A Perceptual Object Based Attention Mechanism for Scene Analysis

ZHAO Xun-po, WANG Lu, HU Zhan-yi

(*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100080)

Abstract The object-based selective attention has been given increasingly importance in psychology domain in recent years, whereas most of the existing models of selective attention in computer vision field are either feature-based or space-based. In this paper, an object-based computational model of selective attention is proposed where "perceptual object" is postulated as the basic attention unit. The proposed attention model consists of following two steps: (1) how to select the first focus of attention in a given image; (2) how to shift the attention within the whole image. Under this model, the contrast, defined as the absolute gray difference between the "perceptual object" and its neighborhood, is used as the measure of the object's salience, by which the attention is determined and shifted. The advantages of this model lie in: Firstly, the model is entirely based on perceptual objects, its results can be easily applied to object detection, image segmentation, and scene analysis. Secondly, the model is of multi-scale, flexible enough to be easily adjusted according to the specific applications. Extensive experiments on real images show that the proposed attention model works reasonably.

Keywords selective attention, regions-of-interest, perceptual object, shift of attention

中图法分类号: TP391.41 **文献标识码**: A **文章编号**: 1006-8961(2006)02-0281-08

一种基于感知物体的场景分析注意机制

赵训坡 王 璐 胡占义

(中国科学院自动化研究所模式识别国家重点实验室,北京 100080)

摘 要 基于物体的选择性注意在心理学领域正日益为广大研究人员所认可,而计算机视觉领域中现有的注意模型大多数是基于特征的,或者是基于空间的。本文给出了一种基于物体的选择性注意计算模型。该模型将"感知物体"作为引起注意的基本单元,并给出了感知物体及其邻域的定义。该注意模型包括两个步骤:(1)在给定图像中选择第一个注视点;(2)在整幅图像中实现注视点的有效转移。在该注意模型中,感知物体与其邻域之间灰度值的绝对差异——对比度,被作为该感知物体显著性的一种度量,并且注视点在图像中的转移顺序是由每个感知物体的显著度的次序来决定的。该模型的优点有:首先,由于该模型是完全基于感知物体的,使得其输出结果可以很容易地应用到物体识别、图像分割和场景分析中;其次,该模型是多尺度的,也就是说,它可以根据实际任务的需要进行适当的调整。大量的真实图像实验表明,所提出的模型具有一定的合理性。

关键词 选择性注意 感兴趣区域 感知物体 注意转移

1 Introduction

Attention is very important to everyone. The environment around us is rich and colorful. The amount of information, which the real and dynamic environment can provide, is very huge. Due to computational limit, we can select only a small fraction

基金项目:国家自然科学基金项目(60375006)

收稿日期:2004-11-15;**改回日期**:2005-04-20

第一作者简介:赵训坡(1974 ~),男。2005 年于中国科学院自动化研究所获模式识别与智能系统专业博士学位。主要研究方向为数字图像处理、图像配准、计算机主动视觉。E-mail: xpzhao@nlpr.ia.ac.cn

of all information for further processing and ignore the others. How can we do that? This all depends on the mechanism of selective attention in our vision system. Without attention, the so-called general-purpose vision is not possible.

In recent years, especially with the emerging interests of active vision, the researchers on computer vision have been increasingly concerned with attention mechanisms. Many computational models or methods of selective attention have been proposed and most of them are either feature-based or space-based. The theory of feature-based selective attention as Allport argued in[1] regards the features of an object as the units to be selected and believes that vision system is limited to the number of distinguished features. For example, a widely used attention operator based on gray level variance proposed by Moravec in[2] is typically a kind of feature-based mechanism. The same is for the method of Lamdan and Yeshurun shown in [3],[4], in which the points of high curvature in the edge map are regarded as the interesting points and the method of Danial proposed in[5] where the feature of symmetry decides the fixation of interests.

The view of space-based attention is relatively popular. It advocates that space is the medium of visual attention. The spotlight metaphor proposed by Posner in[6] is commonly used for space-based attention. In this idea, attention is just like a beam of light and the sub-region lightened by such a mental light is what we attend to at a certain time. Treisman's theory[7] of feature integration is also a kind of space-based theory in which visual features, such as color, orientation and so on, are combined together through space location. Itti in his work[8,9] sets up an influential method based on saliency-map. He computes saliency value for every pixels of the image and then chooses the pixel with the maximal saliency value as the center of the focus of attention while the radius of the selected region is fixed. This method can be regarded as a space-based one. The size of the attended region in another representative model proposed by Tsotsos[10] is greatly influenced by the preset parameters.

However, more and more psychologists believe that attention is object-based. Many experiments of behavior and neurological psychology studies have made clear that "object" plays an important role in visual attention[11,12]. What visual attention selects are perceptual objects derived through perceptual organization. In addition, many famous experiments originally supporting space-based theory are now found more accordant with object-based theory[13,14]. The advantage of the object-based attention model to the space-based mechanism is that the selected region in the object-based model can be adjusted according to the change of the size and the shape of the object, while the region attended in the space-base model can only be adjusted according to the change of the size of the object at most. The problem associated with object-based attention is that it is hard to give a rigorous definition of the object.

In this paper, we attempt to construct an object-based attention model in which the attention units are "perceptual objects".

2 Selective attention model

Our attention model consists of the following two parts: (1) how to select the first focus of attention in a given image; (2) how to shift the attention within the whole image. Before elaborating on the two parts, we will first give some explanations on "perceptual object" and its neighborhood.

2.1 Perceptual object

One point we have to emphasize is that the "object" here refers to a perceptual object rather than a real object or a natural object, such as a person, a cat, or a house. What we are studying here is low-level visual process and specifically data-driven or bottom-up visual attention. Without high-level knowledge or top-down information, it is impossible to put those regions that are perceptually heterogeneous together to compose a real object. For example, at early stage, it is infeasible to group the regions of a red roof and a white wall into a whole region representing a house. It is more likely that we focus on the red region

and the white region separately. After using high-level knowledge to find the relationship between them, we can group them together to form the concept of a house. In this example the red roof and the white wall are perceptual objects whereas the house belongs to a real object. We can see that a real object may consist of multiple perceptual objects.

Although it is hard to give a rigorous definition of a perceptual object, we can get some clues from our common sense. A perceptual object is spatially connected and homogenous in color and intensity and has high contrast compared with its surrounding. Hence a perceptual object should have the following defining characteristics: (1) spatially connected and with a certain size; (2) endogenous feature I—homogeneity in intensity and colors inside the object; (3) exogenous feature E—contrast between the object and its surrounding.

Along these lines, an evaluation function is introduced to assess the potentiality of a given region to be a perceptual object, and its output is defined as the saliency of the region. Here is a definition of a given region's neighborhood: As in fig. 1, a layered neighborhood is introduced. The first layer is composed of the pixels in the 4-neighborhood of the border pixels of the given region, and the second layer is composed of the pixels in the 4-neighborhood of the first layer pixels and so on. For a solid region, its neighborhood is just its outer layers, whereas the neighborhood of a non-solid region includes both its outer and inner layers.

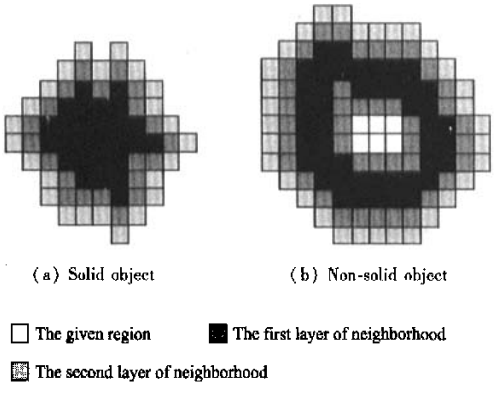

Fig. 1　The definition of the neighborhood of a give region

Suppose a region is composed of N_0 pixels, and the i^{th} layer of its neighborhood has m_i pixels, if:

$$\sum_{i=1}^{k-1} m_i < N_0 \leqslant \sum_{i=1}^{k} m_i \tag{1}$$

then the neighborhood of the region has k layers. In this way, the size of the neighborhood is approximately equal to the size of the given region in terms of pixel components.

Once the neighborhood of a given region is defined, we can calculate the region's saliency S. Assume that the given image consists of N pixels and a region to be evaluated has N_0 pixels with a k—layer neighborhood. We define $\boldsymbol{R}_0$ as the region to be evaluated, $\boldsymbol{R}_N$ the neighborhood of $\boldsymbol{R}_0$, and $\boldsymbol{R}_W$ the rest region of the given image outside $\boldsymbol{R}_0$, as shown in fig. 2.

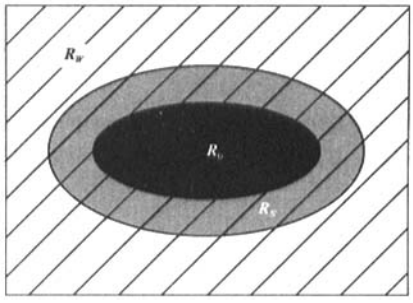

Fig. 2　The saliency computation of a given region

The saliency S is defined as

$$S = E - I \tag{2}$$

where E represents the exogenous feature and I represents the endogenous feature of the given region $\boldsymbol{R}_0$. E can be further divided into two parts. The part C_W is the contrast between $\boldsymbol{R}_0$ and $\boldsymbol{R}_W$ as:

$$C_W = \frac{1}{N - N_0} \sum_{i=1}^{N-N_0} |g_i - \bar{g}|, (g_i \in \boldsymbol{R}_W) \tag{3}$$

$$g = \frac{1}{N_0} \sum_{i=1}^{N_0} g_i, \ (g_i \in \boldsymbol{R}_0) \tag{4}$$

where g_i is the gray value of the i^{th} pixel. The other part C_N is the contrast between $\boldsymbol{R}_0$ and its neighborhood $\boldsymbol{R}_N$,

$$C_N = \sum_{i=1}^{k} f_{ni} C_{ni} \tag{5}$$

where C_{ni} denotes the contrast between the region $\boldsymbol{R}_0$ and the i^{th} layer $\boldsymbol{R}_{Ni}$ of its neighborhood:

$$C_{ni} = \frac{1}{m_i} \sum_{j=1}^{m_i} |g_j - \bar{g}|, \ (g_i \in \boldsymbol{R}_{Ni}) \tag{6}$$

where m_i denotes the number of pixels of the i^{th} layer $\boldsymbol{R}_{Ni}$ of its neighborhood. The research on neurophysiology discovers that the response curve of the receptive field of visual cortex to optical signals is an approximate difference of two Gaussian functions. From this, we assume each neighborhood layer has different contribution to C_N. The nearer is the neighborhood layer to the given region, the more contributions it has. Therefore we give weight f_{ni} to scale the i^{th} neighborhood layer of $\boldsymbol{R}_0$. It is defined as:

$$f_{ni} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{i^2}{2\sigma^2}}, \ (\sigma = k/3) \tag{7}$$

where k is the number of neighborhood layers. Combining the two parts, the exogenous feature E of the given region $\boldsymbol{R}_0$ is computed as:

$$E = f_W C_W + f_N C_N \tag{8}$$

where f_W and f_N are two weights: $f_W = N_0/N$ and $f_N = 1 - f_W$.

The endogenous feature I is computed as $I = f_H C_H$, where C_H denotes the intensity homogeneity of the region $\boldsymbol{R}_0$,

$$C_H = \frac{1}{N_0} \sum_{i=1}^{N_0} |g_i - \bar{g}|, \ (g_i \in \boldsymbol{R}_0) \tag{9}$$

and f_H is its weight determined by the dynamic range of gray values of the evaluated region $\boldsymbol{R}_0$,

$$f_H = \frac{1}{\sqrt{g_{max} - g_{min}} + 1} \tag{10}$$

$$g_{max} = \max[g_1, \cdots, g_{N_0}] \tag{11}$$

$$g_{min} = \min[g_1, \cdots, g_{N_0}] \tag{12}$$

Based on the above equations, the saliency of a given region can be computed, which is used as a gauge to decide whether the region is a perceptual object.

2.2 The first focus of attention in a given image

The foundation of our model is the assumption that the greater is the saliency value S of a region, the more likely the region is to be a perceptual object and attracts our attention. Then our first focus of attention should be the region with the highest saliency value all over the given image. Without any constrains in size and shape, to find out the connected region with the maximum S all over the given image is an NP-complete problem. Due to computational concerns, we turn to find out a sub-optimal algorithm, which could be computationally efficient and capable of locating the region with quasi-maximum S.

Our idea is similar to the method presented by Lindeberg[15]. The difference lies in the exact way to select regions and the algorithm to evaluate their saliency. Our method can be divided into two steps (1) Selecting the regions which are potential perceptual objects in the given image; (2) Evaluating the saliency S of each of these potential regions and selecting the region with the largest saliency value as our first focus of attention.

From common sense, we know that the interested perceptual object is always a small fraction of a given image. Hence the pixel number of a possible perceptual object should be no more than the half of the total pixels of the given image. Meanwhile, considering that too small region generally cannot arouse our attention either, we limit the minimum number of the selected region to be T_A, which is preset in our model. In short, all selected regions satisfy the following two rules: (1) spatially connected; (2) $T_A \leqslant N_0 < N/2$, where N is the number of the total pixels of the given image and N_0 is the number of pixels of a selected region.

The procedure of selecting the first focus of attention can be summarized as:

(1) For a given image, compute $G_{min} = \min[g_1, \cdots, g_N]$, $G_{max} = \max[g_1, \cdots, g_N]$, $g_i \in \boldsymbol{R}_0$;

(2) For all combinations of G_i and G_j, with $G_i < G_j$ and $G_{min} \leqslant G_i$, $G_j \leqslant G_{max}$, use G_i as the low-threshold and G_j as the high-threshold to segment the given image. The connected region with its gray values within $[G_i, G_j]$ and its total number of pixels N_0 satisfying $T_A \leqslant N_0 \leqslant N/2$ will be recorded.

(3) Select the region with the largest saliency value in all the recorded regions in step 2 as our first focus of attention for the whole given image.

Remark: In practice, in order to speed up the computation, rather than enumeate all possible combination of G_i and G_j, we only select those combinations with $G_i = G_{min} + iT_g$ and $G_j = G_{min} + jT_g$, $i < j$. The integer $T_g > 1$ is used as a tradeoff between the goodness of selection and the computational load.

In all our experiments, we choose $T_g = 10$. In addition, based on our experiments, the choice of T_g seems not too sensitive so long as it is not too large.

2.3 Attention shift

The first attended region (perceptual object) is the one with the highest saliency value among all the regions selected in the given image. Then how can our attention be shifted to the next one? Taking into account the principle of IOR (inhibition of return), described by Klein in [16], [17], our attention shift procedure is outlined as follows:

(1) Set a threshold T_S which is the minimum saliency value for a region to arouse our attention.

(2) Take the given image as the first input region F_0 and initialize the iterative parameter $i = 0$.

(3) In the current input region F_i, select the sub-region whose salience value is the highest among all sub-regions. The two cases appear:

(a) If the salience value of a sub-region is higher than T_S, then the sub-region will be our next focus of attention. We take it as the next input region F_{i+1} and record the rest sub-region of F_i outside F_{i+1} as B_{i+1}. Then set $i = i + 1$ and return to step 3.

(b) In the case that there is no sub-region of F_i having saliency value higher than T_S, we will leave the region of F_i and return to the region of B_{i-1} as the next input region. Let $F_i = B_{i-1}$ and $i = i - 1$. If $i = 0$, go to step 4; otherwise return to step 3.

(4) In the case $i = 0$, which means there is no region of the unvisited area of the input image having the saliency value higher than T_S, the whole process is ended.

Taking an example as shown in fig. 3, the first attended region is the black rectangular area in which lie a small rectangle and a circle. Then this area is considered as the input region, and the small rectangle and the circle are successively attended. The fourth attended region is the elliptical area, whereas the fifth attended region is the light gray rectangle in the ellipse. Finally, we obtain the sixth region to attend, which is the light gray circle. The attended order is: 0→1→2→3→4→5→6.

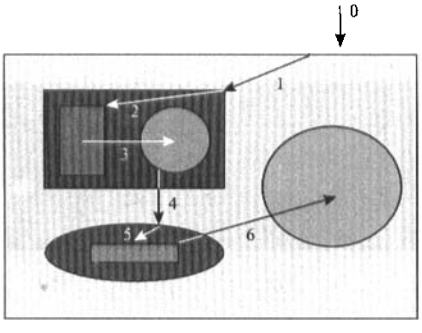

Fig. 3　One simulated example of our model

3 Experiments

We tested our attention model with numerous real images. In all of our experiments we do not consider any high-level guidance and set the thresholds $T_A = 200$ and $T_S = 40$. In each of the examples from fig. 4 to fig. 7, the first picture is the original input gray image and in the following pictures the white region is the current selected focus of attention, what is called "perceptual object" in this paper, while the black is the region which will no longer be further considered, and the gray part is the region to be considered later. At the same time, b, c, d and so on denote the attention shift in the original image, where b1, b2 denote two attention shifts in the white region of b and so on. The order of the pictures is the order of attention shift according to the salience value of perceptual objects. From these results, we can see that our

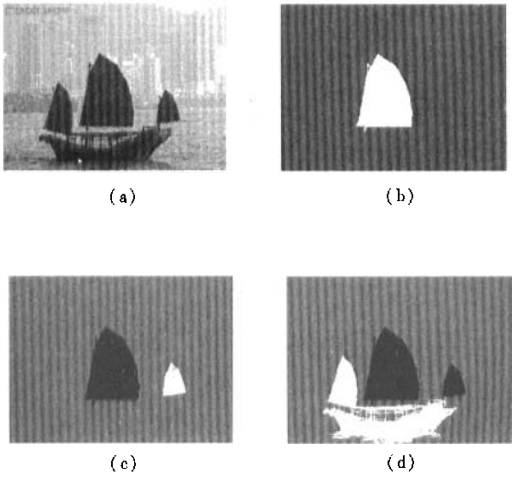

Fig. 4　The first example of real images

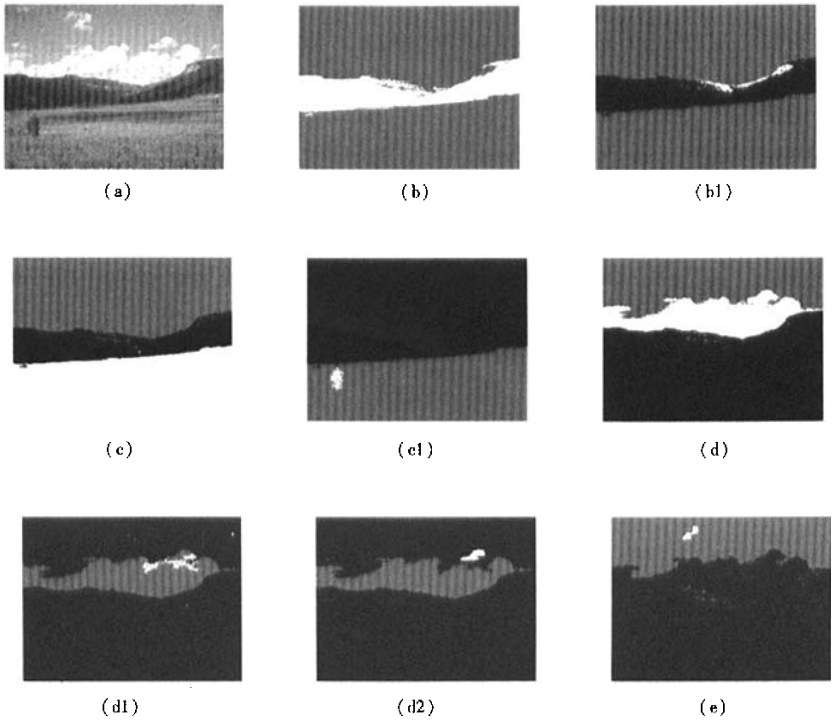

Fig. 5 The second example of real images

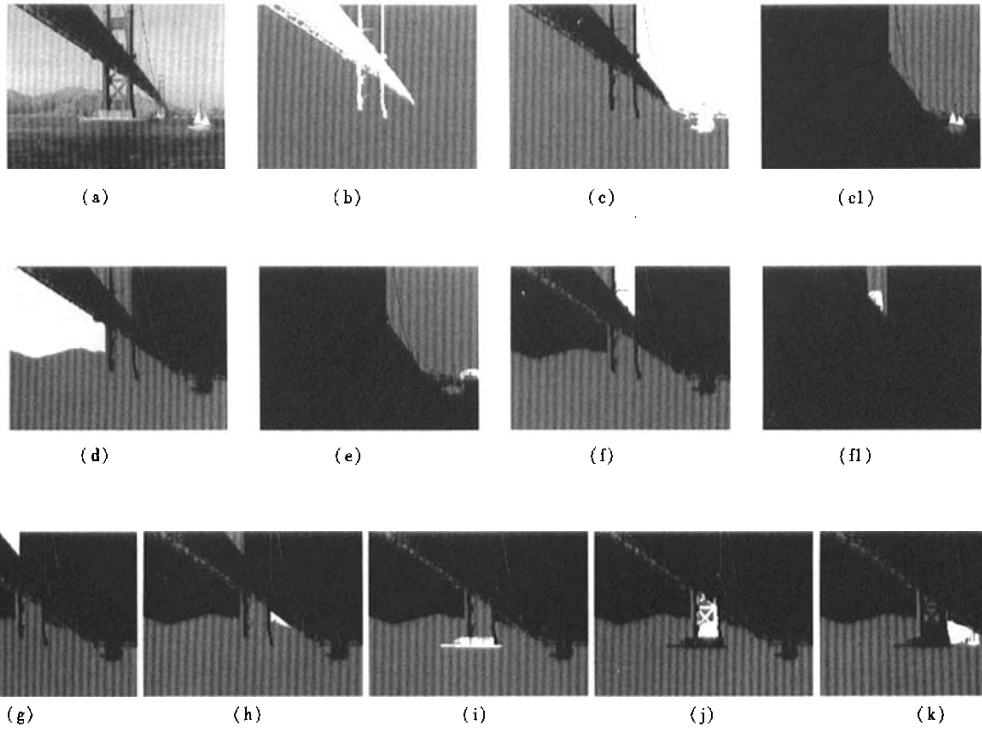

Fig. 6 The third example of real images

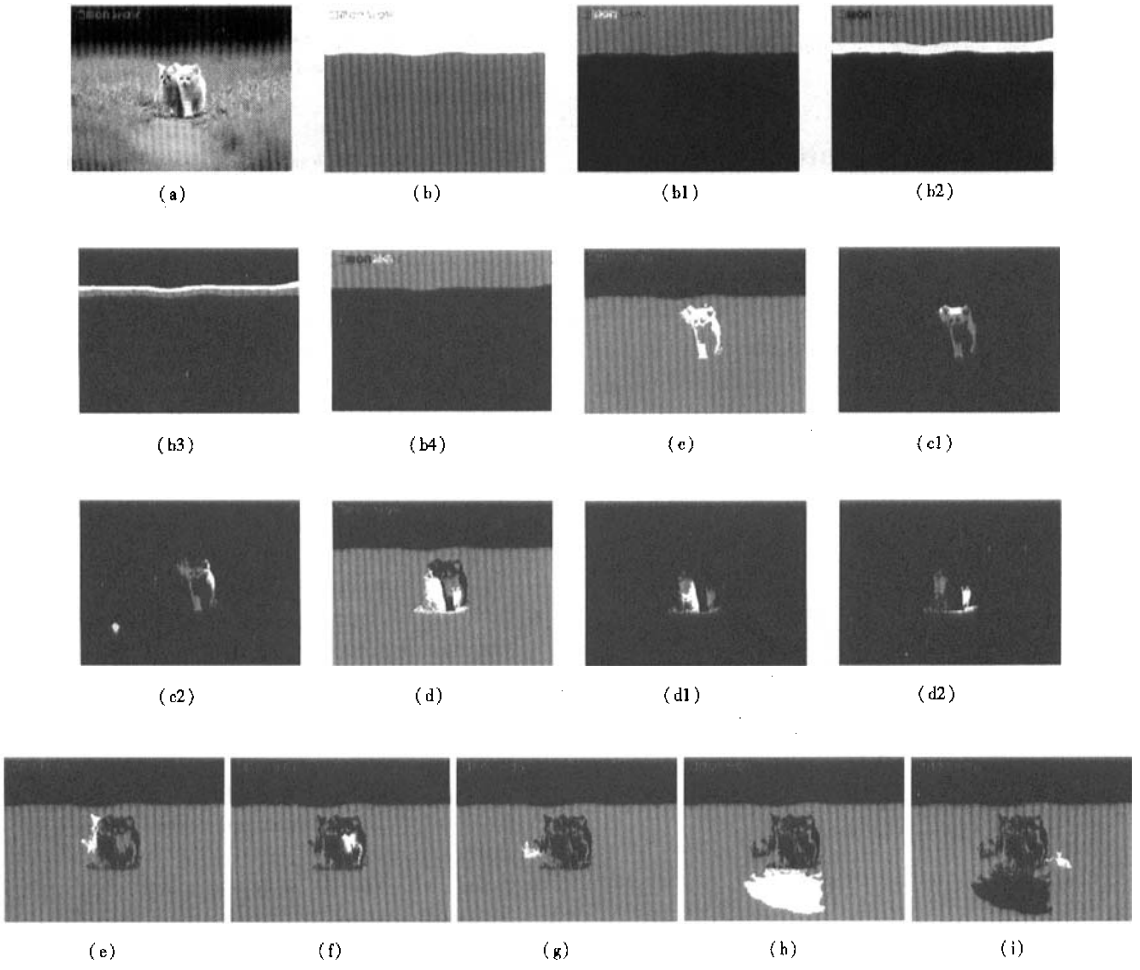

Fig. 7　The fourth example of real images

proposed model agrees with human visual instinct to a large extent. Because no high-level guidance is involved in our model, sometimes the shift order of attention is not consistent with the human vision.

4 Conclusions

In this paper, we have proposed a perceptual-object based visual attention model for selecting regions-of-interest in an image. Under this model, at first, a threshold-like technique is used to segment regions. A region potentially to be a perceptual object must be a connected one with certain size. After that, based on a new saliency operator which takes into account the internal density homogeneity, the contrast to its neighbor, as well as the contrast to the whole given image, the saliency value of each region is evaluated. The region with the highest saliency value is selected as the first focus of attention. Then a reasonable mechanism of attention shift is introduced, where the principle of " inhibition of return " is embedded, which is hierarchical in nature.

The advantages of our mechanism lie in two aspects. The first one is its advantage in theory. Our model is object-based. This coincides with great amount of psychological experiments. Moreover, the perceptual objects obtained in our model can be further used in image segmentation, object recognition and image compression. The second aspect is that our method is of multi-scale. By adjusting the thresholds,

T_A and T_S, we can get into different detailed levels of analysis. The larger are the thresholds, the coarser is the analysis.

5 Limitations and future work

There are two main limitations in our mothod. The first one is that the detection of some perceptual objects is still beyond the ability of our method. For example, in Fig. 8, our method cannot group the pixels into four parts for attention. This is because we consider only the intensity information into our model but not the texture. How to include texture information in our model is one of our future works.

Fig. 8 An example of beyond the ability of our method

The second one is that the border information of a region is not sufficiently considered in our model. Because of this, some regions are selected but they are obviously not perceptual objects judged by humans. In our future work, the border information will be enhanced when a region is selected.

References(参考文献)

1 Allport A. Visual attention [A]. In M. I. Posner, (Ed.), Foundations of Cognitive Science[M]. Cambridge, MA: MIT Press, 1989: 631 ~ 682.

2 Moravec H P. Towards automatic visual obstacle avoidance[A]. In: Proceedings of International Joint Conference on Artificial Intelligence [C], Cambridge, MA, USA, 1977: 584 ~ 590.

3 Lamdan Y, Schwartz J T, Wolfson H. On recognition of 3-d objects from 2-d images [A]. In : Proceedings of IEEE international conference on Robotics and Automation [C], Los Alamitos, CA, 1988: 1407 ~ 1413.

4 Yeshurun Y, Schwarts E L. Shape description with a space-variant sensor: Algorithm for scan-path, fusion, and convergence over multiple scans [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989, **11**(11): 1217 ~ 1222.

5 Danial R, Haim W, Yehezkel Y. Context free attentional operators: the generalized symmetry transform [J]. International Journal of Computer Vision, 1995, **14**(2): 119 ~ 130.

6 Posner M I, Snyder C R, Davidson B J. Attention and the detection of signals[J]. Journal of Experimental Psychology: General, 1980, **109**(2): 160 ~ 174.

7 Treisman A M, Gelade G. A feature-integration theory of attention [J]. Cognitive Psychology, 1980, **14**(1): 107 ~ 141.

8 Itti L, Koch C, Niebur E. A Model of saliency-based visual attention for rapid scene analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, **20**(1): 1254 ~ 1259.

9 Itti L, Koch C. Computational modeling of visual attention [J]. Nature Reviews Neuroscience, 2001, **2**(3): 194 ~ 230.

10 Tsotsos John K, Culhance Sean M, Wai Winky Yan Kei, *et al*. Modeling visual attention via selective tuning [J]. Artificial Intelligence, 1995, **78**(1): 507 ~ 545.

11 Duncan J. Selective attention and the organization of visual information[J]. Journal of Experimental Psychology: General, 1984, **113**(4): 501 ~ 517.

12 Treisman A M. Perceptual grouping and attention in visual search for features and for objects [J]. Journal of Experimental Psychology, 1982, **37A**(4): 571 ~ 590.

13 Kramer A F, Jacobson A. Perceptual organization and focused attention: The role of objects and proximity in visual attention [J]. Perception & Psychophysics, 1991, **50**(3): 267 ~ 284.

14 Gibson B. Visual attention and objects: One versus two or convex versus concave? [J]. Journal of Experimental Psychology: Human Perception and Performance, 1994, **20**(1): 203 ~ 207.

15 Lindeberg T. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method for focus-of-attention [J]. International Journal of Computer Vision, 1993, **11**(3): 283 ~ 318.

16 Klein R M. Inhibitory tagging system facilitates visual search [J]. Nature, 1988, **334**(4): 430 ~ 431.

17 Klein R M, Macinnes W J. Inhibition of return is a foraging facilitator in visual search [J]. Psychological Science, 1999, **10** (4): 346 ~ 352.