Detecting Perceptually Important Regions in an Image Based on Human Visual Attention Characteristic

Kyungjoo Cheoi and Yillbyung Lee

Dept. of Computer Science, Yonsei University
134 Sinchon-Dong, Seodaemun-Gu, Seoul, 120-749, Korea
{kjcheoi,yblee}@csai.yonsei.ac.kr

Abstract. In this paper a new method of automatically detecting perceptually important regions in an image is described. The method uses bottom-up components of human visual attention, and includes the following three components : i) several feature maps known to influence human visual attention, which are computed in parallel directly from the original input image, ii) importance maps, each of which has the measure of "perceptual importance" of local regions of pixels in each corresponding feature map, and are computed based on lateral inhibition scheme, iii) single saliency map, integrated across multiple importance maps based on a simple iterative non-linear mechanism which uses statistical information and local competence of pixels in importance maps. The performance of the system was evaluated over some synthetic and complex real images. Experimental results indicate that our method correlates well with human perception of visually important regions.

1 Introduction

We can say that the main problem in computer vision lies in its limited ability which is caused by enlarging the size of a given image, and the computational complexity followed by it. Actually, computer vision system receives vast amount of visual information, and real-time image capturing at any useful image resolution yields prodigious quantities of visual information. Therefore, analyzing all of inputted visual information for high-level process, such as object recognition, is actually impossible, and is also unnecessary in aspects of using limited computational resources efficiently. Therefore, the mechanism of selecting and analyzing only the information "most" relevant to the current visual task is needed to computer vision system.

It is known that human visual system does not handle all visual information received by the eye but selects and processes only the information essential to the task at hand while ignoring a vast flow of irrelevant details [7]. Many existent experimental evidences about primate report that there are a lot of mechanisms related to the function of "visual selection", and visual attention belongs to this mechanism.

T. Caelli et al. (Eds.): SSPR&SPR 2002, LNCS 2396, pp. 329-338, 2002.

Visual attention is one of the primate's most important intellectual ability that maximizes visual information processing capability by rapidly finding the portion of an image with which the information is most relevant to the current goals(See Colby's work [4] for a neurophysiological review). From these, one usable method of reducing prodigious quantities of visual information of input image is deploying the function of human visual attention within the system. That is, extract the regions of interest from the image which usually constitute a considerably lesser proportion of the whole image, and discard the rest, non-interest regions [12].

This paper describes a new method of automatically detecting salient regions in an image based on the bottom-up human visual attention characteristic. The proposed method can be explained by following three stages (See Fig. 1). First, the input image is represented in several independent feature maps, two chromatic feature maps and one achromatic feature map. Second, all feature maps are converted into corresponding number of importance map by lateral inhibition scheme. The importance map has the measure of "perceptual importance" of local regions of pixels in feature map. Third, all importance maps are combined into a single representation, saliency map. Iterative non-linear mechanism using statistical information and local competence of pixels in importance map is processed on all of importance maps and the output is just simply summed. The saliency map represents the saliency of the pixels at every location in an image by a scalar quantity in relation to its perceptual importance, and guides the selection of attended regions [7].

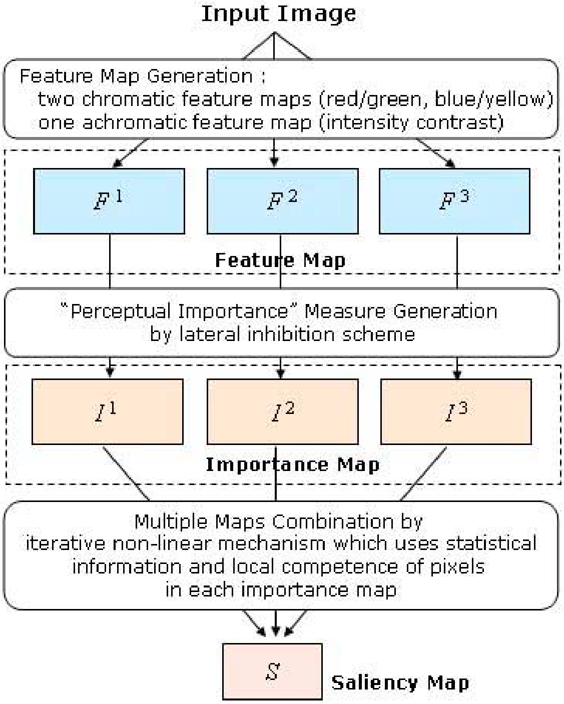

Input Image
Feature Map Generation :
two chromatic feature maps (red/green, blue/yellow)
one achromatic feature map (intensity contrast)
F^1
F^2
F^3
Feature Map
"Perceptual Importance" Measure Generation
by lateral inhibition scheme
I^1
I^2
I^3
Importance Map
Multiple Maps Combination by
iterative non-linear mechanism which uses statistical
information and local competence of pixels
in each importance map
S
Saliency Map

Fig. 1. Overall Architecture of the System Proposed

The organization of the paper is as follows. Related works are given in Section 2. In Section 3, the proposed method for detecting salient regions is explained. Experimental results are shown in Section 4, and concluding remarks are made in Section 5.

2 Related Works

Researches related with visual attention have been studied in two primary approaches according to the ways of directing attention, that is, the *bottom-up*(or *data-driven*) approach and the *top-down*(or *model-driven*) approach.

In bottom-up approach, the system selects regions of interest by bottom-up cues obtained by extracting various elementary features of visual stimuli [3,6,7,13]. And in top-down approach, the system uses top-down cues obtained from a-priori knowledge about current visual task [8]. A hybrid approach combining both bottom-up and top-down cues has also been reported [2,5,9,10].

As bottom-up models do not use any kind of priori knowledge about the given task, they can be employed on a variety of applications without major changes in the architecture. Also the most of what is known about human visual attention is related to the bottom-up cues. Meanwhile, almost all previous top-down systems neglect bottom-up cue, so they are very useful to match specific patterns whose high-level information was presented to the system. In such cases, the system needs training process and also needs partial interaction with the recognition system. Therefore, it is very difficult to extend the top-down system to other applications. Because of these reasons, relatively few studies have been made to provide quantitative top-down systems although the importance of top-down cues of attention has long been emerged.

Treisman's "Feature Integration Theory" [13] which proposed to explain strategies of human visual search has been very influential theory of attention. The first biological plausible computational model for controlling visual attention was proposed by Koch and Ullman [7]. Many successful computational models for bottom-up control of visual attention have the common stages of computing several feature maps and single saliency map. The differences between those models are the differences of the strategies used to create feature maps and the saliency map. Among existing computational models, our system is built at the basis of [6] and [9]. Itti *et al.* proposed the purely bottom-up attention model that consists of saliency map and winner-take-all network [6], and Milanese's model [9] extracts regions of interest by integrating the bottom-up and the top-down cues by a non-linear relaxation technique using energy minimization-like procedure.

At least two main remarks can be made about most of the systems reviewed in this section. The first remark is that, the most of existing systems are in the progress of establishing the concept of visual attention, and they put too much emphasis on the theoretical aspects of human visual attention, not the real aspects. The second remark is that, in many cases, the performance of most of the systems have been evaluated over just synthetic or simple simulated images, so they yield a rare example of a system that can be applied to natural color images.

From these two remarks, we can conclude that the existent systems are not general-purpose enough or widely applied to real actual problem of visual world yet. Our method proposed here is designed to extend the capabilities of previous systems. In doing so it proved that our system was suitable for applications to real color images including noisy images.

3 The System

Our system detects regions of interest by properties of the input image without any a-priori knowledge. As shown in Fig. 1, our system has three main components, the feature map, the importance map, and the saliency map. In this section, these three components are described in detail.

3.1 The Feature Maps

Two kinds of topographic feature maps known to influence human visual attention are generated from the input image : two chromatic feature maps for color contrast and one achromatic feature map for intensity contrast.

The chromatic information is one of the biggest properties of human vision that discriminates an object from others, and psychophysical results also show that it is available for pre-attentive selection. In human vision, the spectral information is represented by the collective responses of the three main types of cones(*R*, *G*, *B*) in retina. These responses are projected to the ganglion cells, and then to the LGN, and to the visual cortex. In this way, we can get both chromatic and achromatic information about the input objects. In V1, there exist three types of cells with center-surround receptive fields, homogeneous receptive fields, and more complex receptive fields which combine above mentioned two types. Among them, the cells with homogeneous receptive fields respond the highest when both the center and the surround receives the same stimuli of a specific wavelength, and this means that they are not spatially selective but responds very strongly to color contrast. From these, two chromatic feature maps which simulate the effect of two types of color opponency exhibited by the cells with homogeneous receptive fields are generated. The process of generating two kinds of chromatic feature maps is as follows.

First, red, green and blue components of the original input image are extracted as *R*, *G*, and *B*, and four broadly tuned color channels are created by

$$r=R-(G+B)/2,\quad g=G-(R+B)/2,\quad b=B-(R+G)/2,\quad y=R+G-2(|R-G|+2) \tag{1}$$

where *r*, *g*, *b*, and *y* denote red, green, blue, and yellow channels respectively. Each channel yields maximal response for pure, fully saturated hue to which it is tuned, and yields zero response both for black and for white inputs.

Second, based on above color channels, two chromatic feature maps are created by

$$F^1 = r - g,\quad F^2 = b - y \tag{2}$$

F^1 is generated to account for red/green color opponency, and F^2 for blue/yellow color opponency.

If no chromatic information is available, the *gray-level*(or *intensity*) image can be used as an achromatic feature map. Gray-level information can be obtained from the chromatic information of the original color input image as I = (R+G+B) / 3, and is used as an achromatic feature map F^3.

$$\mathrm{F}^3 = \mathrm{I} \tag{3}$$

These generated multiple independent feature maps are then normalized in the range of 0~1 in order to eliminate across-modality differences due to dissimilar feature extraction mechanisms, and to simplify further processing of the feature maps.

3.2 The Importance Maps

Since each of the computed feature maps has the special meaning at every locations of input image, we have to assign measure of importance to each of the feature maps in order to detect salient regions based on this. We used center-surround operator, based on the *DOOrG*(Difference-Of-Oriented-Gaussians) model [9] to generate corresponding number of importance maps. This operator is also based on lateral inhibition scheme which compares local values of the feature maps to their surround and enhances those values strongly different from their surroundings' while inhibiting the others. Aguilar and Ross have suggested that the regions of interest are those regions which differ the most [1]. With this operator, the system also can have the effect of reducing noises. The processing of generating importance maps for each available feature map is as follows.

First, construct filter bank h at 8 orientations (Fig. 2) by

$$h_{x',y'}(\theta) = \left| DOOrG_{x',y'}(\sigma, r_{x'/y'}, r_{on/off}) \right| \tag{4}$$

where $DOOrG_{x',y'}(\cdot,\cdot,\cdot)$ denotes 2-D *DOOrG* function. The *DOOrG* model is defined by the difference of two Gaussians of different sizes with the width of positive Gaussian being smaller than the width of the negative one. The two Gaussians may have an elliptic shape characterized by different width of the two Gaussians while the *DoG*(Difference-of-Gaussian) model has isotropic shape of Gaussians. See [9] for more details. If we change a coordinate, it is possible to extend the canonical *DOOrG* model to vary the orientation of the filter. In our system, θ is fixed as $\theta \in \{0, \pi/8, 2\pi/8, \cdots, 7\pi/8\}$ and the values of other parameters are as follows: σ=5.5, $r_{x'/y'}$=1/9, $r_{on/off}$ =4.76, K_1=1/6, K_2=17/60.

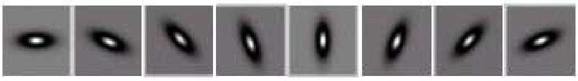


Fig. 2. Generated filter bank $h(\theta)$

Second, for each importance map, convolution is processed over the map with eight $h(\theta)$ filters, and then the results are squared to enhance the contrast. Finally, all computed maps are summed to factor out θ.

Since the importance map computed in this section used filter bank based on the *DOOrG* model at 8 orientations, the system can have the ability of detecting orientation.

3.3 The Saliency Map

In general, the system extracts perceptually important regions based on importance measures provided by importance maps. The difficulty of using these measures resides in the fact that importance maps are derived from different types of feature maps. Since importance map provides different measures of importance for same image, each map may guide different regions as a salient region. To settle down this problem, different measures must be integrated in order to obtain a single global measure of importance for each location of image. In this case, computed global measure could guide the detection of the final salient regions. But, combining information across multiple importance maps is not an easy work. In principle, this could be done by taking high activity pixels over all information [2,11] or by weighted sum of all information [3,14]. However, in the former case, there is no reason why a high intensity region should be more interesting than a low intensity one. And in the latter case, the results highly depend on the appropriate choice of the weights. Also, both do not consider the fact that each importance map represents different measures of importance about the "same" input image.

Here, we propose a simple iterative non-linear combination mechanism which uses statistical information and local competence of pixels in importance maps. Our method promotes those maps in which small number of meaningful high activity areas present while suppressing others. The saliency map has been generated through following three steps.

At the first step, important maps I^k (k=1,2,3), computed in section 3.2, are inputted to the system. Each importance map is convolved with the large size of the *LoG* filter and the result is added with the original input one. Iterate this procedure several times, and corresponding number of IT^k maps are generated as a result. This procedure causes the effect of short-range cooperation and long-range competition among neighboring values of the map. And we can also have the advantage of reducing noises of an image. The *LoG* function we used is given by

$$LoG(x,y) = \frac{1}{\pi\sigma^4}\left[1 - \frac{x^2+y^2}{2\sigma^2}\right] \cdot e^{-\left(\frac{x^2+y^2}{2\sigma^2}\right)} \tag{6}$$

where σ denotes the scale of the Gaussian. We set σ as 3.6.

At the second step, each IT^k map is evaluated iteratively by statistical information of the map to enhance the values associated with strong peak activities in the map while suppressing uniform peak activities. For each IT^k map, update the map by

$$\mathrm{IT}^{\mathrm{k}} = \mathrm{IT}^{\mathrm{k}} \times (\mathrm{GMax}^{\mathrm{k}} - \mathrm{Ave}^{\mathrm{k}})^2 \tag{7}$$

where $GMax^k$ denotes the global maximum value of the map and Ave^k denotes the average value of the map. After this, normalization is processed on each of computed IT^k maps by

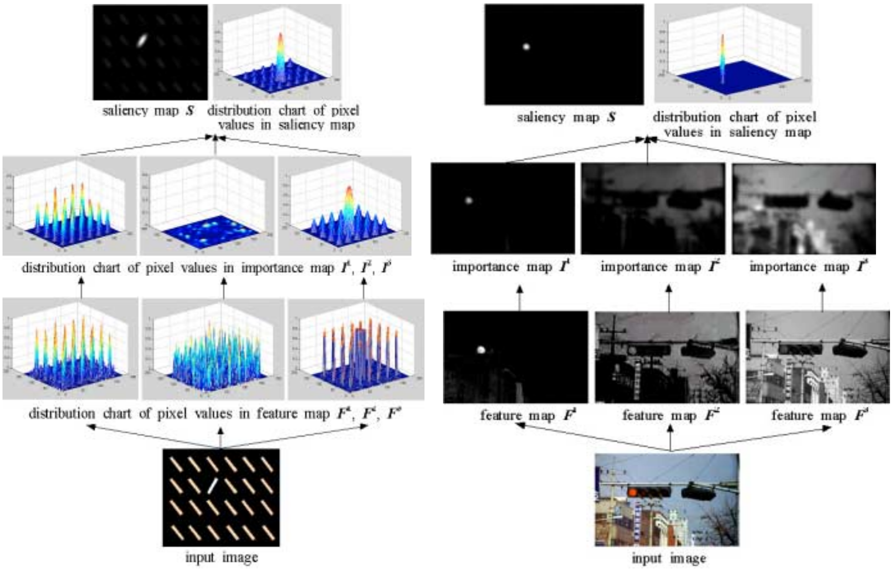
$$IT^k = \frac{IT^k - IT_{\min}}{IT_{\max} - IT_{\min}} \tag{8}$$

where $IT_{\min}$ and $IT_{\max}$ denote the global minimum and the maximum value out of all IT^k maps. Through this, relative importance of an IT^k map with respect to other ones would be remained, and irrelevant information extracted from ineffective IT^k maps would be suppressed. Iterate the procedure of this step several times. Here, we iterated 4 times.

At the third step, computed IT^k maps are summed and normalized to a fixed range of 0~1 to generate saliency map *S*.

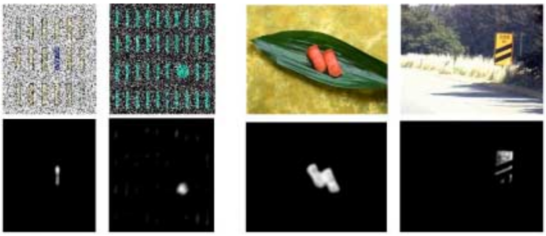
4 Experimental Results and Analysis

To evaluate the performance of our system, we used three kinds of experimental images. In this section, we will describe what kinds of images are we used, and the experimental results, in detail. As explained already, our system was developed in order to solve several problems caused by enlarging the size of input image in computer vision system, through selecting regions of interest which humans think to be perceptually important. Therefore, it is useless to use images which contain only the target. So, we used images of not only the target, but images including complex background or other objects photographed at a great distance. By the way, many previous researchers have been concentrated more on the evaluation of their system's performance on simple synthetic images, not real images. However, this is not proper from the fact that computer vision system actually operates in the real world. Therefore, we cannot neglect the system's performance on complex real images. So, we used various images from simple synthetic images to complex real images. Besides, many images of real visual world may have lots of noises caused by the properties of images themselves, or added through the image acquiring processes. For these reasons, we included our testing with noisy images. With the images selected by above mentioned three criteria, we tested our system. And through experimental results, we've found that our system detects the interest part of an image that a human is likely to attend to, and it has following three properties. First, the system was able to reproduce human performance for a number of pop-out tasks [13], using images of the type shown in Fig. 3(a). A target defined by a simple and unique feature such as color, orientation, size, contrast, etc. distinguishing it without any ambiguity, or isolated, is easily detected at almost constant time independent from the number of the other stimuli. To evaluate the system's performance of this paradigm, we used various images that differed in orientations by 30°, 45°, 90°, in colors by red, blue, green, white, black, yellow, in sizes, and in intensity contrasts. Also, the system was tested with the images of which the background has lighter contrast than those of the target, and vice versa. The system detected the target properly, and some results of these tasks were shown in Fig. 3(a), Fig 4(a). Second, the system could be successfully applied to complex real images. The system was tested with complex real color and gray-level images such as signal lamp image of the type shown in Fig. 3(b) and various images of traffic sign, food, animal, and natural scenes. See Fig 4(b) for example. One major difficulty of deciding whether the result is good or not is that each person may choose a different region as the most salient region. However, if we follow the assumption that the most salient region to which attention goes is an object of interest, the results for complex real images are successful. Third, the system was very strong to noises. See Fig. 4(a) for example.


saliency map S
distribution chart of pixel values in saliency map
distribution chart of pixel values in importance map I^1, I^2, I^3
distribution chart of pixel values in feature map F^1, F^2, F^3
input image
saliency map S
distribution chart of pixel values in saliency map
importance map I^1
importance map I^2
importance map I^3
feature map F^1
feature map F^2
feature map F^3
input image

(a) (b)

Fig. 3. Examples of experimental results for synthetic and real images : (a) orientation pop-out task. Orientation is detected in importance map, and this feature wins among other features through the procedure of saliency map generation (b) detects red signal lamp



(a) (b)

Fig. 4. Some more examples. (a) Noisy Image : (left) color pop-out task(middle:blue, the very left upside and right downside:light green, remainder:yellow):detects blue bar, (right) size pop-out task:detects circle shaped object (b) Non-Noisy Image : (left) detects red sliced raw fish, (right) detects yellow traffic sign

5 Concluding Remarks

In this paper, we proposed a new method of detecting salient regions in an image in order to solve several problems caused by enlarging the size of input image in computer vision system. The proposed method uses only bottom-up components of human visual attention. As shown in experimental results, the performance of the system is very strong to not only synthetic images but also complex real images, although the system employed very simple mechanisms in feature extraction and combination. Also our system can be extended to other vision applications such as arbitrary target detection tasks through just simply modifying feature maps. However, our method needs more experiments and analysis with more complex real and noisy images in order to confirm whether our system can be applicable to other various actual problems. And we are currently doing this kind of job with more complex real and noisy images. In addition, as human visual attention actually depends on both bottom-up and top-down controls, researches to integrate the proposed method with top-down cue still has to be carried out.

References

1. Aguilar, M., Ross, W.:Incremental art:A neural network system for recognition by incremental feature extraction. Proc. of WCNN-93 (1993)
2. Cave, K., Wolfe, J.: Modeling the Role of Parallel Processing in Visual Search. Cognitive Psychology 22 (1990) 225-271
3. Chapman, D.: Vision, Instruction, and Action. Ph.D. Thesis, AI Laboratory, Massachusetts Institute of Technology (1990)
4. Colby:The neuroanatomy and neurophysiology of attention. Journal of Child Neurology 6 (1991) 90-118
5. Exel, S., Pessoa, L.:Attentive visual recognition. Proc. of Intl. Conf. on Pattern Recognition 1 (1998) 690-692
6. Itti, L., Koch, C., Niebur, E.: Model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (1998) 1254-1259
7. Koch, C., Ullman, S.: Shifts in Selective Visual Attention : Towards the Underlying Neural Circuitry. Human Neurobiology 4 (1985) 219-227
8. Laar, P., Heskes, T., Gielen, S.:Task-Dependent Learning of Attention. Neural Networks 10, 6 (1997) 981-992
9. Milanese, R., Wechsler, H., Gil, S., Bost, J.,Pun, T.: Integration of Bottom-up and Top-down Cues for Visual Attention Using Non-Linear Relaxation. Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (1994) 781-785
10. Olivier, S., Yasuo, K., Gordon, C.:Development of a Biologically Inspired Real-Time Visual Attention System. In:Lee, S.-W.,Buelthoff, H.-H., Poggio, T.(eds.):BMCV 2000.Lecture Notes in Computer Science, Vol. 1811. Springer-Verlag, Berlin Heidelberg New York (2000) 150–159
11. Olshausen, B., Essen, D., Anderson, C.: A neurobiological model of visual attention and Invariant pattern recognition based on dynamic routing of information. NeuroScience 13 (1993) 4700-4719
12. Stewart, B., Reading, I., Thomson, M., Wan, C., Binnie, T.: Directing attention for traffic scene analysis. Proc. of Intl. Conf. on Image Processing and Its Applications (1995) 801-805

13. Treisman, A.-M., Gelade, G.-A.: A Feature-integration Theory of Attention. Cognitive Psychology 12 (1980) 97-136
14. Yagi, T., Asano, N., Makita, S., Uchikawa, Y.:Active vision inspired by mammalian fixation mechanism. Intelligent Robots and Systems (1995) 39-47