Contrast-based Image Attention Analysis by Using Fuzzy Growing

Yu-Fei Ma, Hong-Jiang Zhang
Microsoft Research Asia
5F, Beijing Sigma Center, 49 Zhichun Road
Beijing 100080, China
{yfma, hjzhang}@microsoft.com

ABSTRACT
Visual attention analysis provides an alternative methodology to semantic image understanding in many applications such as adaptive content delivery and region-based image retrieval. In this paper, we propose a feasible and fast approach to attention area detection in images based on contrast analysis. The main contributions are threefold: 1) a new saliency map generation method based on local contrast analysis is proposed; 2) by simulating human perception, a *fuzzy growing* method is used to extract attended areas or objects from the saliency map; and 3) a practicable framework for image attention analysis is presented, which provides three-level attention analysis, i.e., attended view, attended areas and attended points. This framework facilitates visual analysis tools or vision systems to automatically extract attentions from images in a manner like human perception. User study results indicate that the proposed approach is effective and practicable.

Categories and Subject Descriptors
I.4.10 [**Image Processing and Computer Vision**]: Image Representation – *Statistical, Hierarchical.*
I.4.6 [**Image Processing and Computer Vision**]: Image Segmentation – *Region growing, partitioning, Pixel classification.*

General Terms
Algorithms, Design, Experimentation, Theory.

Keywords
Attention detection, visual attention model, contrast analysis, fuzzy growing, image analysis.

1. INTRODUCTION
The effective information retrieval in a large image library involves two key issues, one is retrieval accuracy, and the other is the suitable representations in various displays. It is widely recognized that region-based image retrieval is able to achieve higher accuracy than that based on global image features [1]. Also, due to a variety of display screens used by users, especially those small display screens in PDA (Personal Digital Assistant), mobile phone and the size alterable webpage browsers, an adaptive image display scheme is needed [2]. One of the core issues in the two popular applications is to determine the important and representative regions in whole image. Obviously, if the semantics of each region or object are known, this issue will be easily solved. However, this task is beyond the capability of contemporary computer vision systems. On the other hand, the perceptual attention mechanism plays an important role in biological vision and human cognition, which enable humans to filter and prioritize incoming information. If such an attention mechanism can be modeled, it can then be applied to determine the important regions in an image and serve the needs of region-based image retrieval and adaptive image delivery/browsing. According to these crucial requirements, this paper presents a three-level framework, composing of attended view, attended areas and attended points, from top to bottom, for fast image attention analysis.

Attention is at the nexus between cognition and perception. The control of the focus of attention may be goal-driven and stimulus-driven which corresponds to top-down and bottom-up processes in human perception, respectively. The study of attention involved in a few fields, including biology, psychology, neuro-psychology, cognitive science and computer vision. Although the attention mechanism is not completely understood yet, some proven conclusions can be used to guide its applications. The earlier attention research began with William James, who was the first person to outline a theory of human attention [3]. Successively, Broadbent proposed his filter theory of attention in an attempt to explain many of the existing experimental results [4]. The response selection theory of attention was proposed by Deutsch [5], who indicated that a part of attention involves high level processing. This process is called late selection in the later studies. From 1960s, Treisman proposed a series of models that combined early and late selection into a model known as Feature Integration Theory (FIT) [6]. Treisman's recent study believes that early selection is most active when the perceptual load is high, whereas late selection (object-based and location-based) is used when perceptual load is low [7]. Besides, advances in a neurophysiological model of attention were also made by Koch [8, 9].

In recent years, especially with emerging interest in active vision, computer vision researchers have been increasingly concerned with attention mechanisms as well. Consequently, a number of

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

computational attention models were developed, such as the models proposed in [10, 11]. The basic principles behind these efforts are greatly influenced by psychophysical research. Niebur proposed a computational attention model in [12]. He indicated that the so-called "focus of attention" scans the scene both in the form of a rapid, bottom-up, saliency-driven and task-independent manner and in a slower, top-down, volition-controlled and task-dependent manner. Based on the work in [8, 13, 14], Itti proposed a saliency-based visual attention model for scene analysis in [15]. In this work, visual input is first decomposed into a set of topographic feature maps which all feed into a master "saliency map" in a bottom-up manner. Then, a WTA (Winner-Take-All) competition is employed to select the most conspicuous image locations as attended points. In primates, such a map is believed to be located in the posterior parietal cortex as well as in the various visual maps in the pulvinar nuclei of the thalamus.

Another well known computational visual attention model is VISIT proposed by Ahmad [16], which is more biologically-plausible than Itti's. VISIT consists of a gating network which corresponds to the pulvinar and whose output, the gated feature maps, corresponds to the areas V4, IT and MT of the optic nerve, a priority network corresponding to the superior colliculus, frontal eye field and posterior parietal areas, a control network corresponding to the posterior parietal areas, and a working memory corresponding to the prefrontal cortex.

Both Itti's framework and Ahmad's model build up an elegant mapping from computational implementation to biological theories. However, their high computational complexity requires massively parallel method to obtain fast responses. This drawback usually exists in all biological structure based attention models. On the other hand, the study of human attention mechanism is not mature yet. Therefore, if we are only concerned with such high level applications as region-based image retrieval and browsing, it is not necessary to strictly reproduce biological structures by computer algorithms. For example, we have proposed a computational motion attention model for video skimming in [17] and a user attention framework for video summarization in [18]. We also proposed a pure computational algorithm for salient region extraction from video [19], which is mainly based on the relationship between texture and human perception. In [18], we present a comprehensive solution for user attention analysis in video, which integrates a series of algorithms of video, audio and image attention analysis. Itti's model was also employed in this user attention model as static image attention analysis module. Although the work in [19] partly solves the issue of image attention, it focuses on the efficiency of video processing and is not a total solution for image attention analysis. Meanwhile, Itti's model needs massive computations, which makes it impracticable to be implemented on personal computers. Consequently, an alternative, more powerful and more practicable solution for image attention analysis is desirable.

Instead of only verifying human perception mechanism by computer algorithms, we proposed an image attention analysis framework according to the characters of digital images and the capabilities of personal computers in this paper. This framework extracts three-level attentions from image, namely, attended view, attended areas, and attended points. The attended view means the main part of image with balance composition and rich information. The attended areas are those regions which represent important semantics and most possibly attract human attentions. Whereas, the attended points, without semantics, are those locations which perceive local maximum stimulus. Such three-level attentions are able to support a number of applications related to image analysis. In our work, the primary principles about human attention mechanism explored in the aforementioned works are utilized as only high-level guidance. Firstly, we investigated the key factors in human visual perception. The conclusion is that contrast is the most important factor which dominantly influences human visual perception. Therefore, a contrast-based saliency map is proposed as an attention presentation of image, like Itti's work. Based on such saliency map, attended points and attended view are directly extracted. In addition, a *fuzzy growing* process is proposed to simulate the process of human perception, by which the attended areas are extracted from saliency map. The satisfactory user study results show that the proposed approach is an effective and fast solution for image attention analysis.

The rest of this paper is organized as follows: Section 2 introduces the concept and computation method of contrast-based saliency. In Section 3, *fuzzy growing* is discussed in detail. Based on the methods presented in Section 2 and Section 3, Section 4 addresses contrast-based image attention analysis framework as well as attention extraction methods. Additionally, the specific applications of the proposed framework are also discussed at the end of this section. Section 5 displays and discusses the evaluation results obtained from user study experiment. Finally, Section 6 concludes the paper.

2. CONTRAST-BASED SALIENCY MAP

Contrast is an important parameter in assessing vision. Visual acuity measurement in the clinic uses high contrast, that is, black letters on a white background. In reality, objects and their surroundings are of varying contrast. Therefore, the relationship between visual acuity and contrast allows a more detailed understanding of human visual perception [20]. Traditional image processing techniques usually consider an image by three basic properties, color, texture, and shape. These techniques have been successfully applied to a number of applications. However, they cannot provide high level understanding of image, because human usually do not simply understand an image from color, texture, and shape aspects separately. In fact, a basic principle behind the three basic characters is contrast. In other words, whether an object can be perceived or not depends on the distinctiveness between itself and its environment.

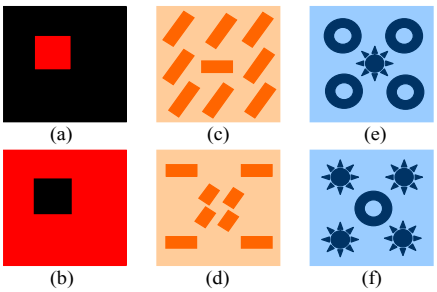

Figure 1. Contrast behind color, texture and shape perception

As shown in Figure 1, there are three pairs of synthesized images. In Figure 1(a), there is a red box on black background. It goes without saying that the attended area is the red box. Red color is

usually considered as bright color which easily attracts human attentions. However, Figure 1(b) cannot support this assumption. Obviously, black box becomes attended area though red background occupies most of image. This phenomenon indicates that the color and size are not most pivotal factor for human perception, although human visual sensitivity is influenced by color and size. On the contrary, color contrast plays an important role in this perception process. Figure 1(c) and (d) show two textured images with oriented dashes. The weak texture area is surrounded by the strong texture patches in (c), while the contrary case is shown in (d), the strong texture area being surrounded by the weak texture patches. Similarly to color, the strength of texture does not greatly influence human perception, but contrast still does. The same conclusion can also be drawn from Figure 1(e) and (f). The complexity of shape is not the main factor in human perception either. From above comparison experiments, we may make a supposition that the regions with high contrast have rich information and are most likely to attract human attentions.

Based on this observation, we proposed a contrast-based saliency measure in this work. There have been a number of methods to compute contrast, such as color contrast and luminance contrast. However, a generic contrast is needed in this work. We define an effectual area perceiving stimulus as *perceive field*, which is equivalent to receptive field in human eye. An image with the size of $M \times N$ pixels can be regarded as a *perceive field* with $M \times N$ *perception units*, if each *perception unit* contains one pixel. The contrast value $\boldsymbol{C}_{i,j}$ on a *perception unit* (i, j) is defined as follows:

$$\boldsymbol{C}_{i,j} = \sum_{\boldsymbol{q} \in \Theta} \boldsymbol{d}\left(\boldsymbol{p}_{i,j}, \boldsymbol{q}\right) \tag{1}$$

where $\boldsymbol{p}_{i,j}$ ($i \in [0, M]$, $j \in [0, N]$) and $\boldsymbol{q}$ denote the stimulus perceived by *perception units*, such as color. $\boldsymbol{\Theta}$ is the neighborhood of *perception unit* (i, j). The size of $\boldsymbol{\Theta}$ controls the sensitivity of *perceive field*. The smaller the size of $\boldsymbol{\Theta}$ is, the more sensitive the *perceive field* is. $\boldsymbol{d}$ is the difference between $\boldsymbol{p}_{i,j}$ and $\boldsymbol{q}$, which may employ any suitable distance measure according to applications.

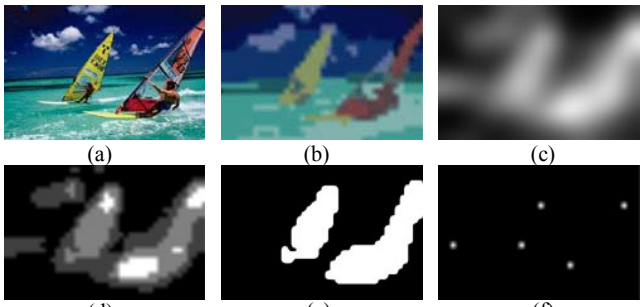

Figure 2. Contrast-based image attention analysis (a) original image; (b) quantized block image; (c) saliency map; (d) fuzzy growing; (e) attended areas; (f) attended points.

By normalizing to [0, 255], all contrasts $\boldsymbol{C}_{i,j}$ on the *perception units* form a saliency map. A sample of contrast-based saliency map is shown in Figure 2(c). In our implementation, the colors in LUV space are used as stimulus on *perceive field*, and the difference $\boldsymbol{d}$ is computed by *Gaussian* distance. The image attention analysis is performed on such local contrast based saliency map, because this kind of saliency map not only reflects color contrast, but also reflects the strength of texture. Moreover, the areas close to the boundary of objects lean to have same or similar contrasts. Therefore, this contrast-based saliency map presents color, texture and approximate shape information at the same time, which provides sufficient information for image attention analysis.

The more samples of saliency map are given in Figure 4. col.(c). Experimental results show that the saliency map generated by our approach is good representation of human attention, and the mechanism embedded it effectively supports the following processes of attention detection.

3. FUZZY GROWING

In order to extract attended areas from saliency map, a method called *fuzzy growing* is proposed. Saliency map is a gray-level image in which the bright areas are considered as attended areas, as shown in Figure 2(c). Obviously, a hard cut threshold is not effective for attended areas extraction, because the variation of gray-levels in saliency map are not consistent, even in one object. Consequently, conventional region growing approaches based on one strict measure cannot solve such problem well. In our implementation, fuzzy theory [21] is employed in region growing process, named *fuzzy growing*, because fuzzy theory has been proven to be effective to imitate human mental behaviors.

According to the definition of a fuzzy event [22], saliency map is regarded as a fuzzy event modeled by a probability space. We assume that saliency map has L gray levels from g_0 to g_{L-1} and the histogram of saliency map is h_k, k=0, …, L-1. Thus, the saliency map is modeled by a triplet ($\boldsymbol{\Omega}$, k, $\boldsymbol{P}$), where $\boldsymbol{\Omega}=\{g_0, g_1, \ldots, g_{L-1}\}$ and $\boldsymbol{P}$ is the probability measure of the occurrence of gray levels, i.e., $\boldsymbol{Pr}\{g_k\} = h_k/\sum h_k$. The membership function, $\mu_S(g_k)$, of a fuzzy set $\boldsymbol{S} \in \boldsymbol{\Omega}$ denotes the degree of some properties, such as attended areas, unattended areas, etc., possessed by gray level g_k. In fuzzy set notation, it is written as

$$\boldsymbol{S} = \sum_{g_k \in \Omega} \mu_S(\boldsymbol{g}_k) / \boldsymbol{g}_k \tag{2}$$

The probability of this fuzzy event can be computed by

$$\boldsymbol{P}(\boldsymbol{S}) = \sum_{k=0}^{L-1} \mu_S(\boldsymbol{g}_k) \boldsymbol{P}_r(\boldsymbol{g}_k) \tag{3}$$

For saliency map, there exist two classes of pixels, attended areas and unattended areas. We define the two classes as two fuzzy sets, denoted by $\boldsymbol{B}_A$, and $\boldsymbol{B}_U$, respectively, which are mutually exclusive. Thus, these two fuzzy sets partition saliency map $\boldsymbol{\Omega}$. In such fuzzy partition, there is no sharp boundary between the two fuzzy sets, which is like human perception mechanism. We adopt fuzzy c-partition entropy as the criterion to measure the fitness of a fuzzy partition. Theoretically, a fuzzy c-partition is determined by 2(c-1) parameters, and the problem becomes to find the best combinations of these parameters, which can be considered a combinatorial optimization problem. Usually, simulated annealing or genetic algorithms is used to solve this problem, but they are all time-consuming. Fortunately, only 2 parameters are needed in our algorithm due to 2-partition. Therefore, we use an exhaust search to find optimal result without involving high computational complexity.

In saliency map $\boldsymbol{\Omega}$, considering the two fuzzy events, attended areas $\boldsymbol{B}_A$ and unattended areas $\boldsymbol{B}_U$, the membership functions of

fuzzy events are defined in (4) and (5), respectively

$$\mu_A = \begin{cases} 1 & \boldsymbol{x} \geq \boldsymbol{a} \\ \dfrac{\boldsymbol{x}-\boldsymbol{u}}{\boldsymbol{a}-\boldsymbol{u}} & \boldsymbol{u} < \boldsymbol{x} < \boldsymbol{a} \\ 0 & \boldsymbol{x} \leq \boldsymbol{u} \end{cases} \tag{4}$$

$$\mu_U = \begin{cases} 0 & \boldsymbol{x} \geq \boldsymbol{a} \\ \dfrac{\boldsymbol{x}-\boldsymbol{a}}{\boldsymbol{u}-\boldsymbol{a}} & \boldsymbol{u} < \boldsymbol{x} < \boldsymbol{a} \\ 1 & \boldsymbol{x} \leq \boldsymbol{u} \end{cases} \tag{5}$$

where $\boldsymbol{x}$ is the independent variable denoting gray level and a and u are the parameters determining the shape of the above two membership functions. The optimal parameters a and u will be obtained, if an optimization objective function is satisfied. The gray-levels greater than a have the membership of 1.0 for fuzzy set B_A, which means the pixels with these gray-levels definitely belong to the attended areas. On the contrary, when the gray levels is smaller than u, the membership for fuzzy set B_A becomes 0, which means the pixels with these gray-levels do not belong to the attended areas. Similarly, B_U has opposite variation form. While, the pixels with the gray-levels between a and u have the membership of (0, 1) for fuzzy sets B_A and B_U according to the definition (4) and (5), respectively.

Assuming that the prior probabilities of the attended areas and the unattended areas are approximately equal, the optimal partition requires the difference between the prior entropies of attended areas and that of unattended areas reaches the minimum. Sahoo *et al*. proposed a minimal difference of entropy as metric to obtain optimal threshold for image segmentation [23]. We modify this measure according to fuzzy set definition as follows,

$$\boldsymbol{\Gamma}(\boldsymbol{a}, \boldsymbol{u}) = [\, \boldsymbol{H_A}(\boldsymbol{a}, \boldsymbol{u}) - \boldsymbol{H_U}(\boldsymbol{a}, \boldsymbol{u}) \,]^2 \tag{6}$$

where $H_A(a, u)$ and $H_U(a, u)$ are the prior entropies of fuzzy sets, attended areas and unattended areas, respectively. They are calculated by

$$\boldsymbol{H_A}(\boldsymbol{a}, \boldsymbol{u}) = -\sum_{k=0}^{L-1} \frac{\Pr(\boldsymbol{g_k})}{\boldsymbol{P}(\boldsymbol{B_A})} \ln \frac{\Pr(\boldsymbol{g_k})}{\boldsymbol{P}(\boldsymbol{B_A})} \tag{7}$$

$$\boldsymbol{H_U}(\boldsymbol{a}, \boldsymbol{u}) = -\sum_{k=0}^{L-1} \frac{\Pr(\boldsymbol{g_k})}{\boldsymbol{P}(\boldsymbol{B_U})} \ln \frac{\Pr(\boldsymbol{g_k})}{\boldsymbol{P}(\boldsymbol{B_U})} \tag{8}$$

where $\boldsymbol{P}(\boldsymbol{B_A}) = \sum_{k=0}^{L-1} \mu_A \, Pr(\boldsymbol{g_k})$ and $\boldsymbol{P}(\boldsymbol{B_U}) = \sum_{k=0}^{L-1} \mu_U \, Pr(\boldsymbol{g_k})$ according to (3).

The global minima of $\boldsymbol{\Gamma}(\boldsymbol{a}, \boldsymbol{u})$ indicates the optimal fuzzy partition, i.e., optimal parameters a and u are found. This criterion can be expressed as:

$$(a, u) = \textbf{arg min } (\boldsymbol{\Gamma}(\boldsymbol{a}, \boldsymbol{u})) \tag{9}$$

With the optimal a and u, *fuzzy growing* process is performed on the saliency map. A number of initial attention seeds are needed first. The criteria for seed selection are: 1) the seeds must have maximum local contrast; and 2) the seeds should belong to the attended areas. Sequentially, starting from each seed, the pixels with the gray-levels satisfying the following criteria will be grouped.

$$C_{i,j} \leq C_{seed} \text{ and } C_{i,j} > s \tag{10}$$

where $s = (a + u)/2$. The probabilities of the gray-level s belonging to attended areas and unattended areas are all 0.5, see (4) and (5). Then, the new group members are used as seeds to do iterative growing. Such *fuzzy growing* process simulates the bottom-up search process of human perception. Figure 2(d) illustrates fuzzy 2-partition of saliency map by three layers which denote the gray level higher than a (highest), s (middle), and u (lowest), correspondingly. Figure 2(e) shows the result of *fuzzy growing*, two main objects in scene being accurately detected and segmented.

4. IMAGE ATTENTION ANALYSIS

With the two techniques introduced in Section 2 and Section 3, we propose a framework for image attention analysis, which integrates bottom-up process and top-down process by simulating human perception. The top-down process is not the scope of this paper, which involves high level semantics understanding or late selection of human perception, such as object detection and recognition. However, as face detection technology has been approaching to mature, a face detection module [24] is integrated into our framework as a component of top-down process. The bottom-up process corresponding to early selection of human perception is the focus of this paper.

As shown in Figure 3, the image attention analysis framework is composed of two parts, bottom-up, and top-down. The former outputs three-level attentions based on saliency map from low level to high level, including attended points, attended areas, and attended view. The latter outputs the result of face detection. The two processes may have some interactions, which have been proven by neurobiology. However, how they interact is not very clear. Therefore, we simply implement a single direction process. That is, the face detection result may improve the outputs of attended areas and attended view.

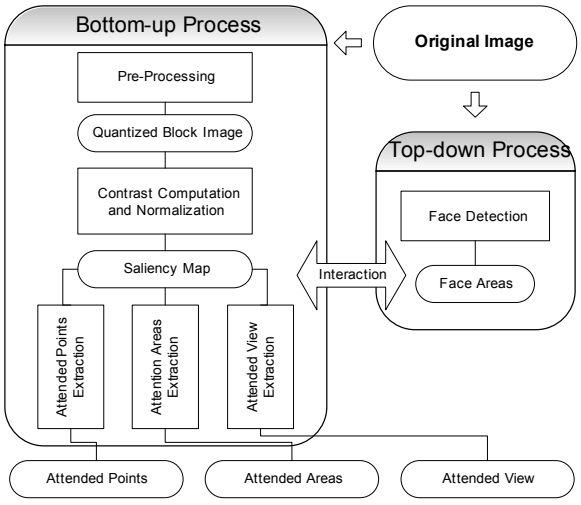

Figure 3. Architecture of image attention analysis

Within the workflow of bottom-up process, an original image passes through a pre-processing module first, by which the original image is transferred to a quantized block image, as shown in Figure 2(b). The pre-processing steps include:

1) Image resizing. First of all, the original image is resized to a uniform image with its aspect ratio unchanged. The advantages of resizing are twofold, that is, all images are considered in the same scale and the computational complexity is effectively reduced.

2) Color space transformation. As LUV space is consistent with human color perception system well, the resized image is transformed from RGB space to LUV space.

3) Color quantization. Human vision perception usually is more sensitive to the changes in smooth areas than in texture areas. So color quantization is performed to make color smooth in texture areas. Here, peer group filtering method [25] is employed.

4) Block dividing. In order to further smooth texture areas and reduce computational cost, the image is divided into the blocks with $n \times n$ pixels, namely, each *perception unit of perceive field* has $n \times n$ pixels. On each *perception unit*, the means of LUV elements are computed separately. In this manner, a quantized block image is obtained.

Based on such quantized block image, or *perceive field*, contrast is calculated on each *perception unit*. Then, the contrasts are smoothed and normalized to [0, 255] to form saliency map, as shown in Figure 2(c). With such contrast-based saliency map, the three-level attentions are extracted.

4.1 Attended Point Extraction

Attended point detection is analogous to the lowest level of human attention, which is directly caused by outside stimulus. Therefore, attended points do not have any semantics, which may be directly detected from saliency map by Winner-Take-All process like the work in [15]. That is, the attended points are those points with local maximum contrast in saliency map. Figure 2(f) shows a sample of attended points in image Figure 2(a).

4.2 Attended Area Extraction

Attended area extraction may be regarded as an extension of attend point detection, including two processes: seed selection and *fuzzy growing*. According to Section 3, the attended areas' seeds are the subset of attended points. We select those points with the contrast greater than a as seeds. Then, from each seed, *fuzzy growing* is performed until no candidate of *perception units* can be grouped. This process simulates early stage of human perception during which human search a semantic object looks like what they have seen. As shown in Figure 2(d) and (e), the reasonable results are obtained. If more than one initial seeds belong to one region/object, the areas grown from these seeds are progressively merged into an area during *fuzzy growing*.

4.3 Attended View Extraction

Motivated by one of Gestalt laws, an attention center as well as an attended view is extracted from the saliency map. The Gestalt school of psychology [26] has played a revolutionary role with its novel approach to visual form. Although the Gestalt theory is a non-computational theory of visual forms, which is a disadvantage for practical engineering applications, it provides some useful guidance.

According to one of Gestalt laws that *visual forms may possess one or several centers of gravity about which the form is organized* [27], we assume that there is a center of gravity in a saliency map, called attention center, which corresponds to the vision center of image. Based on this attention center, the whole image can be organized in the form of information maximization. In our implementation, we define an attended view as a rectangle $\boldsymbol{V}(\boldsymbol{C_0}, \boldsymbol{W}, \boldsymbol{H})$, where $\boldsymbol{C_0}$ denotes attention center, $\boldsymbol{W}$ and $\boldsymbol{H}$ are the width and height of rectangle respectively. If contrast level in saliency map is regarded as density, the attention center is the centroid of saliency map. Similarly, there is a relationship between the size of attended view and the 1st order central moment of saliency map. Specifically, let (x_0, y_0) denote attention center, and $(\boldsymbol{w'}, \boldsymbol{h'})$ denote the 1st order central moment of saliency map, the attention center and the attended view's width and height are computed by (11) and (12) respectively,

$$\begin{cases} x_0 = \dfrac{1}{CM}\displaystyle\sum_{j=0}^{N-1} C_{i,j} \times i \\ y_0 = \dfrac{1}{CM}\displaystyle\sum_{i=0}^{M-1} C_{i,j} \times j \end{cases} \tag{11}$$

where $CM = \displaystyle\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} C_{i,j}$ is the 0th order moment of saliency map.

$$\begin{cases} \boldsymbol{W} = 2\boldsymbol{w} = 2\alpha \cdot \boldsymbol{w}' \\ \boldsymbol{H} = 2\boldsymbol{h} = 2\alpha \cdot \boldsymbol{h}' \end{cases} \tag{12}$$

where $\alpha > 1$ is a constant coefficient. w' and h' are computed by the 1st order central moments of saliency map along x-axis and y-axis respectively, and the 0th order moment CM, expressed by (13).

$$\begin{cases} \boldsymbol{w}' = \dfrac{1}{\boldsymbol{CM}}\displaystyle\sum_{j=0}^{N-1} \boldsymbol{C}_{i,j} \times \left\| \boldsymbol{i} - \boldsymbol{x}_0 \right\| \\ \boldsymbol{h}' = \dfrac{1}{\boldsymbol{CM}}\displaystyle\sum_{i=0}^{M-1} \boldsymbol{C}_{i,j} \times \left\| \boldsymbol{j} - \boldsymbol{y}_0 \right\| \end{cases} \tag{13}$$

The process of attended view extraction can be viewed as the last stage of human perception. That is, when human complete their attention searching, they usually adjust their views according to the attention center of image and attention distributions in whole image. In Figure 4, col.(c) gives some examples of attended view.

Finally, if face detection results were available, attended areas and attended view could be improved according to the semantics provided by face rectangles, such as position of heads and shoulders. Actually, attended areas and attended view are also able to speed up the process of face searching, though it is not implemented in the current system. As our system does not strictly simulate the biological structure of human perception, the computational complexity is effectively reduced so that it can be easily implemented in high level applications of personal computers or integrated into real-time vision systems.

By using the three-level image attention analysis, the performance of information searching in large image library can be greatly improved in accuracy, speed and display aspects. The attended view effectively accelerates feature extraction during image retrieval by extracting a sub-image with the most important information. The attended areas provide more details about important areas for region-based image retrieval. Also, both attended view and attended areas may facilitate users to quickly browse important parts of image in a variety of display screens in different sizes. Although the attended points lack of semantics, they provide users possible search paths on images, which can be utilized to determine the browsing sequence of image regions. Therefore, the proposed image attention analysis framework not only provides effective supports to visual perception systems, but also can be used in such high level multimedia applications as [1] and [2]. In fact, the proposed framework can be employed by

video attention analysis system at the stage of salient region detection in frames, such as [19]. Meanwhile, this framework is also the supplement for our user attention model [18], which may significantly improve the performance of user attention model, both on effectiveness and efficiency.

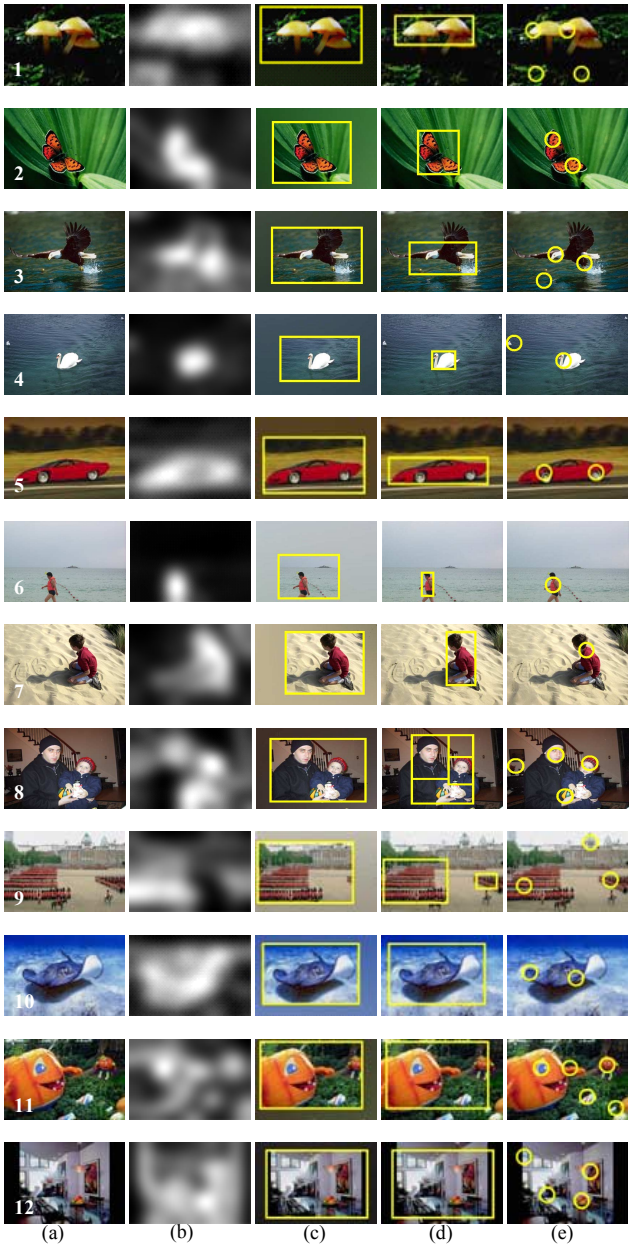

Figure 4A. Some sample results of attention analysis col.(a) original image, col.(b) contrast-based saliency map, col.(c) attended view, col.(d) attended areas, col.(e) attended points.

Figure 4 displays some results output from our system. It can be seen that our system is not sensitive to the scale of attended areas. Attended views successfully crop the main bodies of images, which are usually larger than attended areas. Most of detected attended areas give satisfactory results. For example, in Figure 4A(1-7) and 4B(2), the dominant objects are accurately detected. For some complex cases, our system may output multi-scale or multiple attended areas. For example, in Figure 4A(8), a family photo, three attended areas are detected, including the father, the son and the region containing both the father and the son. If there are multiple clearly separate dominant objects/regions in an image, our system may distinguish the different attended areas, such as Figure 4A(9). However, if the objects/regions are not clearly isolated, the connected objects/regions are grouped, see Figure 4B(1). Besides, we may find that the attended areas are the same as the attended views in Figure 4A(10-12) and 4B(3). In Figure 4A(10), the dominant object occupies almost main body of image, so the attended area equals the attended view coincidently. As the scenes are over diverse in Figure 4A(11-12), our system dose not segment attended areas particularly. Similarly, for portrait images (Figure 4B(3)), the grown attended areas may larger or slightly smaller than attended view due to the complexity of contrast distribution in saliency map, our system uses attended view as attended area. The all attended points detected are reasonable except in portrait images, such as Figure 4B(3). However, they still provide useful cues for attention search.

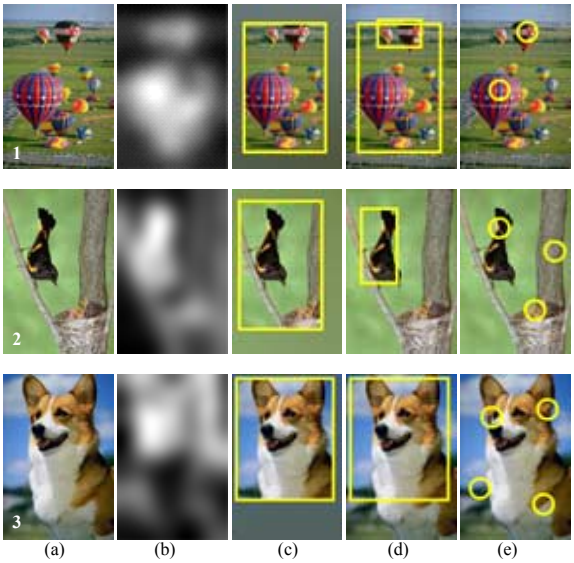

Figure 4B. Some sample results of attention analysis col.(a) original image, col.(b) contrast-based saliency map, col.(c) attended view, col.(d) attended areas, col.(e) attended points.

5. EVALUATIONS AND DISCUSSIONS

Due to the subjectivity of human attention perception, there is not a standardized objective correctness measure for image attention analysis evaluation. Therefore, we have carried out a user study experiment to evaluate the proposed approach. The experiment composes of three parts corresponding to the three levels of attention analysis. 20 human subjects were invited to take part in user study. These subjects are not computer experts and have no any special computer knowledge. They were required to assign one of assessments, GOOD, ACCEPT, or FAILED, to the results output by our system, i.e., attended view, attended areas, attended points, separately. In reality, it is very difficult to define what results are acceptable, because different persons have different criteria. Generally speaking, the acceptable cases include: 1) some not very important areas/points are missed; 2) the locations of view/areas/points are not very accurate.

The testing data used in this experiment are collected from three resources: 750 images are selected from Corel Draw image library by evenly sampling in Corel Draw categories (**Corel Draw**), 561 photos are collected from family albums (**F. Photo**), and 81 portrait photography pictures are provided by professional photographers (**Portrait**). That is, there are totally 1392 images used in our evaluation experiment. During about 5 hours testing session, each subject gave three assessments to each of the 1392 images.

The statistical results of evaluation are displayed in the following three tables by percentage averaging over all subjects and images in each of the three photo categories. Table 1 shows the evaluation results of attended view. We can see that the results marked as "GOOD" account for about 84%, while only 2% of cases are marked as "FAILED". As amateurish family photos are often poor in composition, the effectiveness of attend view adjustment in the category of **F. Photos** is more distinctive than that in other two categories.

Table 1. Attended view evaluation

	GOOD	ACCEPT	FAILED
Corel Draw	0.85	0.14	0.01
F. Photo	0.86	0.13	0.01
Portrait	0.82	0.16	0.02
Avg.	**0.84**	**0.14**	**0.02**

Table 2 lists the results of attended area detection. 67% of images are considered as "GOOD", and the "FAILED" cases made up only 6%. The failed cases are mainly caused by poor exposure of photos, either underexposure or overexposure, and weak contrasts. As the contents in **F. Photo** are often cluttered, the satisfactory of the detected attended areas in this category is the lowest. On the contrary, the best results are obtained in **Portrait** category, because the themes of portraits are very explicit. The collection in **Corel Draw** category includes a variety of photos, so the results are near to the average.

Table 2. Attended areas evaluation

	GOOD	ACCEPT	FAILED
Corel Draw	0.68	0.27	0.05
F. Photo	0.60	0.31	0.09
Portrait	0.74	0.22	0.05
Avg.	**0.67**	**0.27**	**0.06**

Although the attended points do not reflect semantics, we still evaluate them by subjective method to investigate the relationship between low-level stimulus and high level semantics. From Table 3, we can see that 47% of assessments are "GOOD" which is much lower than that of other two evaluations. Meanwhile, the proportion of "FAILED" cases is also higher than that of others, that is, 16%. This result verifies our presupposition that there is a big gap between low level features and high level semantics or human perceptions. However, 84% of cases are acceptable if both the "GOOD and "ACCEPT" cases are considered. It indicates that the attended points have some correlativity with attended regions in images, and are able to guide users to search semantics from certain reasonable starting points.

Table 3. Attended points evaluation

	GOOD	ACCEPT	FAILED
Corel Draw	0.56	0.33	0.12
F. Photo	0.44	0.41	0.15
Portrait	0.40	0.39	0.21
Avg.	**0.47**	**0.37**	**0.16**

The experiment discussed in this section provides a basic verification of correctness for the automatically detected attentions in images. Although, the applications based on the proposed approach need for further evaluations, the encouraging results presented in this paper may boost the applications of the proposed framework to content-based image retrieval, adaptive image delivery and active vision systems.

The aim of this work is to provide a feasible solution for image attention analysis, which is totally different from the other systems focusing on computer vision issues. For example, the goal and result form are all different. Moreover, most of other attention analysis algorithms are time-consuming and not suitable for personal computer implementation. Therefore, we performed a subjective evaluation instead of comparing our system with other attention analysis systems. The satisfied user study results have shown the effectiveness and practicability of the proposed framework.

6. CONCLUSIONS

This paper presents an image attention analysis framework which simulates the two human cognition processes: bottom-up and top-down. This framework outputs three-level attentions including attended view, attended areas, and attended points. The attended points are analogous to the direct perceptions of stimulus. The attended area searching simulates the early selection of human perception. The attended view may be regarded as the result of late selection of human perception. In addition, two issues are discussed in this paper. By investigating the effects of contrast in human perception, a contrast-based saliency map is proposed, which is the base of our framework. The other issue is how to search attended areas in saliency map. A *fuzzy growing* algorithm is developed to simulate early selection of human perception, by which attended areas are accurately located. User study results show that the proposed approaches are effective in accuracy and efficient in computation. The proposed framework can be easily employed or integrated into a variety of vision systems and visual content analysis related multimedia applications.

7. REFERENCES

[1] F. Jing, M. Li, H.-J. Zhang, B. Zhang, "An Effective Region-Based Image Retrieval Framework." *Proc. of* ACM Multimedia 2002, Juan Les Pins, France, December 1-6, 2002.

[2] L.Q. Chen, X. Xie, X. Fan, W.Y. Ma, H.J. Zhang, and H.Q. Zhou, "A Visual Attention Model for Adapting Images on Small Displays," ACM Multimedia Systems Journal, to appear, 2003.

[3] W. James, "The Principles of Psychology," Harvard University Press, 1890.

[4] D.E. Broadbent, "Perception and communication," Pergamon Press, Oxford, 1958.

[5] J. Deutsch, D. Deutsch, "Attetnion: Some theoretical considerations, " *Psychological Review*,70:80~90, 1963.

[6] A. Treisman, S. Gormican, "Feature analysis in early vision: evidence from search asymmetries," *Psychology Review* 95:15-48, 1988.

[7] A. Treisman, "Perception of features and objects," in: *Visual Attention*, Oxford University Press, New York, 1998.

[8] C. Koch, S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, Vol.4, pp. 219-227, 1985.

[9] F. Crik, C. Koch, "Some reflections on visual awareness," in: *Proceedings of the Cold Spring Harbor Symposia on Quantitative Biology*, Volume LV, Cold Spring Harbor Laboratory Press, 1990.

[10] J. K. Tsotsos, S. M. Culhane, W.Y.K. Wai, et al, "Modeling visual attention via selective tuning," *Artificial Intelligence*, 78:507-545, 1995.

[11] J.M. Wolfe and K.R. Cave, "Deploying visual attention: The guided search model," In A. Blake and T. Troscianko, editors, *AI and the Eye*, chapter 4, page 79-103. John Wiley & Sons Ltd., 1990.

[12] E. Niebur, C. Koch, "Computational architectures for attention," R. Parasuraman, (Ed.), *The attentive brain*, Cambridge, MA: MIT Press, pp.163-186, 1998.

[13] R. Milanese, S. Gil and T. Pun, "attentive Mechanism for dynamic and static scene analysis," *Optical Engineering*, Vol.34, no.8, pp2428-2434, Aug. 1995.

[14] S. Baluja and D.A. Pomerleau, "Expectation-based selective attention dor visual monitoring and control of a robot vehicle," *Robotics and Autonomous System*, vol.22, no.3-4, pp.329-344, Dec.1997.

[15] L. Itti, C. Koch, E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. on* Pattern Analysis and Machine Intelligence, 1998.

[16] S. Ahmad, "VISIT: A neural model of covert attention," *Advances in Neural Information Processing Systems*, Vol.4, p.420-427, San Mateo, CA: Morgan Kaufmann, 1991.

[17] Y.F. Ma, H.J. Zhang, "A Model of Motion Attention for Video Skimming," *Proc. of* International Conference on Image Processing 2002, Rochester, N.Y., U.S., 2002.

[18] Y.F. Ma, L. Lu, H.J. Zhang, M.J. Li, "A User Attention Model for Video Summarization," *Proc. of* ACM Multimedia 2002, Juan Les Pins, France, Dec. 1-6, 2002.

[19] Y. Li, Y.F. Ma, H.J. Zhang, "Salient Region Detection and Tracking in Video," *Proc. of* IEEE International Conference on Multimedia & Expo 2003, Baltimore, Maryland, U.S, July 6-9, 2003.

[20] Lamming D. (1991) Contrast Sensitivity. Chapter 5. In: Cronly-Dillon, J., Vision and Visual Dysfunction, Vol 5. London: Macmillan Press.

[21] G.J. Klir, B. Yuan, "Fuzzy Sets and Fuzzy Logic: Theory and Applications," Published by Prentice Hall, Upper Saddle River, NJ, 1995.

[22] L.A. Zadeh, "Probability measures of fuzzy events," J. Math. Anal. Appl. 23, 421-427, 1968.

[23] P.K. Sahoo, D.W. Slaaf, T.A. Albert, "Threshold selection using a minimal histogram entropy difference," *Optical Engineering* 36(7), 1976-1981.

[24] S. Z. Li, et al., "Statistical Learning of Multi-View Face Detection," *Proc. of* European Conference on Computer Vision 2002, Copenhagen, Denmark, May, 2002.

[25] Y. Deng, C. Kenney, *et al.*, "Peer group filtering and perceptual color image quantization," *Proc. of* IEEE International Symposium on Circuits and Systems, Vol.4, p.21-24, 1999.

[26] L. Zusne, "Contemporary theory of visual form perception: III," *The global Theories*, chapter 4, p108-174. Academic Press, 1970.

[27] K. Koffka, "Principles of Gestalt psychology," Harcourt Brace Jovanovic, 1935.