A NONLINEAR FEATURE FUSION BY VARIADIC NEURAL NETWORK IN SALIENCY-BASED VISUAL ATTENTION

Zahra Kouchaki[1] and Ali Motie Nasrabadi[2]
[1]*Department of Biomedical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran*
[2]*Department of Biomedical Engineering, Shahed University, Tehran, Iran*

Keywords: Saliency Map, Visual Attention, Nonlinear Fusion, Neural Network, Object Detection.

Abstract: This study presents a novel combinational visual attention system which applies both bottom-up and top-down information. This can be employed in further processing such as object detection and recognition purpose. This biologically-plausible model uses nonlinear fusion of feature maps instead of simple superposition by employing a specific Artificial Neural Network (ANN) as combination operator. After extracting 42 feature maps by Itti's model, they are weighed purposefully through several training images with their corresponding target masks to highlight the target in the final saliency map. In fact, the weights of 42 feature maps are proportional to their influence on finding target in the final saliency map. The lack of bottom-up information is compensated by applying top-down information with available target masks. Our model could automatically detect the conceptual features of desired object only by considering the target information. We have tried to model the process of combining 42 feature maps to form saliency map by applying the neural network which resembles biological neural network. The Experimental results and comparing our model with the basic saliency model using 32 images of test dataset indicate a noticeable improvement in finding target in the first hit.

1 INTRODUCTION AND RELATED WORKS

One of the effective abilities of the human is visual attention system which directs the human vision to the most intersecting parts of a scene. These parts are called salient regions and their saliencies are corresponding to how much attention can focus on them. When we don't have special goal, the low-level visual features could attract our attention which then will be sent in higher cognitive areas for further processing such as object recognition. However, attention is also dependent on top-down features such as prior knowledge which is extracted from higher brain areas. Both bottom-up and top-down cues contribute in directing attention toward the most salient points. Selective visual attention has so many applications in computer vision such as automatic target detection, navigational aids and robotic control (Itti and Koch, 2001). Till now, many computational models of visual attention are proposed which simulate human visual attention based on bottom-up information. In terms of psychology, Treisman and Gelade (1980) proposed the theory of feature integration in visual attention which is the basic theory for the most bottom-up models such as (Itti, Koch, and Niebur, 1998), (Koch and Ulman, 1985), (Sun and Fisher, 2003). In all these models, bottom-up low-level cues such as color, intensity and orientation could detect salient points by their contrast. As it is proved psychologically, top-down information is also effective in directing low-level visual cues toward salient regions (Wolfe, 1994). Some scientist has studied on top-down visual attention recently (Wolfe, 1994), (Navalpakam, Rebesco and Itti, 2005), (Frintrop, 2006). All of the mentioned models of bottom-up and top-down visual attention have applied linear fusion of feature maps which does not seem plausible biologically. Although some scientists have studied in the way of combining feature maps (Itti and Koch, 2001), (Walter, Itti, Riesenhuber, Poggio and koch, 2002), most of them have employed linear fusion of feature maps. In (Itti and Koch, 2001), four different approaches of fusion of feature maps were presented. Among the four strategies, the approach of *linear combination with learned weights* had the best performance in finding the target. However, it is still a linear fusion and also

Kouchaki Z. and Motie Nasrabadi A..
A NONLINEAR FEATURE FUSION BY VARIADIC NEURAL NETWORK IN SALIENCY-BASED VISUAL ATTENTION.
DOI: 10.5220/0003859204570461
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2012), pages 457-461
ISBN: 978-989-8565-03-7

all the features obtain positive weights even if they may erode visual attention. As we have shown in our previous work, nonlinear fusion of feature maps sounds more reasonable biologically. In (Kouchaki, Nasrabadi and Maghooli, 2011), we proposed a novel nonlinear feature fusion strategy to fuse three conspicuity maps through *Fuzzy Interface System* which had better results in comparison with the basic saliency model in detecting desired object. However, it combines three conspicuity maps rather than 42 feature maps that could be more effective. Moreover, in (Bahmani, Nasrabadi, Hashemi Golpayegani, 2008), a combinational approach of *multiplicative weighted feature maps* was proposed which multiply 42 feature maps after weighing them purposefully as in (Itti and Koch, 2001). Although a remarkable improvement was achieved, the simplest nonlinear function was employed. Unquestionably, the real biological system of visual attention is more complicated than a simple multiplication. In this study, we have tried to indicate the biological process of consisting saliency map through 42 feature maps. In order to select a nonlinear function, which could show the details of creating saliency map from 42 feature maps more reasonably, we thought of *Artificial Neural Network (ANN)* as it considerably resembles the biological neural network.

After extracting 42 feature maps by Itti's model (Itti et al, 1998), we applied them as the inputs of the network. The 42 feature maps were weighed automatically through training process by considering target masks as the desired output of the network. In fact, we compensated the lack of bottom-up information with considering target information as top-down cues to adjust desired weights.

The rest of this paper is as follows. In section 2, we discuss about the basic bottom-up model of visual attention. Then we present our methodology in section 3. The details of the *Variadic Neural Network* structure will be discussed in section 4. Section 5 presents the details of our proposed model. Experimental results are discussed in section 6. Finally, section 7 concludes the paper.

2 THE BASIC BOTTOM-UP MODEL

This part discuss about the details of computing the bottom-up saliency map which proposed by Itti et al (1998). Whereas an image is placed at the input of the Itti's model, it is filtered by a low-pass filter. After low pass filtering, different spatial scales are generated in three different channels of colour, intensity and orientation by *Dyadic Gaussian Pyramids*. These Gaussian Pyramids subsample the input colour image in different scales. After that, the feature maps are constructed in three different channels of colour, intensity and orientation with "centre-surround "operation. Subtraction between fine and coarse scales images, which is a point-by-point subtraction, yields 42 feature maps that consist of 12 colors, 6 intensities and 24 orientation maps. All the feature maps in each channel are linearly fused into a conspicuity map which finally leads to three conspicuity maps. Each conspicuity map is an indication for one of the three features. After linear combination of three conspicuity maps, the final saliency map is formed which is based on the bottom-up cues.

3 METHODOLOGY

In this study we want to promote some of the computational weaknesses of the bottom-up visual attention models for the object detection purpose. In this study, we thought of designing a nonlinear fusion kernel for combining 42 feature maps which can indicate the biological details of forming saliency map. Furthermore, the feature maps should be weighed purposefully to be fit for object detection purpose (Walther, 2006). As a result, we assumed that *Artificial Neural Network* could be a good choice for nonlinear fusion of 42 feature maps as it resembles biological neural network. However, in the beginning, combination of 42 images with the big size through neural network seemed impossible due to having 42 huge sized images as the inputs to the network. But, finally, we found the *Variadic Neural Network* (McGregor, 2007), as a suitable network which could meet our needs in this respect due to accepting *n*-dimensional vectors as its inputs. The top-down information could be considered in the model by training the network using available target masks. The supervised neural network could be trained using the target information to weigh the 42 feature maps purposefully. As we know, searching the desired object which is known previously for the viewer is easier and faster than searching it without prior knowledge. As we know, one of the important factors for modelling the human visual attention is considering the learning ability. We have considered this matter with training the network. The proposed visual attention structure is illustrated in Figure 1. As shown, after deriving 42

feature maps by Itti's model (Itti et al, 1998), they are considered as the inputs of the network and the output of the network is the final saliency map.

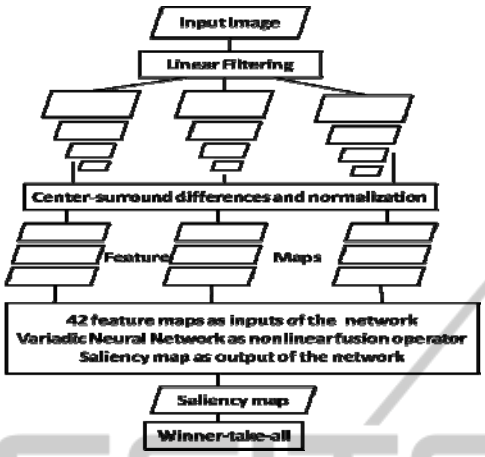

Figure 1: The proposed visual attention structure using Variadic Neural Network (Itti et al, 1998).

4 VARIADIC NEURAL NETWORK STRUCTURE

Variadic Neural Network (*VNN*) structure which is extracted from multi-layer perceptron architecture was suggested by McGregor in 2007. The inputs of the network are vectors of the same length with arbitrary dimension. The number of parameters in the network and its architecture do not vary by changing the dimension of the input's vectors. There are variadic links which connect two variadic neurons. The inputs and outputs of a link are vectors which their sizes are optional. There are two link parameters. The *multiplicative weight parameter* (*w*) is the same as the weight parameter in the traditional feed-forward network. But there is also an *interaction-weight parameter* (θ) which causes the interaction between different elements of an input vector. In the *Variadic Neural Network* (*VNN*) structure, the inputs to the neuron are the vectors of real numbers instead of real scalar in the traditional network. The variadic neuron sums the variadic inputs to generate vector activation. After adding the bias to the vector activation and passing through the *hyperbolic tangent* functions, the variadic output of the neuron will be obtained. More details could be seen in (McGregor, 2007).

5 MODEL

As it is illustrated in Figure 1, in the first stage of our proposed model, 42 feature maps are extracted using Itti's model (Itti et al, 1998). Since the inputs of the network in the *Variadic Neural Network* are *n*-dimensional vectors, we should change the training images to the training vectors. So, after extracting 42 feature maps with the size of (30×40) pixels from each 32 training images, we should change them to 32 vectors of 1200-dimentional. In addition, 32 target masks with the size of (640×480) pixels corresponding to the training dataset should be resized to (30×40) pixels and then to the vectors of 1200-dimentional to be acceptable for placing as desired output of the network. The same process should be done for the test dataset except for the target masks. The test process does not need the target mask. As illustrated in Figure 1, two stage process of generating saliency map incorporates to one-stage nonlinear combination process by applying the neural network. Three parameters of (w, θ, B) were weighed after training the network. The weights of the network parameters are in direction of highlighting the target in the final saliency map. We can use these obtained weights for the application of object detection in cluttered scenes. These weights effectively indicate the influence of each feature map to find our target in the final saliency maps. The trained network is capable of finding the target (emergency triangle) in the test images containing the emergency triangle with arbitrary background. After entering a test image to our proposed model, first, the bottom-up cues are derived and then will be sent to the network. Here, the network acts such as higher brain area which compensate the lack of information about the target.

6 EXPERIMENTAL RESULTS

6.1 Database and Parameters

Image datasets are from Itti's lab at USC. This dataset consists of 64 images of emergency triangle with natural environment background. The 32 images with available target masks are applied as training dataset. Another 32 images of emergency triangle without target mask are utilized as the testing dataset.

The results presented here are based on the batch mode training using the *RPROP* algorithm (Riedmiller and Braun, 1992), which is a fast second-order gradient method. The RPROP algorithm parameters were as follows: $\eta 0 = 0.0001$, $\eta- = 0.5$, $\eta+ = 1.2$, $\Delta_{min} = 10-8$, $\Delta_{max} = 50$. We

initialized bias, weight and interaction-weight parameters with a normal distribution with zero mean and the variance was inversely proportional to the node fan-in. It should be noted that the multiplicative weight parameter w, is multiplied with each image. But there is also an interaction-weight parameter θ which allows the interaction between different pixels of a feature map. The network was selected with 42 inputs of 1200-dimentional, a hidden layer with 6 nodes and one output of 1200-dimentional in the output layer. We trained the network for 4000 epochs when the error did not change any more. After training, (42×6) weights (w) and (42×6) interaction-weights (θ) parameters were obtained which connect the 42 inputs to the 6 neurons of hidden layer. In addition, (6×1) weights (w) and (6×1) interaction-weights parameters were generated which connect the 6 neurons of hidden layer to the one neuron of the output layer. The value of these parameters can represent the influence of each feature map on the final saliency map.

6.2 The Results of Implementing the Algorithm

As could be observed in the Figure 2, two samples of 32 images of test dataset are illustrated on the left side and their corresponding saliency map is shown on the right side. In our strategy, each saliency map is generated as the *output of the network* after nonlinear fusion of 42 feature maps by the *trained network*. As shown in Figure 2 for two images, our model founded emergency triangle in the first hit in the 28 images of the 32 test images which is a remarkable result. The model could not find the target in the first hit just in four images that are shown in Figure 3. As shown in Figure 3, the target is detected in the fifth, second, third and sixth hit in the Figures.3.a, b, c, and d, respectively.

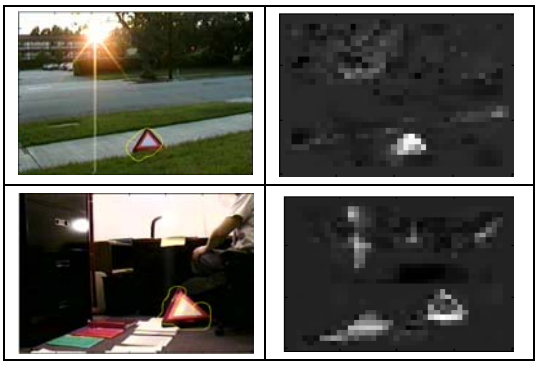

Figure 2: Left: Two samples of test images; Right: saliency maps corresponding to the left column images using our proposed method.

6.3 Comparison of Two Models

We have proved the effectiveness of our model with comparing our *nonlinear fusion approach using neural network* with the *simple superposition* in the basic saliency-based visual attention model with the same test dataset. As could be seen in Table 1, we defined four parameters of *No of FHD*, *No of UST*, *Mean* and *Standard Deviation* (*STD*) for comparing the two mentioned models. The number of first hit detection (*No of FHD*) demonstrates the number of trials in which the model could detect the target in the first hit without any mistakes. The number of unsuccessful trials (*No of UST*) is the number of trials which the model could not detect the target before five hit .After distinguishing the first point, the next salient point are detected by Inhibition of Return (IOR) process. As illustrated in Table 1, our method detects the target in 28 images of 32 test images. Our model could not detect the target before five executions just in one image which is shown in Figure 3.d. So, we had one unsuccessful trail (UST) based on our parameter definition. However, in Itti's model we found five unsuccessful (UST) trails. Moreover, two parameters of *Mean* and *STD* were employed to compare two models. These parameters demonstrate the *average mean* and *standard deviation* of false detections before finding the target in 32 images of test dataset. It is completely evident that our nonlinear fusion method has remarkable improvement in comparison with the basic saliency model. On top of that, the value of weights after training could be employed for the detection and recognition of emergency triangle in every arbitrary image. Another point is that, the network could be trained for any arbitrary dataset. When our images are more complicated than this database, we may need more number of training images to train the network. Although the training process is complicated and time consuming, after training and obtaining the efficient weights, these weights can be utilized to obtain the target. Furthermore, for other training patterns and other target, the training parameters and the number of neurons in hidden layer could be changed.

Table 1: Comparison of two models.

Fusion methods	No. of FHD	No. of UST	Mean	STD
Simple superposition	9	5	2.18	2.32
neural network	28	1	0.34	1.09

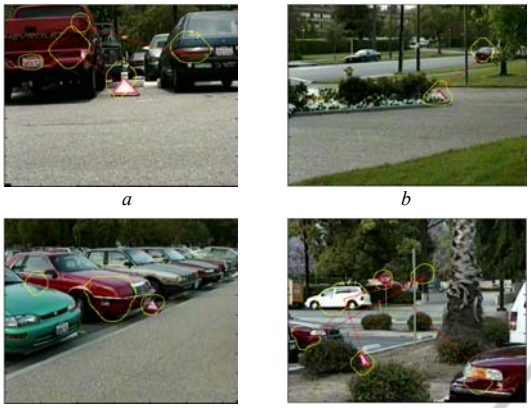

Figure 3: The images in which the model could not find the target in the first hit.

7 CONCLUSIONS

The proposed combinational structure applies both bottom-up and top-down information to detect the pre-learned objects. After extracting bottom-up features by Itti's model, they then will be sent to the variadic neural network as its input. The supervised variadic network was trained through several training images with their corresponding target masks and suitable parameters are weighed purposefully to highlight the target. The amounts of parameters indicate the influence of each feature map in finding the target on the final saliency map. Hence, the neural network can simultaneously weigh the feature maps and fuse them nonlinearly which is more convincing biologically. The noticeable improvement in first hit detection was achieved which is desirable for object detection purpose. As a future work, we wish to implement our model with other database and also propose a model for multiple object detection.

REFERENCES

Bahmani, H., Nasrabadi, A. M., and Hashemi Gholpayegani, M. R. (2008). Nonlinear data fusion in saliency-based visual attention, *4th International IEEE Conference in Intelligent System*.

Frintrop, S. (2006). VOCUS: A Visual Attention System for Object Detection and Goal-directed Search, LNAI, 3899, Springer Berlin/Heidelberg. ISBN: 3-540-32759-2.

Itti, L. and Koch, C. (2001). Computational modeling of visual attention, *Nature Reviews Neuroscience*, 2(3), 194–203.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis, *IEEE Trans PAMI*, 20(11), 1254–1259.

Itti, L., and Koch, C. (2001). Feature combination strategies for saliency-based visual attention systems, *Journal of Electronic Imaging*, 10(1), 161–169.

Kouchaki, Z., Nasrabadi, A. M., Maghooli, K. (2011). An improved model of saliency-based visual attention using fuzzy interface system, The *6th International conference on computer science and convergence information technology*, 204-207.

McGregor, S. (2007). Neural network processing for multi set data". In *Proc. Artificial neural networks — ICANN2007, 17th international conference*, 4668, 460–470.

Navalpakam, V., Rebesco, J. and Itti, L. (2005). Modeling the influence of task on attention, *Vision Research*, 45(2), 205–231.

Riedmiller M. and Braun, H. (1992). RPROP — A fast adaptive learning algorithm, technical report, *Universitat Karlsruhe*.

Sun, Y. and Fisher, R. (2003). Object-based visual attention for computer vision, *Artificial Intelligence*, 146(1), 77–123.

Treisman, A. M. and Gelade, G. A. (1980). A feature integration theory of attention, *Cognitive Psychology*, 12, 97–136.

Wolfe, J. (1994). A revised model of visual search, *Psyconomic Bulletin Review*, 1(2), 202–238.

Walter, D. Itti, L., Riesenhuber, M., Poggio T., and koch, C. (2002). Attentional selection for object recognition-A gentle way, *LNCS*, 25, 472–279

Walther, D. (2006). Interactions of visual attention and object recognition, *PhD Thesis, California Institute of Technology*.