PolyV id="Point1" PointX="88" PointY="65"/>
      <PolyV id="Point2" PointX="120" PointY="65"/>
      <PolyV id="Point3" PointX="136" PointY="93"/>
      <PolyV id="Point4" PointX="120" PointY="121"/>
      <PolyV id="Point5" PointX="88" PointY="121"/>
      <PolyV id="Point6" PointX="72" PointY="93"/>
    </PolygonROI>
  </AO>
</Model>

```

Fig. 3. An example description of image attention model built from Figure 1

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 generate each attention object automatically and then build up the entire image attention model. For instance, the *ROI* of many semantic objects such as human faces and text units can be detected using some visual feature analysis algorithms and the *AV* and *MPS* values can be generated based on some predefined rules.

3. Automatic attention modeling

By analyzing an image, we can extract many visual features (including color, shape, and texture) that can be used to generate a saliency-based attention model as [18]. In addition, special objects like human faces and texts tend to attract most of user’s attention. In this section, we discuss in detail a variety of visual attention models we used for modeling image attention and a framework to integrate them together.

3.1. Saliency attention model

Itti *et al.* have defined a saliency-based visual attention model for scene analysis [10]. In this paper, we adopt the approaches in [10] to generate the three channel saliency maps, *color contrasts*, *intensity contrasts*, and *orientation contrasts* by using the approaches proposed and then build the final saliency map using the iterative method proposed in [11].

As illustrated in [16], the saliency attention is determined by the number of saliency regions, and their brightness, area, and position in the gray saliency map, as shown in Figure 4. However,

in order to reduce adaptation time, similar to [16], we binarize the saliency map to find the regions that most likely attract human attention. That is,

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

where $B_{i,j}$ denotes the brightness of pixel point (i, j) in the saliency region R , $W_{saliency}^{i,j}$ is the position weight of that pixel. Since people often pay more attentions to the region near the image center, a normalized Gaussian template centered at the image is used to assign the position weight.



Fig. 4. Examples of saliency attention detection

We use some heuristic rules to calculate the *MPS* of salient region. For example, bigger regions can be scaled down more aggressively than smaller ones, so their *MPS* is smaller in ratio. Since saliency maps are always in arbitrary shapes with little semantic meanings, a set of *MPS* ratios are predefined as the general *MPS* thresholds.

3.2. Face attention model

Face is one of the most salient characters of human beings and the appearance of dominant faces in images certainly attracts viewers' attention. Therefore, face attention model should be integrated into the image attention model to enhance the performance.

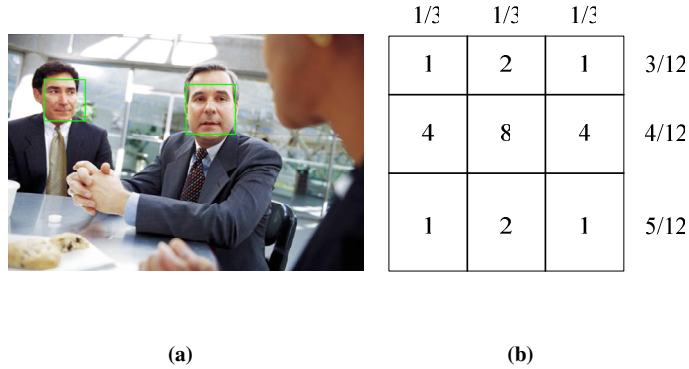


Fig. 5. (a) An example of face detection and (b) the position weight of the face model

By employing the face detection algorithm in [14], we obtain the face information including the number of faces, and the pose, region, and position of each face. In our current system, seven face poses (with out-plane rotation) can be detected, from the frontal to the profile. Figure 5(a) shows an example of face detection in a photo. We observe that the importance of a detected face is usually reflected by its region size and position. Hence,

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

where $Area_{face}$ denotes the size of a detected face region and W_{face}^{pos} is the weight of its position defined in Figure 5(b), which is the same as [16].

To our experience, a perceptible human face must be larger than a threshold to be recognized. So the MPS of face attention model can be predefined as an absolute area size. In our experiments, we define the MPS of face to be $25 \times 30 = 750$ pixels.

3.3. Text attention model

Similar to human faces, text regions also attract viewers' attention in many situations. Thus, they are also useful in deriving image attention models. There have been so many works on text detection and recognition and the localization accuracy can reach around 90% for text larger than 10 points.

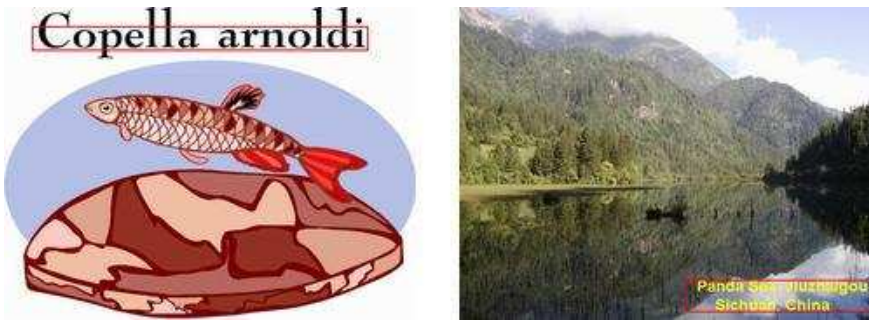


Fig. 6. Two examples of text detection

By adopting the text detection module in [3], we can find most of the informative text regions inside images. Two examples of text detection are shown in Figure 6. Similar to the face attention model, the region size is also used to compute the attention value of a text region. In addition, we include the aspect ratio of region to the calculation in the consideration that important text headers or titles are often in an isolated single line with large heights whose aspect ratios are quite different from text paragraph blocks. The attention value is defined as

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

where $Area_{text}$ denotes the size of a detected text region, and $W_{AspectRatio}$ is the weight of its aspect ratio generated by some heuristic rules. Different from face attention model, the position of text seldom indicates its importance, so that the position weight is not included here. The unit of text region can be a block, a sentence or even a single word depending on the granularity of text detection modules.

Similar to the face attention model, the MPS of a text region can also be predefined according to the font size, which can be calculated by text segmentation from the region size of text. For example, the MPS of normal text can be assigned from a specific 10 points font size in height. This threshold can be adjusted according to different users or environments.

3.4. Attention model adjustment

To combine multiple visual attention measurements, we need an adjustment on each attention value before integrating them together. Currently, we use a rule-based approach to modify the values because of its effectiveness and simplicity. For instance, if human faces or texts with large area are detected, their attention values should be considered more important than the saliency model because of the rich semantic meaning they carry. Otherwise, the saliency model is the determinative factor in the final attention model. The attention value of AO in each model should be first normalized to $(0, 1)$ and the final attention value is computed as

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

where w_k is the weight of model k and $\overline{AV_i^k}$ is the normalized attention value of AO_i detected in the model k , e.g. saliency model, face model, text model, or any other available model.

Based on our observation, images from different applications do not catch viewers' attention in the same way. For example, human faces in a home photo may attract most of viewers' attention, while they are far less noticeable in a scenery picture where the main purpose is to show the beautiful scene instead of some small faces captured in background. Thus, the weight w_k of attention values in each model needs to be adjusted for different image purposes.

We classify image functions into five different classes: *news pictures*, *home photos*, *sports shots*, *artistic graphs*, and *scenery pictures*, among which image authors or publishers can manually select an appropriate one to achieve the best adaptation results. For images belonging to each of these categories, different rules will be used to generate weights for attention value adjustment.

It is also worth noticing that when adapting images contained in a composite document such as a Web page, the image contexts are quite influential to user's attention. Thus, it is important to accommodate this variation in modeling image attentions. In our previous work [2], an efficient Function-based Object Model (*FOM*) was proposed to understand author's intention for each object in a Web page. For example, images in a Web page may have different functions, such as information, navigation, decoration or advertisement, etc. By using *FOM* analysis, the context of an image can be detected to assist image attention modeling.

Currently, the time cost for automatic modeling process is a bit large (usually 5~10s for an 800x600 image on our PII 450 test bed with 128M memory). However, the automatic modeling results can be saved with the original images in the syntax we proposed in Section 2.4 for reuse. Therefore, we calculate the attention model only once at the first time.

4. Attention-based image adaptation

Based on the previously described image attention model, the image adaptation problem can be better handled to accommodate user's attention. In the following we will discuss a technique to transform the problem into integer programming and a branch-and-bound algorithm to find the optimal image adaptation under resource constraints on client devices.

4.1. Information fidelity

Information fidelity is the perceptual ‘look and feel’ of a modified version of content object, a subjective comparison with the original version. The value of information fidelity is between 0 (lowest, all information lost) and 1 (highest, all information kept just as original). Information fidelity gives a quantitative evaluation of content adaptation that the optimal solution is to maximize the information fidelity of adapted content under different client context constraints. The information fidelity of an individual AO after adaptation is decided by various parameters such as spatial region size, color depth, ratio of compression quality, etc.

For an image region R consisting of several AOs , the resulting information fidelity is the weighted sum of the information fidelity of all AOs in R . Since user’s attention on objects always conforms to their importance in delivering information, we can directly employ attention values of different AOs as the informative weights of contributions to the whole perceptual quality. Thus, the information fidelity of an adapted result can be described as

$$IF(R) = \sum_{ROI_i \subset R} AV_i \cdot IF_{AO_i} \quad (6)$$

4.2. Adapting images on small displays

Given the image attention model, now let us consider how to adapt an image to fit into a small screen which is often the major limitation of mobile devices. We address the problem of making the best use of a target area T to represent images while maintaining their original spatial ratios. Various image adaptation schemes can be applied to obtain different results. For each adapted result, there is a corresponding unique solution which can be presented by a region R in the original image. In other words, an adapted result is generated from the outcome of scaling down its corresponding region R . As screen size is our main focus, we assume the color depth and compression quality does not change in our adaptation scheme.

Because in most situations the target area is rectangular and smaller than the original region of adapted result, the region of result R is a rectangle in the following discussion. But note that our model does not require the target area to be a rectangle.

According to Equation (6), an objective measure for the information fidelity of an adapted image can be formulated as follows:

$$\begin{aligned} IF(R) &= \sum_{ROI_i \subset R} AV_i \cdot IF_{AO_i} \\ &= \sum_{ROI_i \subset R} AV_i \cdot u\left(r_R^2 \cdot \text{size}(ROI_i) - MPS_i\right) \end{aligned} \quad (7)$$

where $u(x)$ is defined as

$$u(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

$\text{size}(x)$ is a function which calculates the area of a *ROI*, and r_R denotes the ratio of image scaling down, which can be computed as

$$r_R = \min\left(\frac{\text{Width}_T}{\text{Width}_R}, \frac{\text{Height}_T}{\text{Height}_R}\right) \quad (8)$$

Here, Width_T , Height_T , Width_R , and Height_R are the widths and heights of target area T and solution region R , respectively. As can be seen from Figure 7, when adapting an image to different target areas, the resulting solution regions may be different.

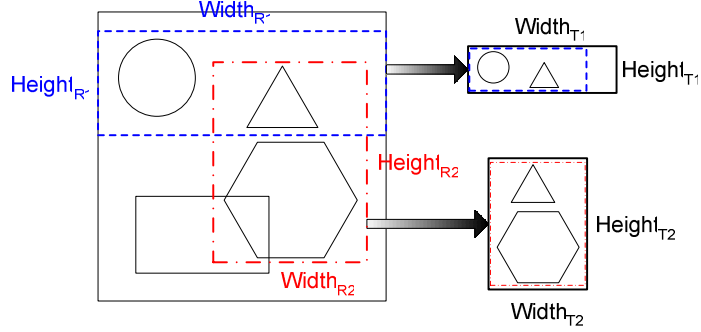


Fig. 7. Different solution regions for different target areas

We can use this quantitative value to evaluate all possible adaptation schemes to select the optimal one, that is, the scheme achieving the largest *IF* value. Taking the advantage of our image attention model, we transform the problem of making adaptation decision into the problem of searching a region within the original image that contains the optimal *AO* set (i.e. carries the most information fidelity), which is defined as follows:

$$\max(\text{IF}(R)) = \max_R \left\{ \sum_{ROI_i \subset R} AV_i \cdot u(r_R^2 \cdot \text{size}(ROI_i) - MPS_i) \right\} \quad (9)$$

4.3. Image adaptation algorithm

As we can see, for an image with width m and height n , the complexity for finding the optimal solution of (9) is $O(m^2n^2)$ because of the arbitrary location and size of a region. Since m and n may be quite large, the computational cost could be expensive. However, since the information fidelity of adapted region is solely decided by its attention objects, we can greatly reduce the computation time by searching the optimal *AO* set before generating the final solution.

4.3.1. Valid attention object set

We introduce I as a set of *AOs*, $I \subset \{AO_1, AO_2, \dots, AO_N\}$. Thus, our first step of optimization is to find the *AO* set that carries the largest information fidelity after adaptation. Let us consider R_I , the tight bounding rectangle containing all the *AOs* in I . We can first adapt R_I to the target area T , and then generate the final result by extending R_I to satisfy the requirements.

To our notice, not all of the AOs within a given region R can be perceptible when scaling down R to fit a target area T . To reduce the solution space, we define a valid attention object set as

Definition 2: An attention object set I is valid if

$$\frac{MPS_i}{size(ROI_i)} \leq r_l^2, \quad \forall AO_i \in I \quad (10)$$

where r_l (r_l is equivalent to r_R in Equation (8) for simplicity) is the ratio of scaling down when adapting the tight bounding rectangle R_l to T , which can be computed as below:

$$\begin{aligned} r_l &= \min \left(\frac{Width_T}{Width_l}, \frac{Height_T}{Height_l} \right) \\ &= \min \left(\frac{Width_T}{\max_{AO_i, AO_j \in I} |Right_i - Left_j|}, \frac{Height_T}{\max_{AO_i, AO_j \in I} |Bottom_i - Top_j|} \right) \end{aligned} \quad (11)$$

Here, $Width_l$ and $Height_l$ denote the width and height of R_l , while $Left_i$, $Right_i$, Top_i , and $Bottom_i$ are the four bounding attributes of the i^{th} attention object.

r_l in Definition 2 is used to check scaling ratio, which should be greater than $\sqrt{MPS_i / size(ROI_i)}$ for any AO_i belonging to a valid I . This ensures that all AO included in I is perceptible after scaled down by a ratio r_l . For any two AO sets I_1 and I_2 , there has $r_{l_1} \geq r_{l_2}$, if $I_1 \subset I_2$. Thus, it is straightforward to infer the following property of validity from Definition 2.

Property 1: If $I_1 \subset I_2$ and I_1 is invalid, then I_2 is invalid.

With our definition of valid attention object set, the problem of Equation (9) can be further simplified as follows:

$$\begin{aligned} \max_I (IF(R_l)) &= \max_I \left(\sum_{AO_i \in I} AV_i \cdot u(r_l^2 \cdot size(ROI_i) - MPS_i) \right) \\ &= \max_I \left(\sum_{AO_i \in I} AV_i \right) \quad \forall \text{ valid } I \subset \{AO_1, AO_2, \dots, AO_N\} \end{aligned} \quad (12)$$

As can be seen, this is a typical integer programming problem and the optimal solution can be found by a branch and bound algorithm.

4.3.2. Branch and bound process

As shown in Figure 8, let us consider a binary tree in which

- a) Each level presents the inclusion of a different AO ;
- b) Each node denotes a different set of AOs ;
- c) Each bifurcation means the alternative of keeping or dropping the AO of next level.

Thus, the height of this AO tree is N , the number of AO s inside the image, and each leaf node in this tree corresponds a possible I .

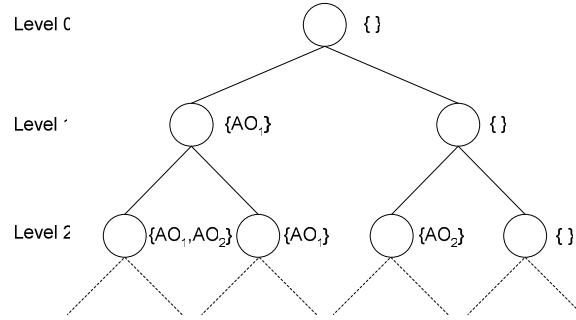


Fig. 8. The binary tree used for searching the optimal solution

For each node in the binary AO tree, there is a boundary on the possible IF value it can achieve among all of its sub-trees. Obviously, the lower boundary is just the IF value currently achieved when none of the unchecked AO s can be added, that is, the sum of IF values of AO s included in current configuration. And the upper boundary is the addition of all IF values of those unchecked AO s after current level, in other words, the sum of IF values of all AO s in the image except those dropped before current level.

Whenever the upper bound of a node is smaller than the best IF value currently achieved, the whole sub-tree of that node will be truncated. At the same time, for each node we check the ratio r_I of its corresponding AO set I to verify its validity. If it is invalid, according to Property 1, the whole sub-tree of that node will also be truncated. By checking both the bound on possible IF value and the validity of each AO set, the computation cost is greatly reduced.

We also use some techniques to reduce the time of traversal as listed below:

- Arrange the AO s in a decreasing order of their AV s at the beginning of search, since in most cases only a few AO s contribute the majority of IF value.
- While traveling along to a new level k , first check whether AO_k is already included in current configuration. If so, just go on travel the branch of keeping AO_k and prune the one of dropping AO_k and all sub-branches.

4.3.3. Transform to final adapted solution

After finding the optimal AO set I_{opt} , we can generate different possible solutions according to different requirements by extending $R_{I_{opt}}$ while keeping I_{opt} valid.

If an image has some background information which is not included in the attention model, the adapted result should present a region as large as possible by extending $R_{I_{opt}}$ as shown in

Figure 9(a). The scaling ratio of final solution region should be $r_{I_{opt}}^{\max} = \max_{AO_i \in I_{opt}} (MPS_i / \text{size}(ROI_i))$

in order to keep I_{opt} valid as well as to obtain the largest area. Therefore, we extend $R_{I_{opt}}$ to a region determined by $r_{I_{opt}}^{\max}$ and T , within the original image.

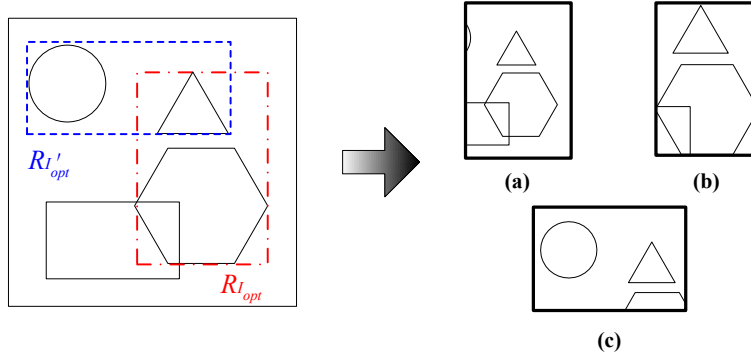


Fig. 9. Different solutions generated from a same image according to: (a) larger area, (b) higher resolution, and (c) larger area with rotation

In other cases, as shown in Figure 9(b), the adapted images may be more satisfactory with higher resolution than larger area. Thus, we should extend $R_{I_{opt}}$ similarly while keeping the scaling ratio at $r_{I_{opt}}$ instead of $r_{I_{opt}}^{\max}$. However, it is worth noticing that in this situation, the scaled version of whole image will perhaps never appear in adapted results.

To our observation, sometimes a better view can be achieved when the screen is rotated by a 90 degree as shown in Figure 9(c) where I_{opt}' carries more information than I_{opt} . In this case, we compare the result with the one for the rotated target area, and then select the better one as the final solution.

The complexity of this algorithm is exponential with the number of attention objects within an image in the worst case. However, our approach can be conducted efficiently, because the number of attention objects in an image is often less than a few dozens and the attention values are always distributed quite unevenly among attention objects. The experimental results in Section 5 verified the efficiency of this algorithm.

5. Experimental results

We have implemented an adaptive image browser to validate the performance of our proposed schemes. With the image attention model and the corresponding adaptation algorithm, the browser down-samples the resolution and relocates the viewing region to achieve the largest information fidelity while preserving satisfying perceptibility. This browser provides not only the adapted view of important regions, but also the “cropped” parts of original image by scrolling, enabling the users to have the overall view of entire image. An image can be adapted to arbitrary display sizes as well as several typical different display resolutions in a set of devices including Desktop PC, Hand-held PC, Pocket PC, TV browser, and Smartphone.

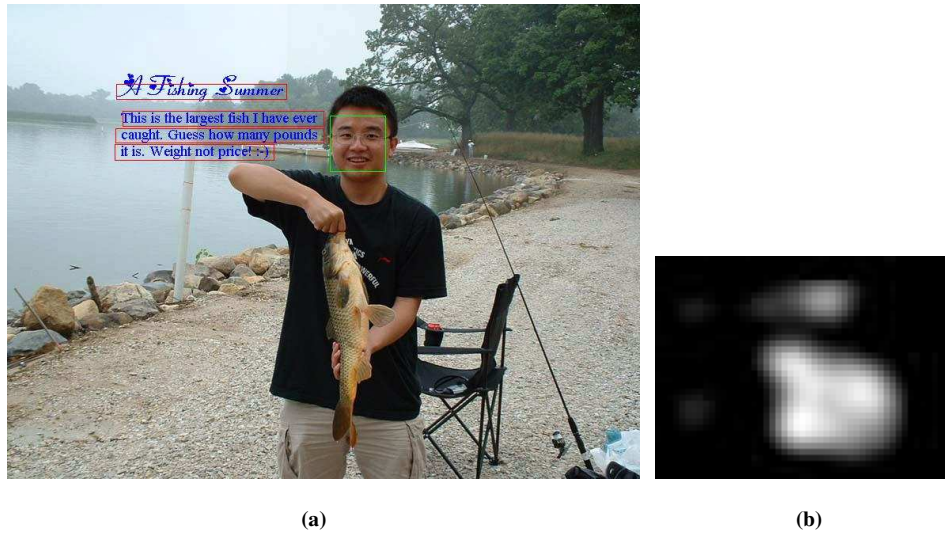


Fig. 10. The attention model of a home photo. (a) face and text detected, (b) saliency attention modeled



Fig. 11. An example comparing the proposed attention model with the conventional approaches to adapt image. (a) direct down-sampling for Pocket PC, (b) attention-based method for Pocket PC, (c) a better adapted result by rotation for Pocket PC, (d) direct down-sampling for Smartphone, and (e) attention-based method for Smartphone

Figure 10 shows an image from a personal photo collection with all three models (face, text, and saliency) built automatically. The adapted images are shown in Figure 11, compared with the direct down-sampling method. As can be seen in Figure 11(a), the most informative text in the image is hardly recognizable because of down-sampling when fitting into the typical screen size of Pocket PCs (240x320 pixels). In contrast, our method provides a much clearer view of the text region together with the salient face as shown in Figure 11(b). If we take rotation into consideration, a better result is achieved in Figure 11(c) where both the text and salient face are visible. However, when the image is to be adapted for a further smaller display, such as Smartphone with 120x160 screen in Figure 11(d), the text regions are not perceptible any more due to the limitation of scaling ratio. In this case, we perform a search for the optimal adaptation based on the new constraint, which results in a solution of the center region that contains the detected face and the brightest saliency regions as shown in Figure 11(e). The highest information fidelity is achieved by this solution.

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

Fifteen volunteers were invited to give their subjective scores on the adapted results by our approach while comparing with results from direct down-sampling, the most common method. We choose test data from various types of images with different sizes many of which are obtained from popular Web sites, such as MSN, YAHOO, USATODAY, etc. 56 images, with sizes varying from 388x242 to 1000x1414, were selected as our test dataset

The images were divided into two groups. The images of the first group were classified and modeled manually, while the images of the second group were automatically modeled without category classification. Two separate experiments were conducted to compare the automatic approach with the manual one. In both experiments, the subjects are required to give an assessment of being better, worse, or no difference to the adapted results by our approach compared with those by direct down-sampling method. Experimental results of manual and automatic modeling are listed in Table 1 and Table 2, respectively, where the percentages denote the average proportions of subjects' assessments.

As can be seen, in the first experiment, more than 71% of subjects consider our solution better than the conventional method and only 16% of them consider worse. However, it is worth noticing that in the scenery class, our approach had a quite low score compared with the direct scaling down method. It is actually reasonable because a scenery picture typically uses the entire picture to present a scene and its important information spreads out all over the picture. The result in the second experiment on automatic modeling is also satisfactory, although it is a bit worse than the manual one due to the immature detecting technologies.

We also conducted an experiment on the efficiency of our algorithm by logging the computational time costs while making 10 times adaptation for each image. Here we only include the time cost for the searching procedure introduced in Section 4. Our test bed is a Dell OptiPlex GX1p with PII 450 CPU, 128M memory and Windows 2000 Professional system. We got an average time cost at 22 microseconds, i.e. about 45,000 images per second, with variation from 6

to 95 microseconds. Without code optimization, our technique is already fast enough to be employed in real-time adaptive content delivery system on either proxy servers or content servers, or even on client devices.

Table 1. The results of evaluation for image adaptation based on manual attention modeling

Image Class	Better	No Diff.	Worse
News Picture	80.67%	7.33%	12.00%
Home Photo	75.24%	16.19%	8.57%
Sports Shot	65.00%	11.67%	23.33%
Artistic Graph	66.67%	20.00%	13.33%
Scenery	26.67%	16.67%	56.67%
Total	71.28%	12.05%	16.67%

Table 2. The results of evaluation for image adaptation based on automatic attention modeling

	Better	No Diff.	Worse
Assessment	67.56%	12.00%	20.44%

6. Conclusion

In this paper, we proposed a novel solution for adapting image contents based on user attention to fit into heterogeneous client display sizes. The main contributions of our work are two-fold: a new scheme to model user attention in viewing images, and an algorithm to utilize such modeling in image adaptation and browsing.

Most existing work on image adaptation is mainly focusing on saving file size, while our point is to adapt to all context constraints among which screen size is the most critical one. Our approach is not only scaling, compressing and cropping images, but also help locating perceptually important regions when the user is browsing images. Compared with [13], which is the most relevant work that we know, our proposed scheme provides a better performance because of the proposed image attention model together with the automatic modeling process and the developed efficient search algorithm.

In addition to adapting images for small-form-factor devices, our approach can be easily applied to other various applications such as thumbnail generation. We are currently developing an authoring tool to assist the generation of different attention models for an image. It will provide an editor interface for authors, publishers, or viewers to customize the models. Also we are considering improving those adaptation parameters in our attention model by learning user feedbacks instead of rule-based approaches. With the satisfactory results from our experiments, we plan to extend the attention model to other media types, such as video and Web pages. For example, by incorporating the attention model for video summarization in [16], we could apply the principle of our proposed image adaptation to attention-based video clipping. If we consider a Web page containing multiple spatial blocks as an image containing multiple attention objects, our image attention model can be used to adapt the layout of a Web page to assist Web browsing on small-form-factor devices. We will continue to investigate these directions in our future work.

7. Acknowledgements

We would like to express our special appreciation to Yu Chen for his 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.

References

1. Chandra S, Gehani A, Ellis CS, Vahdat A (2001) Transcoding Characteristics of Web Images. Proc. SPIE (Multimedia Computing and Networking 2001) 4312: 135-149
2. Chen JL, Zhou BY, Shi J, Zhang HJ, Wu QF (2001) Function-based Object Model Towards Website Adaptation. Proc. 10th Int. World Wide Web Conf., May 2001, Hong Kong, China, pp. 587-596
3. Chen XR, Zhang HJ (2001) Text Area Detection from Video Frames. Proc. 2nd IEEE Pacific-Rim Conf. on Multimedia (PCM2001), October 2001, Beijing, China, pp. 222-228
4. Christopoulos C, Skodras A, Ebrahimi T (2000) The JPEG2000 Still Image Coding System: An Overview. IEEE Trans. on Consumer Electronics 46(4): 1103-1127
5. Fox A, Gribble S, Brewer EA, Amir E (1996) Adapting to Network and Client Variability via On-Demand Dynamic Distillation. Proc. 7th Int. Conf. on Architectural Support for Programming Languages and Operating Systems, October 1996, Cambridge, USA, pp. 160-170
6. Han R, Bhagwat P, Lamair R, Mummert T, Perret V, Rubas J (1998) Dynamic Adaptation in an Image Transcoding Proxy for Mobile Web Access. IEEE Personal Communications 5(6): 8-17
7. ISO/IEC JTC1/SC29/WG11/N4242 (2001) ISO/IEC 15938-5 FDIS Information Technology – Multimedia Content Description Interface – Part 5 Multimedia Description Schemes. July 2001, Sydney, Australia
8. ISO/IEC JTC1/SC29/WG11/N4674 (2002) MPEG-7 Overview. March 2002, Jeju, Korea
9. ISO/IEC JTC1/SC29/WG11/N4819 (2002) MPEG-21 Digital Item Adaptation. May 2002, Fairfax, USA
10. Itti L, Koch C, Niebur E (1998) A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence 20(11): 1254-1259
11. Itti L, Koch C (1999) A Comparison of Feature Combination Strategies for Saliency-Based Visual Attention System. Proc. SPIE (Human Vision and Electronic Imaging IV), 3644: 473-482
12. Itti L, Koch C (2001) Computational Modeling of Visual Attention. Nature Reviews Neuroscience 2(3): 194-203
13. Lee K, Chang HS, Chun SS, Choi L, Sull S (2001) Perception-based Image Transcoding for Universal Multimedia Access. Proc. 8th Int. Conf. on Image Processing (ICIP-2001), October 2001, Thessaloniki, Greece, vol. 2, pp. 475-478
14. Li SZ, Zhu L, Zhang ZQ, Blake A, Zhang HJ, Shum H (2002) Statistical Learning of Multi-View Face Detection. Proc. 7th European Conference on Computer Vision (ECCV 2002), May 2002, Copenhagen, Denmark, vol. 4, pp. 67-81
15. Ma WY, Bedner I, Chang G, Kuchinsky A, Zhang HJ (2000) A Framework for Adaptive Content Delivery in Heterogeneous Network Environments. Proc. SPIE (Multimedia Computing and Networking 2000) 3969: 86-100
16. Ma YF, Lu L, Zhang HJ, Li MJ (2002) An Attention Model for Video Summarization. To appear in ACM Multimedia, December 2002
17. Mohan R, Smith JR, Li CS (1999) Adapting Multimedia Internet Content for Universal Access. IEEE Trans. on Multimedia 1(1): 104-114

18. Salah AA, Alpaydin E, Akarun L (2002) A Selective Attention-based Method for Visual Pattern Recognition with Application to Handwritten Digit Recognition and Face Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(3): 420-425
19. Smith JR, Mohan R, Li CS (1998) Content-based Transcoding of Images in the Internet. *Proc. 5th Int. Conf. on Image Processing (ICIP-98)*, October 1998, Chicago, USA, vol. 3, pp. 7-11
20. World Wide Web Consortium (1999) Web Content Accessibility Guidelines 1.0. May 1999, <http://www.w3.org/tr/wai-webcontent/>