

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

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

<sup>1</sup> Dept. of Automation, University of Science and Technology of China, Hefei, 230027, P.R. China)

<sup>2</sup> Microsoft Research Asia, 5/F Sigma Center, No. 49 Zhichun Road, Beijing, 100080, P.R. China

Part of this work was originally presented at the 9th International Conference on Multimedia Modeling (MMM'03).

**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 with a view to saving file size under constraints in client environment and have hardly paid attention to user perceptions of the adapted result. Meanwhile, the major limitation on the user's delivery context is moving away 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. 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; DS, description scheme

### 1 Introduction

As Internet content, 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 increasingly large variety of clients so as to attract customers. At the same time, more and more client users in a heterogeneous environment want all the information to be suitable for universal access [16,21], 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 as visual content, it is essential to make images 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 PCs, pocket PCs, and Smartphone, are experiencing a population boom with Internet mobile clients (with the exception of the original desktop PCs) because of their portability and mobility. Although these client devices are becoming increasingly powerful in both computing power and data storage, nevertheless, low-bandwidth connections and small displays – the two crucial limitations on accessing the current Internet – are still a great obstacle to their widespread use. The bandwidth condition is expected to be greatly improved with the development of 2.5-G and 3-G wireless networks, while the display size is more likely to remain unchanged due to the mobility requirement of these devices. In this paper, we would like to focus on the latter, that is, adapting images for devices with limited screen size.

Much attention has focused on visual content adaptation. Related fields come from quite different viewpoints 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 various resolutions to terminals with different capabilities in terms of display and bandwidth. In the MPEG-7 Multimedia Description Schemes [8,9], Media Profile DS 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, Media Transcoding Hints DS and Media Quality DS, of the Media Profile DS are designed to provide information for content adaptation and reduce its computational complexity by specifying transcoding

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

Correspondence to: Xing Xie (e-mail: xingx@microsoft.com)

ing hints of the media being described and representing both subjective quality ratings and objective quality ratings, respectively. Currently, MPEG-21 has started to work on defining an adaptation framework called Digital Item Adaptation [10] 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 data-type-specific content adaptation was proposed in [6]. In particular, adaptation such as image distillation (i.e., compression) 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 et al. [18,20] 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. Based on the analysis results, the transcoding system chooses 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 [7] proposed a framework for determining when/whether/how to transcode images in an 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], which focused 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 [14] 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 many approaches for adapting images, most of them focus only on compressing and caching contents in the Internet in order to reduce 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 pointing out that the algorithms involving large numbers of adaptation rules or significant computation are impracticable in the on-the-fly systems of adaptive content delivery.

Aiming at solving the limitations in current algorithms while avoiding semantic analysis, in this paper we present a generic image adaptation framework based on a viewer attention model. Attention is a neurobiological conception. It means the concentration of mental powers on an object, a close or careful observing or listening. 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 for content analysis in adaptive content delivery by providing the exact information to facilitate the decision making.

The computational attention methodologies have been studied by some researchers. The authors of [13] reviewed re-

cent works on computational models of focal visual attention and presented a bottom-up, saliency-based 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 dynamic neural network [11,12]. A selective attention-based method for visual pattern recognition was presented in [19] together with promising results when applying this method in handwritten digit recognition and face recognition. Recently, Ma et al. [17] 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 image content to optimally match the various screen sizes of client devices based on modeling of viewer 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 the image attention model. In Sect. 3, several methods for automatic modeling of user attention on visual features are presented. Section 4 describes in detail a new approach of image adaptation based on the attention model. The performance of adaptation is evaluated in a user study experiment, with the results reported in Sect. 5. Finally, Sect. 6 provides concluding remarks and discussions on future work.

## 2 Image attention model

The visual attention model for an image is defined as a set of attention objects. 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 AO often represents a semantic object such as a human face, a flower, an automobile, a text sentence, etc. Generally, most perceptible information of an image can be located inside a handful of attention objects; at the same time these AOs catch the most attentions of a user. Therefore, the image adaptation problem can be treated as manipulating AOs to provide as much information as possible under resource constraints.

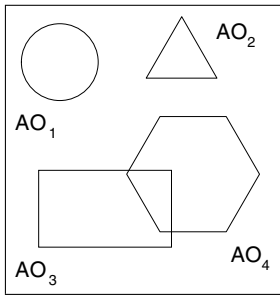
We assign three attributes to each AO – are region-of-interest (ROI), attention value (AV), and minimal perceptible size (MPS). Each is introduced in detail below.

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

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

where

$AO_i$ ,	the $i$ th AO within image
$ROI_i$ ,	ROI of $AO_i$
$AV_i$ ,	AV of $AO_i$
$MPS_i$ ,	MPS of $AO_i$
$N$ ,	total number of AOs in image



**Fig. 1.** Various attention objects in an image

### 2.1 Region-of-interest

We borrow the notion of *region-of-interest (ROI)* from JPEG 2000 [4], which is referred to in our model as a spatial region or segment within an image that corresponds to an AO. As shown in Fig. 1, ROIs can be in arbitrary shapes. The ROIs of various AOs 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  $\{Center.x, Center.y, Radius\}$ . In fact, since most existing images and displays are rectangular, a tight bounding rectangle can be used to cover nongeometric ROIs without any information loss. Also, this simplification accelerates the computation of optimization search in image adaptation.

### 2.2 Attention value

Since different AOs carry varying amount of information, they are of varying importance. Usually, the more attentions an object attracts, the more important it is in delivering information. Therefore, we introduce *attention value (AV)*, a quantified value of user attention on an AO, as an indicator of its 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 for a 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.

The information of an AO relies heavily on its area of presentation. If an AO is scaled down too much, it may not be perceptible enough to let users catch the information its author intends 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 subsampled 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 AO 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 with  $75 \times 90$  pixels. The author or publisher may define its MPS to be  $25 \times 30$  pixels, which is the smallest resolution to show the face region without severely degrading its perceptibility.

### 2.4 Attention model description

Following the Media Profile DS in MPEG-7 [9], the attention model description scheme syntax is described in Fig. 2. It is extensible to various region shapes and prospective attributes in other models.

The description in Fig. 3 presents the attention model of the sample image shown in Fig. 1 containing four attention objects in the shapes of rectangle, circle, triangle, and hexagon.

The image attention model can be manually assigned by authors or publishers. This, however, could be labor intensive. A more plausible approach is to generate each AO 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 in [19]. In addition, special objects like human faces and texts tend to attract most of a 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.

### 3.1 Saliency attention model

Itti et al. have defined a saliency-based visual attention model for scene analysis [11]. In this paper, we adopt the approaches in [11] 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 [12].

As illustrated in [17], 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 Fig. 4. However, in order to reduce adaptation time (as in [17]), we binarize the saliency map to find the regions that most likely attract human attention, i.e., using a binarization threshold on brightness, which is estimated adaptively. Thus the saliency attention value 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  $R$ , and  $W_{saliency}^{i,j}$  is the position weight of that pixel. Since people often pay more attention to the region near the image center, a normalized Gaussian template centered at the image is used to assign the position weight.

```

<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

We use some heuristic rules to calculate the MPS of a salient region. For example, larger 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

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

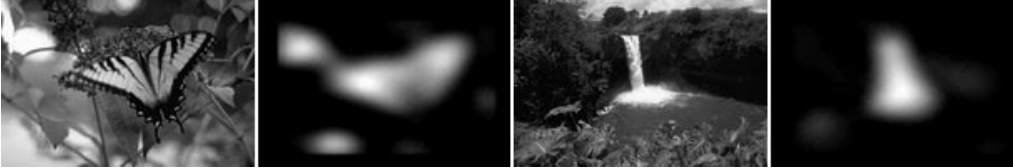
By employing the face detection algorithm in [15], we obtain face information including the number of faces and the pose, region, and position of each face. In our current system, seven face poses (with out-plane rotation) can be detected,

```

<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 sample description of image attention model built from Fig. 1



**Fig. 4.** Examples of saliency attention detection

from the frontal to the profile. Figure 5a 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 Fig. 5b, which is the same as [17].

In our experience, a perceptible human face must be larger than a threshold to be recognized. So the MPS of the 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

Like human faces, text regions also attract viewers' attention in many situations. Thus they are also useful in deriving image attention models. There has been much work done on text

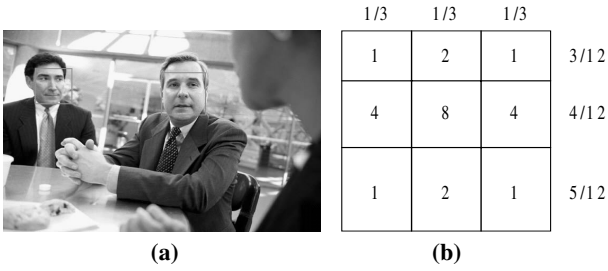
detection and recognition, and localization accuracy can reach around 90% for text larger than 10 point.

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 Fig. 6. Like the face attention model, the region size is also used to compute the AV of a text region. In addition, we include the aspect ratio of a region in the calculation because 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 AV 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. Unlike the 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.

Like the face attention model, the MPS of a text region can also be predefined according to the font size, which can



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

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-point 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 to adjust each attention value before integrating them. 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 a large area are detected, their AVs 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. We normalize the final adjusted AVs as:

$$AV_i = w_k \cdot AV_i^k / \sum_i AV_i^k \quad (5)$$

where  $w_k$  is the weight of model  $k$  and  $AV_i^k$  is the attention value of  $AO_i$  detected in 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 viewers' attention, while they are far less noticeable in a landscape picture where the main purpose is to show the beautiful scenery instead of some small faces captured in the 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*. 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 AV adjustment.

It is also worth noting that when adapting images are contained in a composite document such as a Web page, the image contexts are quite influential on user 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 an author's intention for each object in a Web page. For example, images on 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 cost for automatic modeling process is considerable (usually 5 ~ 10 s for an  $800 \times 600$  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 Sect. 2.4 for reuse. Therefore, we calculate the attention model only once 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 a 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 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 retained). Information fidelity gives a quantitative evaluation of content adaptation. The optimal solution is to maximize the information fidelity of adapted content under various 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 a 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, we now consider how to adapt an image to fit onto a small screen. 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 that 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 do not change in our adaptation scheme.

Because in most situations the target area is rectangular and smaller than the original region of the adapted result, the

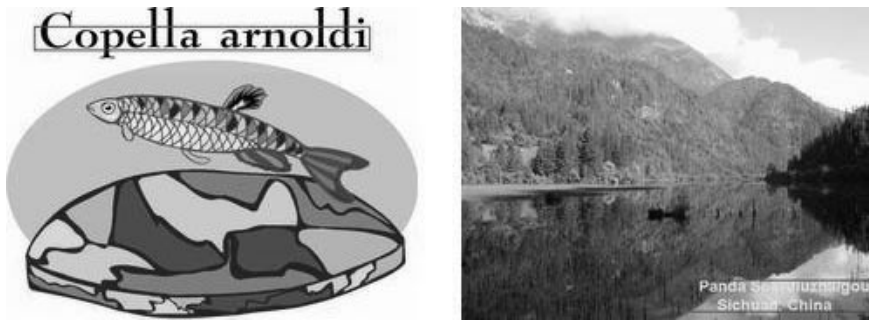


Fig. 6. Two examples of text detection

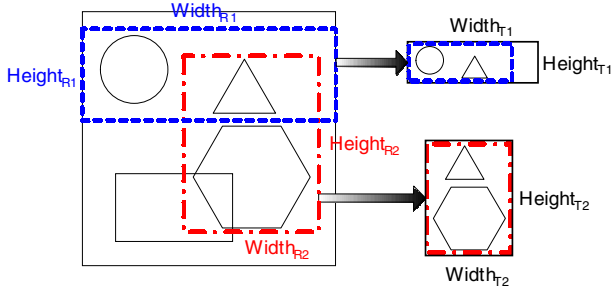


Fig. 7. Different solution regions for different target areas

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. If the target area is not rectangular or is even nongeometric, which is, however, quite rare in the real world, some modifications are necessary to deal with those AOs with arbitrary shapes.

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

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

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 that 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,  $\text{Width}_T$ ,  $\text{Height}_T$ ,  $\text{Width}_R$ , and  $\text{Height}_R$  are the widths and heights of target area  $T$  and solution region  $R$ , respectively. As seen in Fig. 7, when adapting an image to different target areas, the resulting solution regions may be different. It is worth noting that, although the rectangle AO overlaps with the hexagon AO in Fig. 7, they are treated independently.

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. Using our image attention model, we transform the problem of making an 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:

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

### 4.3 Image adaptation algorithm

As we can see, for an image with width  $m$  and height  $n$ , the complexity of finding the optimal solution of Eq. 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 the adapted region is decided solely 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 in optimization is to find the AO set that carries the largest IF 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.

In fact, not all of the AOs within a given region  $R$  are perceptible when scaling down  $R$  to fit a target area  $T$ . To reduce the solution space, we define a valid AO set as

**Definition 2:** An AO set  $I$  is valid if

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

where  $r_I$  ( $r_I$  is equivalent to  $r_R$  in Eq. 8 for simplicity) is the ratio of scaling down when adapting the tight bounding

rectangle  $R_I$  to  $T$ , which can be computed as follows:

$$r_I = \min \left( \frac{Width_T}{Width_I}, \frac{Height_T}{Height_I} \right)$$

$$= \min \left( \frac{Width_T}{\max_{AO_i, AO_j \in I} |Right_i - Left_j|}, \right. \quad (11)$$

$$\left. \frac{Height_T}{\max_{AO_i, AO_j \in I} |Bottom_i - Top_j|} \right) \quad (12)$$

Here,  $Width_I$  and  $Height_I$  denote the width and height of  $R_I$ , while  $Left_i$ ,  $Right_i$ ,  $Top_i$ , and  $Bottom_i$  are the four bounding attributes of the  $i$ th AO.

$r_I$  in Definition 2 is used to check the 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 AOs included in  $I$  are perceptible after being scaled down by a ratio  $r_I$ . For any two AO sets  $I_1$  and  $I_2$ , there is  $r_{I_1} \geq r_{I_2}$ , 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 AO set, the problem of Eq. 9 can be further simplified as follows:

$$\max_I (IF(R_I)) \quad (13)$$

$$= \max_I \left( \sum_{AO_i \in I} AV_i \cdot u(r_I^2 \cdot size(ROI_i) - MPS_i) \right)$$

$$= \max_I \left( \sum_{AO_i \in I} AV_i \right) \quad \forall \text{ valid } I \subset \{AO_1, AO_2, \dots, AO_N\}$$

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 Fig. 8, let us consider a binary tree in which

- each level presents the inclusion of a different AO,
- each node denotes a different set of AOs, and
- each bifurcation means the alternative of keeping or dropping the AO of the next level.

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

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

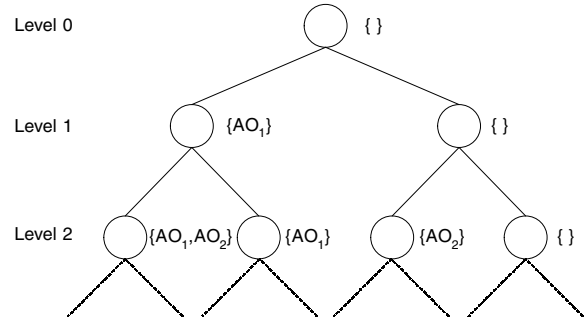


Fig. 8. The binary tree used for searching the optimal solution

Whenever the upper bound of a node is smaller than the best IF value currently achieved, the whole subtree 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 subtree of that node will also be truncated. By checking both the bound on possible IF values 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 AOs in decreasing order of their AVs at the beginning of the search, since in most cases only a few AOs contribute the majority of IF values.
- While traveling to a new level  $k$ , first check whether  $AO_k$  is already included in the current configuration. If it has been included, the subtree which does not include  $AO_k$  will be pruned.

### 4.3.3 Transform to final adapted solution

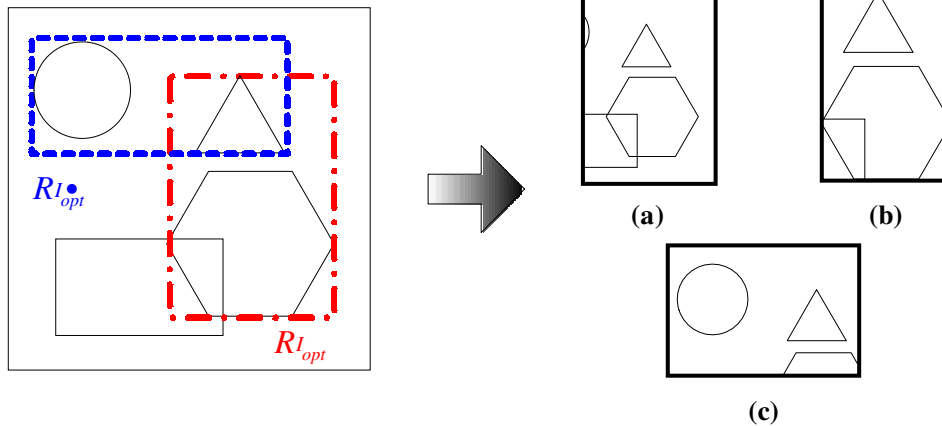
After finding the optimal AO set  $I_{opt}$ , we can generate different possible solutions according to different requirements by extending  $R_{I_{opt}}$  while keeping  $I_{opt}$  valid.

If an image has some other information that 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 Fig. 9a. The scaling ratio of a final solution region should be  $r_{I_{opt}}^{\max} = \max_{AO_i \in I_{opt}} (MPS_i / size(ROI_i))$  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.

In other cases, as shown in Fig. 9b, the adapted images may be more satisfactory with a higher resolution than a 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 noting that in this situation, the scaled version of the whole image will perhaps never appear in the adapted results.

Our observations indicate that sometimes a better view can be achieved when the screen is rotated by  $90^\circ$  as shown in Fig. 9c, 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.

In the worst case, the complexity of this algorithm increases exponentially with the number of AOs within an image. However, our approach can be conducted efficiently because



**Fig. 9.** Various solutions generated from the same image according to: **a** larger area, **b** higher resolution, and **c** larger area with rotation

the number of AOs in an image is often less than a few dozen and the attention values are always distributed quite unevenly among AOs. The experimental results in Sect. 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 IF while preserving satisfactory. This browser provides not only the adapted view of important regions, but also the “cropped” parts of the original image by scrolling, which enables users to have an overall view of the 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, handheld PC, pocket PC, TV browser, and 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 Fig. 11, compared with the direct down-sampling method. As seen in Fig. 11a, the most informative text in the image is hardly recognizable because of down-sampling when fitting into the typical screen size of pocket PCs ( $240 \times 320$  pixels). In contrast, our method provides a much clearer view of the important text region as shown in Fig. 11b. If we take rotation into consideration, a better result is achieved in Fig. 11c, 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  $120 \times 160$  screen in Fig. 11d, the text regions are no longer perceptible due to the limitation of the 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 Fig. 11e. The highest IF is achieved by this solution.

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

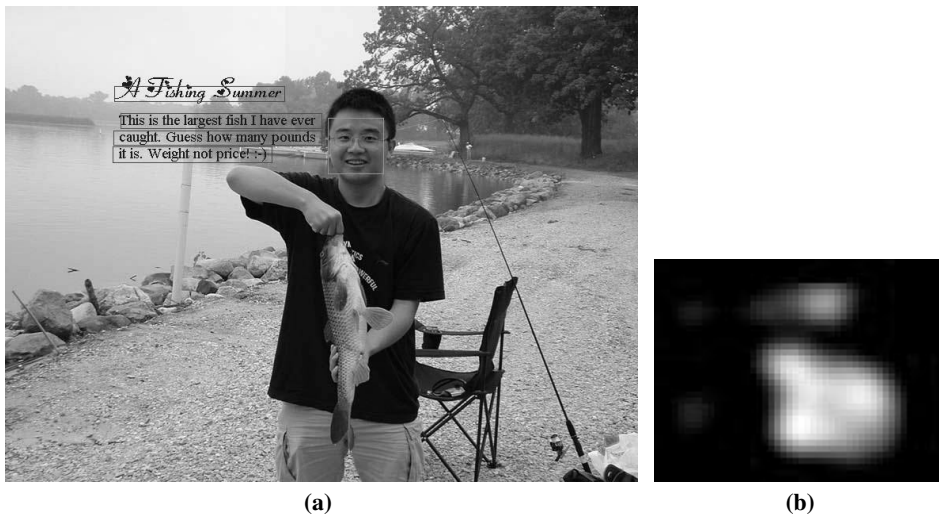
Fifteen volunteers were invited to give their subjective scores on the adapted results by our approach while com-

paring results from direct down-sampling, the most common method. We chose test data from various types of images with different sizes, many of which were obtained from popular Web sites such as MSN, YAHOO, USATODAY, etc. Fifty-six images, with sizes varying from  $388 \times 242$  to  $1000 \times 1414$ , were selected as our test dataset.

The images were divided into two groups. The first group was classified and modeled manually, while the second group was automatically modeled without category classification. Two separate experiments were conducted to compare the automatic approach with the manual one. In both experiments, the subjects were required to give an assessment of being better, worse, or no different than the adapted results by our approach compared with those by the direct down-sampling method. Experimental results of manual and automatic modeling were 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 considered our solution better than the conventional method and only 16% of them considered it worse. However, it is worth noting that in the scenery class, our approach had quite a low score compared with the direct scaling-down method. This is in fact 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 was also satisfactory, although it was 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. Here we only included the time cost for the searching procedure introduced in Sect. 4. Our test bed was a Dell OptiPlex GX1p with PII 450 CPU, 128 MB memory, and Windows 2000 Professional system. We got an average time cost at  $22 \mu\text{s}$  over ten runs, i.e., about 45,000 images per second, with variation from 6 to  $95 \mu\text{s}$ . Without code optimization, our technique was already fast enough to be employed in a real-time adaptive content delivery system on either proxy servers or content servers, or even on client devices.



**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 adapting images. **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. **e** Attention-based method for Smartphone

**Table 1.** The results of evaluation of 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%
<b>Total</b>	<b>71.28%</b>	<b>12.05%</b>	<b>16.67%</b>

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

	Better	No Diff.	Worse
<b>Assessment</b>	<b>67.56%</b>	<b>12.00%</b>	<b>20.44%</b>

## 6 Conclusion

In this paper, we proposed a novel solution for adapting image content based on user attention to fit into heterogeneous client display sizes. The main contributions of our work are twofold: 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 preserving the file size, while our aim is to adapt to all context constraints, among which screen size is the most critical one. Our approach involves not only scaling, compressing, and cropping images, but also helping to locate perceptually important regions when the user is browsing images. Compared with [14], which is the most relevant work that we know, our scheme provides better performance by combining the proposed image attention model 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 various other applications such as thumbnail generation. More advanced UI approaches can be employed to improve the usability, as in [5]. We are currently developing an authoring tool to assist in the generation of different attention models for an image. It will provide an editor interface for authors, publishers, or viewers to customize the models. We are also considering improving those adaptation parameters in our attention model by learning user feedback instead of rule-based approaches. With the satisfactory results of 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 [17], 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 AOs, 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.

*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. Finally, the authors are very grateful to all the reviewers for their valuable comments.

## References

- Chandra S, Gehani A, Ellis CS, Vahdat A (2001) Transcoding characteristics of Web images. *Proc SPIE (Multimedia Comput Network 2001)* 4312:135–149
- Chen JL, Zhou BY, Shi J, Zhang HJ, Wu QF (2001) Function-based object model towards website adaptation. In: *Proceedings of the 10th international World Wide Web conference*, Hong Kong, May 2001, pp 587–596
- Chen XR, Zhang HJ (2001) Text area detection from video frames. In: *Proceedings of the 2nd IEEE Pacific-Rim conference on multimedia (PCM2001)*, Beijing, October 2001, pp 222–228
- Christopoulos C, Skodras A, Ebrahimi T (2000) The JPEG2000 still image coding system: an overview. *IEEE Trans Consumer Electron* 46(4):1103–1127
- Fan X, Xie X, Ma WY, Zhang HJ, Zhou HQ (2003) Visual attention based image browsing on mobile devices. In: *Proceedings of the IEEE international conference on multimedia and expo (ICME 03)*, Baltimore, July 2003
- Fox A, Gribble S, Brewer EA, Amir E (1996) Adapting to network and client variability via on-demand dynamic distillation. In: *Proceedings of the 7th international conference on architectural support for programming languages and operating systems*. Cambridge, MA, October 1996, pp 160–170
- Han R, Bhagwat P, Lamaire R, Mummert T, Perret V, Rubas J (1998) Dynamic adaptation in an image transcoding proxy for mobile Web access. *IEEE Pers Commun* 5(6):8–17
- ISO/IEC JTC1/SC29/WG11/N4242 (2001) ISO/IEC 15938-5 FDIS Information technology – multimedia content description interface – Part 5: Multimedia description schemes. Sydney, July 2001
- ISO/IEC JTC1/SC29/WG11/N4674 (2002) MPEG-7 Overview. Jeju, Korea, March 2002
- ISO/IEC JTC1/SC29/WG11/N4819 (2002) MPEG-21 Digital item adaptation. Fairfax, VA, May 2002
- Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans Pattern Analysis Mach Intell* 20(11):1254–1259
- Itti L, Koch C (1999) A comparison of feature combination strategies for saliency-based visual attention system. *Proc SPIE (Hum Vis Electron Imag IV)* 3644:473–482
- Itti L, Koch C (2001) Computational modeling of visual attention. *Nat Rev Neurosci* 2(3):194–203
- Lee K, Chang HS, Chun SS, Choi L, Sull S (2001) Perception-based image transcoding for universal multimedia access. In: *Proceedings of the 8th international conference on image processing (ICIP-2001)*, Thessaloniki, Greece, October 2001, 2:475–478
- Li SZ, Zhu L, Zhang ZQ, Blake A, Zhang HJ, Shum H (2002) Statistical learning of multi-view face detection. In: *Proceedings of the 7th European conference on computer vision (ECCV 2002)*, Copenhagen, May 2002, 4:67–81
- Ma WY, Bedner I, Chang G, Kuchinsky A, Zhang HJ (2000) A framework for adaptive content delivery in heterogeneous network environments. *Proc SPIE (Multimedia Comput Network 2000)* 3969:86–100
- Ma YF, Lu L, Zhang HJ, Li MJ (2002) A user attention model for video summarization. In: *Proceedings of the 10th ACM international conference on multimedia*, Juan-les-Pins, France, December 2002, pp 533–542

18. Mohan R, Smith JR, Li CS (1999) Adapting multimedia internet content for universal access. *IEEE Trans Multimedia* 1(1):104–114
19. 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 Pattern Analysis Mach Intell* 24(3):420–425
20. Smith JR, Mohan R, Li CS (1998) Content-based transcoding of images in the Internet. In: Proceedings of the 5th international conference on image processing (ICIP-98), Chicago, October 1998, 3:7–11
21. World Wide Web Consortium (1999) Web content accessibility guidelines 1.0. May 1999, <http://www.w3.org/tr/wai-webcontent/>