A Selective Attention Based Method for Target Detection

KYUNGJOO CHEOI, YILLBYUNG LEE
Department of Computer Science
Yonsei University
134 Sinchon-dong, Seodaemun-gu, Seoul 120-749
KOREA

Abstract: - This paper describes a new selective attention based method which can be used effectively in wide range of target detection tasks without using any a-priori knowledge about the target. In our proposed system, several basic features are extracted directly from original input visual stimuli, and these features are integrated based on their local competitive relations and statistical information. Through integration process, unnecessary features for detecting the target are spontaneously decreased, while useful features are enhanced. The performance of our system was evaluated over some simple synthetic images which were generated by computer, and complex real images of natural environment taken from different domains which also include severe amounts of noise.

Key-Words: - Selective Attention, Target Detection, Feature Extraction, Feature Integration, local competition

1 Introduction

In most of the classical methods in image analysis, high-level features of the whole size of image are computed first, and then each of the local set of features are compared with the collection of stored prototypes, in a serial manner. Those methods are extensively slow and too application specific [5]. By the way, biological system appear to employ a serial strategy by which an attentional spotlight rapidly selects regions of interest in the scene, rather than process whole visual scenes [1]. This function is called visual attention, and it enables the Primates to interpret natural scenes in real-time, despite of the limited speed of the biological architecture available for such tasks.

Although many effective and fast computer vision systems for target detection exist, most of them were finely tuned to specific targets. Thus they typically fail in detecting other types of objects except already defined target to which the systems were finely tuned. So, it is very difficult to extend this kind of previous systems to some other types of target detection tasks. Besides, most of them also do not detect the target when severe amouts of noise are exist in input image.

Because of these reasons, we now propose a target(or 'salient object') detection system which can be used for general purpose, not for specific type of the target, and which successfully can be worked with not only normal images which are not courrputed with noise but also with the images which are corrupted with heavy amounts of noise. Our system is based on the bottom-up processing of human visual attention, and we will show that our system is very useful in target detection tasks.

Before explaining our system, we will explain the paradigms which describe human focus of visual attention briefly in Section 2 with some previous biological plausible systems for target detection in computer vision. Then, our proposed methodology will be explained in Section 3, and experimental results will be shown in Section 4. Finally, discussion and conclusion will be made in Section 5.

2 Visual Attention and Previous Works

In engineering, human visual attention mechanism can be very efficiently applied to the problems of background /foreground separation, object recognition, and so forth. However, it is not utilized enough yet. A fundamental problem to be solved is that how to make the system focus its attention to regions of interest. What features are important for focusing the attention?

A number of paradigms which describe human focus of visual attention have been developed over the years by researchers in psychology [2,4,6,7,10,13]. Among them, most literature agrees that the attention selection mechanism consists of two functionally hierarchical stages, pre-attentive and attentive stage. In early pre-attentive stage, all visual stimuli in the

entire visual field are processed in parallel without capacity limitation. And in attentive limited capacity stage, only one item or at best a few items are processed at a time. One of the most popular models on visual attention is the one proposed by Treisman [14]. He suggested a *feature integration theory*, that views the perception of objects on the basis of above two-stage metaphor as a process. It says that the basic 'features' like edges, orientation, width, size, color, brightness, etc., are detected in the pre-attentive stage, and these basic features are 'integrated' in the attentive stage in order to be perceived as objects in the world.

Some biologically plausible systems for target detection in computer vision have been proposed [8,9,11,12,13]. As the systems in [12] and [13] were applied only to synthetic images or other simple images containing characters, it is very hard to extend the systems to such applications which use complex natural color images as input. And in [11], even if the researcher tried to evaluate their system with real images, they presented rare experimental results of natural color images still less noisy images. Also, the performance of the systems in [8] and [9] need more systematic analysis with more noisy images.

Our system proposed here finds the target only by the properties of input visual stimuli. There is no doubt that high level features can be very useful in detecting the target in situations where a template of a target is known a priori. And in such cases, the system can be modeled with high accuracy. But, even if this high level knowledge should eventually improve its performance, the system must be able to work even without it. Because, in the general case little is known about the contents of the scene, and such high-level information cannot be used. In our approach, we used only bottom-up component of visual attention, so we can easily extend the system to various applications without major changes of the architecture. Moreover, as the system does not require any high-level information, it can be used for general purpose. We have tested our system on a wide variety of images, ranging from simple synthetic images to natural outdoor scenes, including noisy images. The performance of the system proved very robust to these kinds of various images.

3 The Methodology

The detailed view of our system is shown in Fig.1.

In our system, input image is processed through two steps : feature extraction, feature integration. In the first step, very simple early visual feature maps such as intensity contrast and color contrast are first extracted directly from visual stimuli, in parallel(EF^1 for intensity contrast, EF^2 for red/blue opponency, and EF^3 for green/yellow opponency). These independent three feature maps are then reorganized into F^1, F^2, and F^3, respectively by oriented ON-center and OFF-surround operator. Each map has one more feature, orientation, than early visual feature maps, and also has enhanced pixel values which are largely different from their surroundings'. Thereafter, these reorganized feature maps are propagated to the next step and integrated into a saliency map by statistical information and competitive relations of the pixels in each reorganized feature map. Through integration process, unnecessary features for detecting the target are spontaneously decreased while useful features are enhanced. And the system selects the most different features among other features as a target.

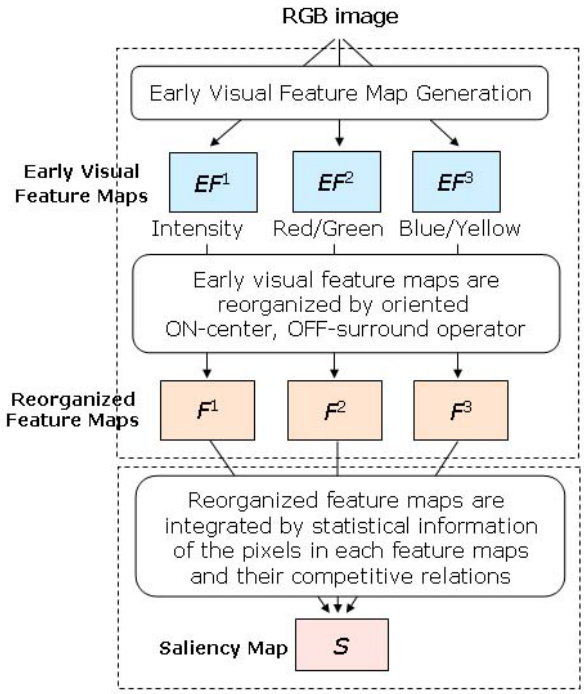


Fig.1 Detailed view of the system

3.1 Feature Extraction

In feature extraction step, an input image is initially decomposed into three independent feature maps. Each map represents the value of a certain attribute computed on a set of low-level features. Here, one achromatic feature map(F^1)and two chromatic feature

maps(F^2,F^3) are generated.

Before generating three feature maps, early visual feature maps(EF^1,EF^2,EF^3) are generated first, and then these maps reorganized into three complete feature maps. EF^1 is generated by intensity information of an input image. The red, green and blue component of an original input image are first extracted as R, G, and B, respectively, and then an intensity image I is obtained as $(R+G+B)/3$. Intensity image I becomes a first early visual feature map EF^1. The next two early visual feature maps EF^2 and EF^3 are modeled with the two types of color opponency exhibited by the cells with homogeneous type of receptive fields in visual cortex which respond very strong to color contrast. To generate EF^2 and EF^3, broadly tuned color channels are extracted as r, g, b, and y by $r=R-(G+B)/2$, $g=G-(R+B)/2$, $b=B-(R+G)/2$, $y=R+G-2(|R-G|+2)$, respectively, first. Each of which indicates red, green, blue, and yellow channel respectively, and yields maximal response for pure hue to which it is tuned [8]. Next, EF^2 is generated to account for red/green color opponency by $EF^2=r-g$, and EF^3 for blue/yellow color opponency by $EF^3=b-y$.

All generated three independent feature maps are then processed by Eq. 1 in order to extract orientations from each feature maps and also to enhance the regions of pixels whose values are largely different from their surroundings'.

$$F_{x,y}^{k} = \sum_{\theta}\left(\sum_{m,n} EF_{m,n}^{k} \cdot h_{x-m,y-n}(\theta) \right)^2 \qquad (1)$$

To generate F^k in Eq. 1, three early visual feature maps EF^k (k=1,2,3) are first normalized in the range [0,1] in order to eliminate across-modality differences due to dissimilar feature extraction mechanisms. Then normalized maps are convolved with the bank of $h(\theta)$ filter at 8 orientations($\theta \in \{0, \pi/8, 2\pi/8, \cdots, 7\pi/8\}$), which were generated by

$$h_{x,y}(\theta) = \left| \begin{array}{l} K_1 \cdot G_{x,y}(\sigma, r_1 \cdot \sigma, \theta) \\ \qquad - K_2 \cdot G_{x,y}(r_2 \cdot \sigma, r_1 \cdot r_2 \cdot \sigma, \theta) \end{array} \right| \qquad (2)$$

where two $G_{x,y}(\cdot,\cdot,\cdot)$ denote 2-D oriented Gaussian function, K_1 and K_2 denote positive constant, r_1 denote the eccentricities of the two Gaussians, and r_2 denote the ratio between the widths of the ON and OFF Gaussians [11]. After convolution, the results are squared to enhance the contrast, and finally we take the summation of the results to eliminate the orientation parameter θ.

3.2 Integration Process

Although many features which influence visual attention have been identified, little quantitative data exist regarding the exact weighting of the different features and their relationship. Some features are very important, but it is very difficult to define exactly that how much the one feature is more important than another [15]. A particular feature may be more important than another in one image, while in another image the opposite may be true. Therefore, multiple feature maps have to be integrated into a unique saliency map in order to obtain a single representation. The idea of *saliency map* which is an explicit two-dimensional map that encodes the saliency of objects in the visual environment, was introduced by Koch and Ullman [10] to accomplish pre-attentive selection [8].

Since the feature maps were derived from different visual modalities, combining multiple feature maps is not an easy work. Previously, this would be done by weighted sum of all information in the map [2]. However, in this case the performance of the system may highly rely on the appropriate choice of the weights. So, we suggest a simple integration process which promotes those maps in which a small number of meaningful high activity areas are present while suppressing the others. This process is composed of following three steps.

First, each computed feature map is convolved with the large size of the LoG filter and the result is added with the original input one by

$$\hat{F}_{x,y}^{k} = \sum_{m,n}\left(F_{m,n}^{k} \cdot LoG_{x-m,y-n} \right) + F_{x,y}^{k} \qquad (3)$$

where

$$LoG_{x,y} = \left[\frac{x^2 + y^2 - 2\sigma^2}{2\pi\sigma^6} \right] \cdot e^{-\left(\frac{x^2+y^2}{2\sigma^2} \right)} \qquad (4)$$

This operation is iterated 4 times, and it causes the effect of short-range cooperation and long-range competition among neighboring values of the map, and also reduces the noise.

Second, the processed map is processed by

$$S_{x,y}^{k} = \frac{\tilde{F}_{x,y}^{k} - MIN}{MAX - MIN} \qquad (5)$$

where

$$\begin{aligned}
\tilde{F}^k_{x,y} &= \hat{F}^k_{x,y} \times (MAX^k - AVE^k)^2, \\
MAX &= \max(\tilde{F}^1_{x,y}, \tilde{F}^2_{x,y}, \tilde{F}^3_{x,y}), \\
MIN &= \min(\tilde{F}^1_{x,y}, \tilde{F}^2_{x,y}, \tilde{F}^3_{x,y}), \\
MAX^k &= \max(\hat{F}^k_{x,y}), \\
AVE^k &= \text{average}(\hat{F}^k_{x,y})
\end{aligned} \tag{6}$$

This operation enhances the values associated with strong peak activities in the map while suppressing uniform peak activities, by the statistical information of the pixels in the map. Comparing the maximum value to the average value over all pixels in each map enables us to know how different the most activation location is from the average. We can use this information to promote those in which a small number of meaningful high activity areas are present while suppressing the others. And by comparing the map with other maps, the relative importance of a feature map with respect to other ones is retained, while irrelevant information extracted from ineffective feature map is suppressed.

Finally, three S^k maps are just simply summed into a single saliency map S by

$$S_{x,y} = S^1_{x,y} + S^2_{x,y} + S^3_{x,y} \tag{7}$$

4 Experimental Results

We have tested our system with various images, ranging from simple synthetic images to complex real images of natural environment. And we have also tested with the images corrputed with heavy amouts of noise.

Fig. 2 shows one example of our system's overall working. In this case, we applied the system to traffic sign detection task, and the target is yellow trafic sign. First, raw color image is inputted to the system, and is decomposed into three early visual feature maps, EF^1 for intensity contrast, EF^2 for red/blue opponency, and EF^3 for green/yellow opponency, in parallel. These maps are then reorganized into F^1, F^2, and F^3. Each reorganized feature map has orientation features addition to pre-computed features, and also has enhanced pixel values which are largely different from their surroundings'. These reorganized feature maps are then integrated into a saliency map by statistical information and competitive relations of the pixels in each enhanced feature map. The integrated saliency map shows that the most salient region is yellow traffic sign, and it indicates that our system works very well.

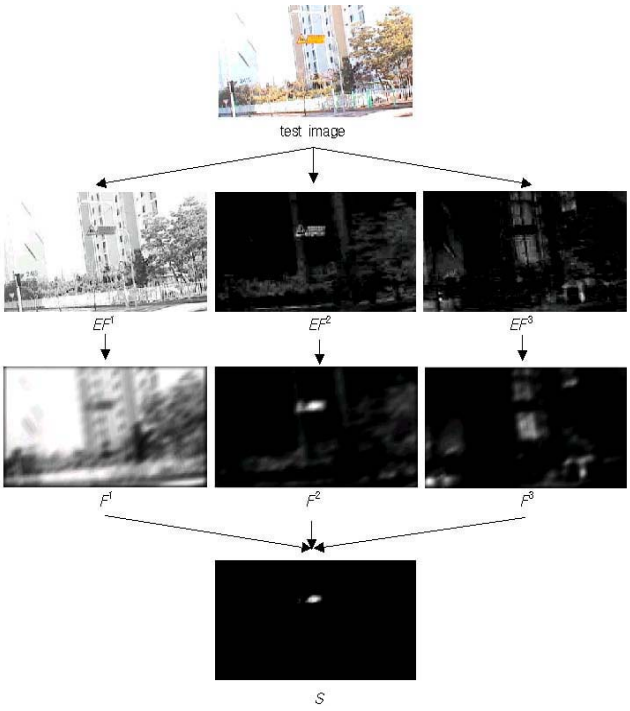


Fig.2 An example of overall working of our system

We divided the results into two groups, which were generated with non-noisy and noisy image. In what follows, we will show these results.

3.1 Results with Non-Noisy Images

As synthetic images, we used various images in which the 'targets' are differed in orientations, in colors, in sizes, in shapes, or in intensity contrast, with a set of 'distractors'. In such cases, the system detected the targets immediately. Fig.3 shows some results with these kinds of images. Fig.3(a)~(b) shows the results of shape pop-out task. The input image in Fig.3(a) has background of lighter contrast than those of foreground, and the test image in Fig.3(b) is vice versa. Fig.3(c) shows the results of orientation pop-out task, and Fig.3(d) shows those of color pop-out task.

As real images, each of the images contains the target object such as signboard, signal lamp, traffic sign, mailbox, placard, and so forth as well as distractors such as strong local variations in illumination, textures, or other non-targets. Fig.4 shows some results with color images of natural environments. Three typed test images shown in Fig.4 are arranged in order of complexity degree of their background and the quality. Fig.4(a) shows the result with an image who has very complex background.

Fig.4(b) shows the result with an image whose complexity degree of its background is simpler than that of an image shown in Fig.4(a), and which has strong local variations in illumination. Fig.4(c) shows the result with an image in which relatively simple background is contained and which was photographed inside the building.

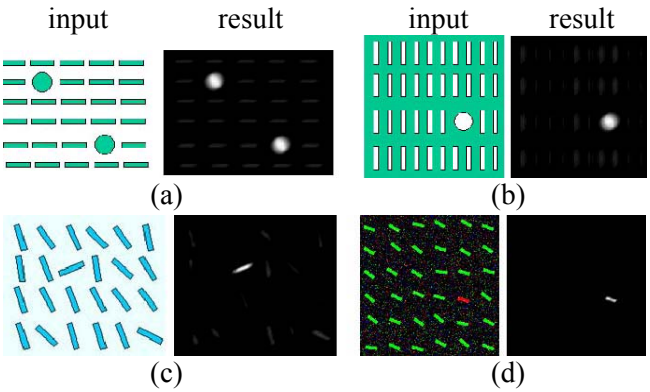


Fig.3 Some results with synthetic images which were not corrupted with noise. (a)~(b) : shape pop-out task, (c) : orientation pop-out task, (d) : color pop-out task

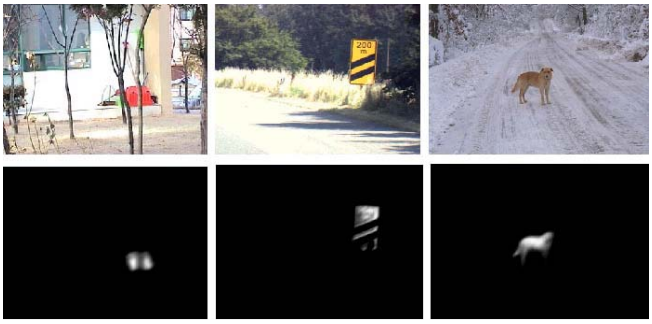
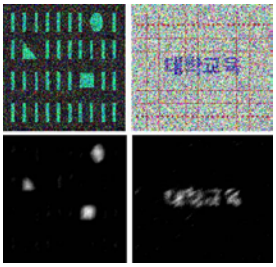


Fig.4 Some results with real images which were not corrupted with noise. The target is (a) red basket, (b) yellow traffic sign, (c) yellow dog

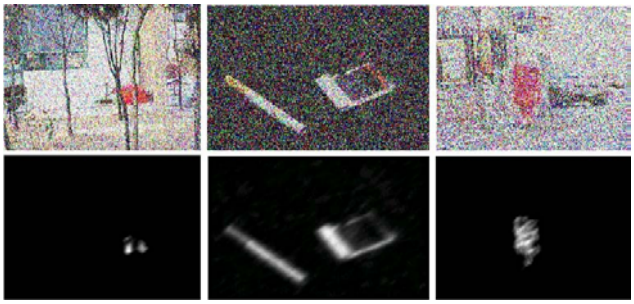
3.2 Results with Noisy Images

As noisy images, we intentionally added heavy amounts of noise(noise density : 40%~90%) to the images which were described and tested in section 3.1. Added noise has some properties : color noise(e.g. Fig.5) or black and white noise(e.g. Fig.6), and gaussian distribution noise(e.g. Fig.5(a), Fig.6(a)) or uniform distribution noise(e.g. Fig.5(b), Fig.6(b)).

Fig.5 and Fig. 6 shows the results with the images corrupted with heavy amounts of noise. The test image in Fig. 5 and Fig. 6 were corrupted with color noise and black and white noise respectively, with different

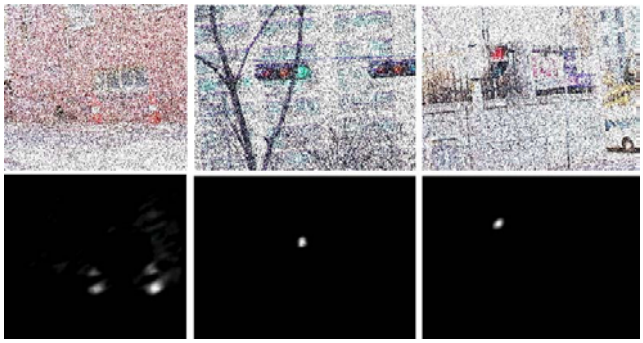


(a) gaussian distribution noise

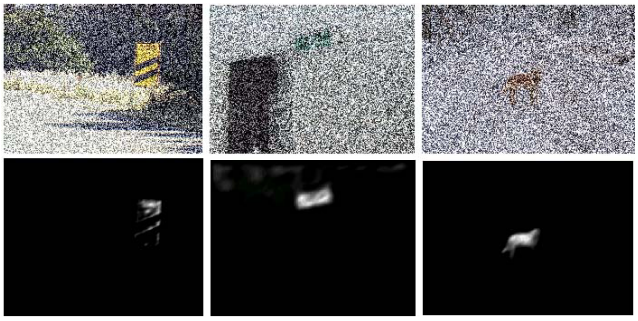


(b) uniform distribution noise

Fig. 5. Color Noise : the target is (a) (left) three different shaped objects(noise density:50%) (right) blue characters(noise density:70%), (b) (left) red basket(noise density:30%) (center) detecting cigarette and the light(noise density:90%) (right) detecting red mail box(noise density:90%)



(a) gaussian distribution noise (noise density:40%)



(b) uniform distribution noise (noise density:60%)

Fig.6 Black and white Noise : the target is (a) (left) red emergency triangle (center) green light (right) red light, (b) (left) yellow traffic sign (center) green emergency lamp (right) yellow dog

distribution property.

As shown in Fig. 5 and Fig.6, the results were very promising, and the system proved very robust to those kinds of noise. Besides, if the noise were absent from the image, the results would be much better than the result shown in Fig.5 and Fig.6 as matter of course(See Fig. 4 for example).

5 Discussion and Conclusion

We reported in this paper a new selective attention based method for target detection. The system proposed herein identifies the regions of an image which contain the most "interesting" features by only bottom-up information of input visual stimuli. It does not require any a-priori knowledge about the target. As our modeling approach is bottom-up, we cannot account for the system's top-down effects as a matter of course. However, there are many potential ways to extend our system. Modifying the feature maps with the simulated top-down knowledge which is trained by neural processing and which is presented in the feature extraction process, might be the one way.

Our system is composed of two main stages, feature extraction and integration stage. Several basic features are extracted directly from visual stimuli first, and these extracted features are integrated based on their local competitive relations and statistical information. We have studied the overall behaviors of the model with input from two large classes of noisy images. The first are visual scenes constructed analogously to the stimuli typically presented in psychophysical studies of visual search. The second are color images of natural environment taken from different domains. As shown in experimental results, the performance of the system was good, and it shows the promise that it could be successfully used as a target detector in complex real images for general purpose.

References:

[1] M.Bear, B.Cornors, and M.Paradiso,*Neuroscience exploring the brain*. Williams and Wilkins, USA, 1996

[2] K.Cave, and J.Wolfe, Modeling the Role of Parallel Processing in Visual Search, *Cognitive Psychology*, Vol.22, 1990, pp.225-271.

[3] D.Chapman, *Vision, Instruction, and Action*. Ph.D Thesis. Massachusetts Institute of Technology, 1990

[4] J.Duncan, and J.Humphreys, Visual search and stimulus similarity, *Psychological Reviews,* Vol.96, 1989, pp.433-458.

[5] O.Gallat, P.Gaussier, and J.Cocquerez, A Model of the Visual Attention to Speed up Image Analysis. *Proceedings of International Conference on Image Processing* , Vol.1, 1998, pp.246-250.

[6] J.Jonides, Further toward a model of the mind's eye's movement, *Bulletin of the Psychonomic Society*, Vol.21, No.4, 1983, pp.247-450.

[7] B.Julesz, and J.Bergen, Textons, the fundamental Elements in Preattentive Vision and Perception of Textures, *Bell Systems Technical Journal*, Vol.62, 1983, pp.1619-1643.

[8] L.Itti, and C.Koch, A saliency-based search mechanism for overt and covert shifts of visual attention, *Vision Research*, Vol.40, No.10-12, 2000, pp.1489-1506.

[9] L.Itti, C.Gold, and C.Koch, Visual Attention and Target Detection in Cluttered Natural Scenes, *Optical Engineering*, Vol.40, No.9, 2001, pp.1784-1793.

[10] C.Koch, S.Ullman, Shifts in Selective Visual Attention : Towards the Underlying Neural Circuitry. *Human Neurobiology*, Vol. 4, 1985, pp.219-227.

[11] R.Milanese, H.Wechsler, S.Gil, J.Bost, and T.Pun, Integration of Bottom-up and Top-down Cues for Visual Attention Using Non-Linear Relaxation, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp.781-785.

[12] M.Mozer, *The Perception of Multiple Objects : a Connectionist Approach*, MIT Press, Cambridge, MA, 1991

[13] P.Sandon, Simulating Visual Attention. *Journal of Cognitive NeuroScience*, Vol.2, No.3, 1990, pp.213-231.

[14] A.Treisman and G.Gelade, A Feature-integration theory of Attention, *Cognitive Psychology*, Vol.12, No.1, 1980, pp.97-136.

[15] W.Osberger and A.J.Maeder, Automatic identification of Perceptually important regions in an image. *Proceedings of Fourteenth International Conference On Pattern Recognition*, Vol. 1, 1998, pp. 701-704.