A Biologically-Inspired Approach for Object Search

Mohammad Saifullah

***Abstract*— In this paper a biologically-inspired approach for object search is introduced. This approach is based on the visual information processing in the human brain and more specifically along the two visual processing pathways of the visual cortex. According to this approach different processes, with similar representational structure, work in parallel toward their local tasks, while at the same time, their mutual interaction leads to achievement of larger global goals. The model based on this approach provides a platform where bottom-up and top-down cues are computed and integrated in small incremental steps and lead to emergence of attention that selects an appropriate object. The two important principles of visual information processing, i.e., constraint satisfaction and inhibition play the key role in this model. The model is implemented with an interactive neural network. Simulation results demonstrate the practicality as well as the strength of this approach for object search tasks.**

***Index Terms*— Biologically-Inspired Approach, Visual Search, Visual Attention, Context, Neural Network**

I. Introduction

Object search in cluttered scenes is a difficult and challenging problem. The root of this problem lies in the fact that the required object may appear at different locations within the image among all sorts of similar or dissimilar distractors. Traditionally, computer vision approaches use a sliding window method to locate an object in a given image. In this method the whole image of a scene is scanned at all possible positions and scales to locate the object of interest. Though it is a useful technique, in terms of computational resources, the technique is usually very expensive. On the other hand, humans avoid, as far as possible, this brute-search strategy by employing the inherent mechanism of selective attention, and filter out the most salient and task-relevant objects in cluttered scenes.

In humans, visual attention is considered to be a two-step process [1][2]. In the first step, called bottom-up attention, a saliency map is generated by using bottom-up image-based cues, where a saliency map represents the most visually salient parts of the scene. Bottom-up attention is involuntarily and is initiated by simple features, such as, color, contrast, illumination and motion, etc. In the second step, top-down task specific cues are used to modulate the bottom-up saliency in favor of the most task specific parts of the scene. Top-down attention, unlike bottom-up attention, is initiated voluntarily by the cognitive areas of the brain. These two processes work in parallel and their interaction result in selection (focus of attention) of task specific parts of the input for further processing.

The visual information processing in humans can easily be understood in terms of the two stream hypothesis [3]. These streams are the ventral stream or 'What' pathway and the dorsal stream or 'Where' pathway. The ventral stream is responsible for recognizing and identifying objects by processing their visual properties, such as color and shape. The dorsal deals with the visual-motor control over the objects by processing object position, size and motion. These two pathways interact with each other and one consequence of this interaction lead to, among other things, attention on a specific location within input.

A number of computational models of attention [4][5][6][7] has been presented. Most of them are inspired by Koch and Ulman's computational architecture for pre-attentive attention to model the bottom-up mediated attention guidance. Despite the strong evidences of top-down modulation of visual processing in the human brain, there are only a few computational models [8][9][10][11] that take into account top-down information processing in the biological vision for solving complex task dependent search problems.

An important cue for task based search comes from the context of the object itself. In the real world there is a strong relationship between objects and their surroundings. Experimental studies [12][13] have also shown that humans use contextual associations of objects for performing detection and recognition more efficiently. Computational models of attention have demonstrated that context play a significant role in guiding the attention towards the most probable locations and thus improve efficiency of the search task [14][15].

In most previous work, the computation for attention is performed in standalone discrete steps. Such that different cues are calculated separately and at the end these cues are combined with some scheme to select the required object or its location. We believe that though quite useful, in terms of specific search results, these approaches lack the true spirit of the biological information processing. And, consequently lacks the robustness in performance, as well as flexibility in integrating different information processing modules, that is required for dealing with different situations.

In this paper we will present a biologically-inspired model of context-based attention for object search. In this model focus of attention will emerge as an outcome of interaction between bottom-up and top-down cues. The most important aspect of this model is an integrated and adaptable computational framework that enables different processing modules to interact with each other at different levels.

Manuscript received April 10, 2012; revised April 14, 2012.

Mohammad Saifullah is with the Department of Computer and Information Science, Linkoping University, SE-58183 Linkoping, Sweden (e-mail: mohammad.saifullahliu.se).

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

The work in this paper is related to earlier biologically-based artificial neural network approaches, such as MAGIC [16], developed by Behrmann and colleagues, which uses bidirectional connections to model grouping of features into objects within the ventral pathway. SLAM presented by Phaf et al., models the earliest stages of visual processing. Hamker [10] uses top-down connections in a recurrent network to mediate task information and thereby influence the control of eye movements. In the same way, Tsotsos et al., [17] developed a partially recurrent network for controlling eye movement. Sun and Fisher [7] also developed a model for object based attention for eye movement control.

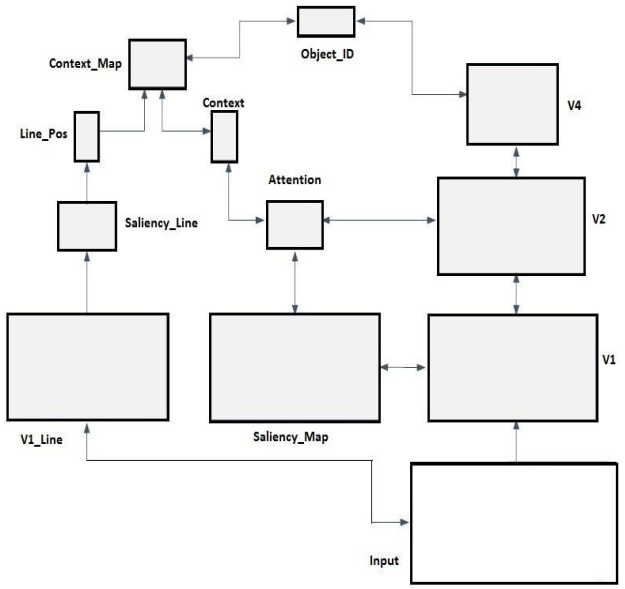

Fig. 1. The model for object search.

II. Our Approach

Our approach in this work is inspired by the modular and parallel processing architecture of the human brain. More specifically, we model visual information processing along the two pathways of the human visual system. According to our understanding of the brain, there are different individual processes that go on in parallel having different local goals, but at the same time these processes interact with each other for achieving a global goal. Moreover, these processes must have a representational structure which facilitates interaction at all necessary levels. This information processing paradigm of the brain requires information flow not only in bottom-up or top-down direction but all possible directions among different modules of the system. In this work this kind of omnidirectional processing is realized by developing a fully recurrent neural network model. This kind of omnidirectional processing has an obvious risk of uncontrolled avalanche effect of unit activations, and needs to be controlled by some kind of inhibition.

We have developed a model of attention, related to the earlier work of [18][19], that considers attention as a natural consequence of interaction between two visual pathways. A context module is added that interactively encodes the contextual cues for objects by interacting with the two pathways. During search the context module finds cues and facilitates the search by biasing the focus of attention towards the most probable location for the objects' presence.

III. The Model

Basically, the model is composed of two separate but interacting modules, namely: the Object Recognition module and the Spatial module. These modules perform a constraint satisfaction style of bidirectional processing of the input. Information oscillates many times between parts of the same module as well as between the two modules, at appropriate levels, before reaching a stable state.

A. The Object Recognition Module

The Object recognition module mimics the role of the ventral pathway and encodes scale, position and size invariant representations of the input objects along its hierarchical structure [20]. Local features are extracted at the lowest level of the hierarchy and more complex, invariant representations, are developed at levels higher-up. As this module encodes identity of the objects, top-down object based, or feature based, modulation is performed along this pathway.

B. The Spatial Module

The Spatial module plays the role of the dorsal pathway of the human visual system and encodes spatial information of objects. This module is divided into two sub modules; the Object sub-module and the Context sub-module.

C. The Saliency Sub-module

The Saliency sub-module is sensitive to positions of objects in the input scene. The layers in this sub-module register different locations of the objects by virtue of their salient property. This module interacts with the object module to pop-up the most salient and task relevant object.

D. The Context Sub-module

The architecture of the Context module is based on our premise that the human brain devises different strategies to deal with different situations. The basic processing principals remain the same, i.e., the Ventral part deals with identity and the dorsal part take care of the position of the object, but the strategy changes from situation to situation and person to person. Moreover, the computing framework has the flexibility to integrate different strategies for achieving the main computational goal. For example, in case of simply encoding the position of the object within a given scene, with respect to the input image's frame, the context module just perform a mapping between the location and identity of a given object that is used to search the same object in the future. In just another scenario, where position of the object depends on the position of another object or an object tends to appear at a position relevant to another object, a somewhat different strategy is devised. For example, when we search for an aeroplane, we almost always lift our head and search for it above the horizon. Our action shrinks the search area but require an estimate of the skyline or the line above where the probability of finding the aeroplane is much higher than at other places. This strategy is based on previous experiences that are encoded in the brain as an association between aeroplanes and their possible locations.

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

IV. Task Scenario for the Model

In order to demonstrate the feasibility of the model and to analyze its internal dynamics, a simple search task was designed. The choice of the task was made with the aim that it should facilitate evaluating information processing in different modules of the network, and the interaction between them.

The task is to search for an object with the help of a cue. For example, in the case of a flying aeroplane the skyline provides an important cue. A flying aeroplane has the highest probability to be found above the skyline, somewhere in the sky. Likewise, the probability of locating a far distance animal, human or vehicle etc. is highest at or around the skyline. In order to perform these search tasks efficiently, humans need to locate the skyline and start to search an appropriate area with reference to the skyline.
For simulation purpose, the task was simplified by taking three simple objects and a horizontal line as cue. One of the three objects has the tendency to appear at the same location where the horizontal line appears, while the second object has an opposite tendency. The third object has no constraints and can appear anywhere in the input.

V. Neural Network for Simulation

A. Network Architecture

A network based on the model of attention described above is developed for the task of context-based visual search (Figure 1). This network is a combination of three sub-networks. Each of the three sub-networks implements one of the modules of the model, and are named after these modules, i.e., the Object Recognition network and the Spatial network. The Spatial network can be further subdivided into the Saliency network and the Context network. The Object recognition network is a bidirectionally-connected hierarchical network, composed of five layers: Input, V1, V2, V4 and Object_ID, with layer sizes 114x114, 54x54, 54x54, 21x21 and 3x1 respectively. The units in layers V1 and V2 are divided into groups of 1x4 and 9x9 units respectively. Each unit within the same group in V1 was looking at the same spatial part in the image, that is, all units within a group had the same receptive field. Similarly, all units within the same group in V2 received input from the same four groups in V1. These sending groups in V1 were adjacent to each other and covered a contiguous patch of the visual image. Object_ID is an output layer and its size depends on the number of categories used for simulations. The Input layer serves as feeding input to the network, V1, V2 and V4 are hidden layers and Correct_ID is an output layer for the Object recognition network. Layers V2, V4 and Object_ID are bidirectionally connected in the hierarchy, while Input, V1 and V2 are connected in a feed-forward fashion.

The Saliency network contains two layers; the Saliencey_map, and the Attention layer. These layers are bidirectionally connected to each other, using all-to-all connections. The Saliency_map layer identifies the salient locations within the input and the Attention layer selects the most salient location from these. The Object Recognition and Saliency networks are bidirectionally connected by connecting the Saliency_map and the Attention layers to the V1 and V2 layers respectively. All these layers are connected to each other in a topographic manner.

The Context network is composed of five layers, namely, V1_Line, Saliency_Line, Line_Position, Context_Map and Context layers, having the sizes, 114x114, 3x3, 3x1, 8x8 and 3x3 units respectively. The V1_Line layer draws its input directly from the Input layer. Ideally the network should get its input from the V1 layer, but for implementation convenience a separate layer, V1_Line, is created. The layers of this network are connected in a feed forward and topographic manner with each other. The Context and Saliency networks interact via the Context and Attention layers. The Context_Map layer is connected to the Object_ID layer of the Object Recognition network.

It should be noted that an inhibitory mechanism, both at the group level as well as at the layer level is used to control the activation dynamics in the network. For this purpose, a k-Winners-Take-All (kWTA) like mechanism [18] is implemented that allow a specific number of units, in order of their decreasing strength, to be active at a given time.

B. Network Algorithm

The network was developed in Emergent [21], using the biologically plausible algorithm Leabra[18]. Each unit of the network had a sigmoid-like activation function:

$$y_j = \frac{\gamma[V_m - \Theta]_+}{\gamma[V_m - \Theta]_+ + 1}, \qquad [z]_+ = \begin{cases} z \text{ if } z \geq 0 \\ 0 \text{ if } z < 0 \end{cases} \tag{1}$$

Where

γ = gain, V_m = membrane potential and Θ = firing threshold.

Learning was based on a combination of Conditional Principal Component Analysis (CPCA), which is a Hebbian learning algorithm and Contrastive Hebbian learning (CHL), which is a biologically-based alternative to back propagation of error, applicable to bidirectional networks [18]:

$$\text{CPCA: } \Delta_{\text{hebb.}} = \varepsilon y_j (x_i - w_{ij}) \tag{2}$$

Where

ε = learning rate, x_i = activation of sending unit i, y_j = activation of receiving unit , w_{ij} = weight from unit i to unit j $\in [0, 1]$.

$$\text{CHL: } \Delta_{err} = \varepsilon(x_i^+ y_j^+ - x_i^- y_j^-) \tag{3}$$

Where

x^-, y^- = act when only input is clamped, x^+, y^+ = act when also output is clamped.

$$\text{L\_mix: } \Delta w_{ij} = \varepsilon[\, c_{\text{hebb}} \Delta_{hebb} + (1 - c_{\text{hebb}}) \Delta_{err}] \tag{4}$$

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Where

c_{hebb} = proportion of Hebbian learning

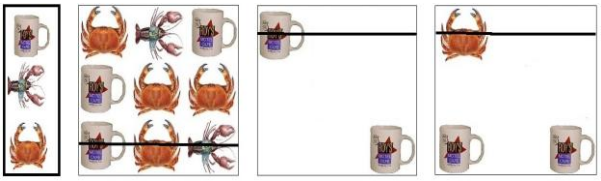

Fig. 2. Data for training and testing. The first box shows the three object categories used for training. The next three boxes show the stimuli used for testing.

VI. Data Set And Training

To demonstrate the feasibility of the approach we took three object categories (Figure 2) from the Caltech-101 data set [22]. For each object category, three images were selected. Each object image was converted to gray scale before detecting the edges in the image. As these images contain little clutter and the objects cover almost the whole image, all edges belong to a single object. Each image was resized to 30 x 30 pixels. The size of the input to the network is 114x114. The object size is approximately one ninth of the actual input size, so that each object could appear in one of nine locations in the input.

The training of the two networks, i.e., the Object recognition network and the Spatial network was performed separately. The Object Recognition network learns shape representations of input objects. For training this network, each object was presented to the network at all nine locations within the input image, one location at a time, so that network could learn the appearance of the objects in a position/location invariant manner.

The Spatial network estimates the position of the object of interest and cues from the input images and learns mapping between those. Only three layers of this network take part in learning, these layers are the Line_Position, the Context_Map and the Context layer. These three layers with the help of the Object_ID layer learn the mapping between the position of the line and the location of an object. Learning the two networks is performed by a combination of Hebbian and error-driven learning algorithms.

VII. Results and Discussion

After training, the network was tested on the task of object search with the help of a cue (in this task a horizontal line). The Attention layer of the network was considered as the output of the network for the search task, as it combines all information from the different parts of the network and selects the object to be processed for ultimate recognition along the ventral pathway.

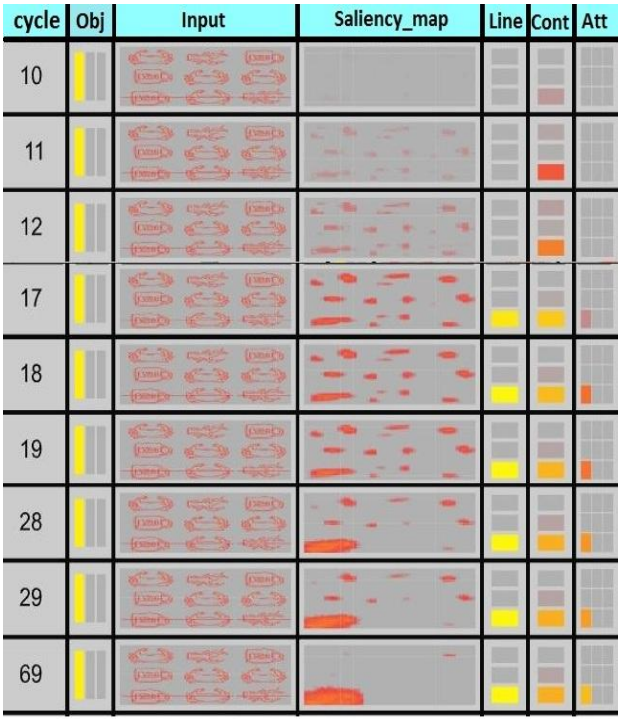

Fig. 3. Consecutive snapshots of activations in the various network layers (each layer is made up of a matrix of units, and the activation values of these matrices are shown here). The recorded changes in activation for different processing cycles illustrates how context-based focus of attention emerges as a result of interactions among layers of different modules. For each graph, in the order from left to right, the columns represent: Number of processing cycles (how far computation of activation has gone), activations in the Object_ID layer, Input layer, Saliency_Map layer, Line_Position layer, Context layer and Attention layer of the network. All these layers are topographically related to each other. Yellow (light) colours denote high activation values and red (dark) colours low activation. Gray (neutral) colour means no activation.

A. *Simulation #1 - Context Help Object Search*

In this stimulus (Figure 3) all of the three objects (cup, crab and crayfish) and a cue were used. Objects appear at all nine locations in the input. At some locations, objects are out of context while at other they appear according to their context. For example, the cup in the last row of the stimulus is according to context as it appears with the line while the cup in the second row is out of context. In this simulation, the task was to find the cup in the input. The network was required to locate the object cup in the input by injecting object identity information from the top via the Object_ID layer. There are two cups in the stimulus but due to the contextual cue the object should be found in the top row. As there are three objects in the top row in this task, it requires feature based modulation to locate the cup. Figure 3 shows a graphical description of the activations in the different layers of the network while solving this task. In the early cycles of processing (cycle: 10) the Salincy_map layer shows no activations, but as soon as it gets input from the V1 layer it begins encoding salient regions in the input. Parallel to this processing, the Context network detects lines (Cycle: 17) and sends signals to the Hidden layer of the network. The Hidden layer combines the input from the Object_ID layer and the Line_Pos layer and activates the most probable location at the Context layer. In parallel, the Attention layer interacts with the Saliency_map layer and V2 and activates the proper unit at the Attention layer.

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Fig. 4. Cycle-wise activations of various network layers for simulation#2.

B. *Simulation #2 - Object with Context Gets Priority*

The stimulus used for this simulation (Figure 4) is composed of two similar objects, e.g., two cups, and a cue. This simulation demonstrates how context influences attention and thereby prioritize a specific location. The task for the network is to find the cup. Since there are two cups in the stimulus it is interesting to see which location that is selected by the network. Consider Figure 4. In the early cycles of this simulation (Cycle:15) the Saliency_map layer shows both of the objects. But as soon as the cue is detected (Cyle: 17) and signals are propagated to the network, the Context layer activates the most probable unit on the basis of prior experience. The attention layer receives signals of almost the same strength from the Saliency_map and the V2 layers for the two objects. But, the Context layer modulates the Attention layer in favour of the cued object. The end result is that the Attention layer selects the cued object.

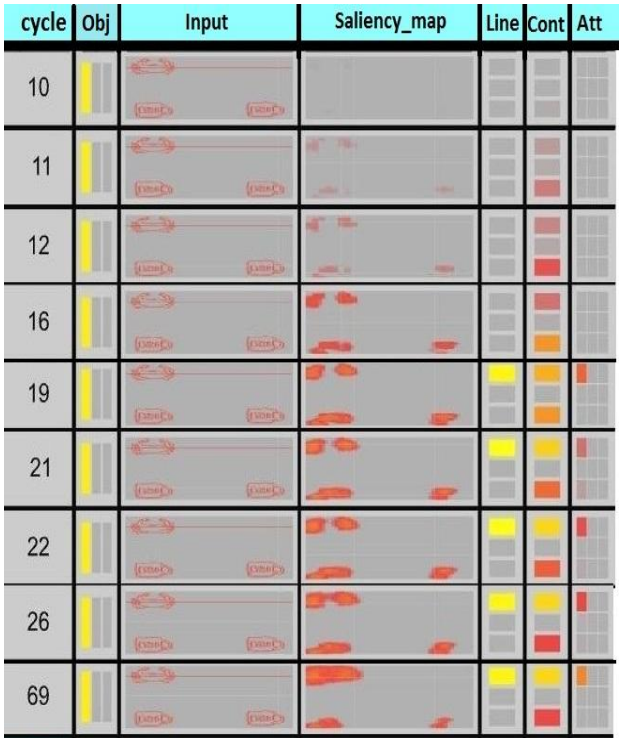

Fig. 5. Cycle-wise activations of various network layers for simulation#3.

C. *Simulation #3 - If Objects Appear Out of Context*

This stimulus in this simulation contains three objects (two cups and a crab) and a cue (Figure 5). All three objects were presented out of context as during training the cup always appears with the line and the crab at other locations. Now it can be seen from Figure 4, that the network was required to locate the object cup in the input. In this situation if there would be no cue the most usual response from the network should be to focus attention on any of the two object cups, because of the top-down feature based modulation in the ventral network. But, the presence of the line as a cue makes a difference here (Cycle: 16-69). The line provides contextual cues for the location of the required object and biases the network towards a few locations in the input (Cycle 23-69) which are more probable than the others for an efficient search. The result is that the position of the crab is selected. Though it is not the correct location but it is selected due to the strong contextual cue.

VIII. Conclusion

In this paper a biologically-inspired approach for object search was presented. The model used on this approach selects the most probable location for a given object by simulating the phenomenon of attention. The attention focus in this model emerges as an interaction between the two information processing pathways and, interaction between top-down expectations and bottom-up visual information under the guiding principles of constraint satisfaction and inhibitory competition. The training and testing of the model demonstrate its practicality.

References

[1] J. Braun and B. Julesz, "Withdrawing attention at little or no cost: detection and discrimination tasks," *Attention, Perception, & Psychophysics*, vol. 60, no. 1, pp. 1–23, 1998.

[2] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[3] L. G. Ungerleider and J. V. Haxby, "[] What'and [] where'in the human brain," *Current opinion in neurobiology*, vol. 4, no. 2, pp. 157–165, 1994.

[4] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 11, pp. 1254–1259, 1998.

[5] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry.," *Hum Neurobiol*, vol. 4, no. 4, pp. 219–27, 1985.

[6] G. Backer, B. Mertsching, and M. Bollmann, "Data-and model-driven gaze control for an active-vision system," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1415–1429, 2001.

[7] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artificial Intelligence*, vol. 146, no. 1, pp. 77–123, 2003.

[8] J. M. Wolfe, "Guided search 2.0 A revised model of visual search," *Psychonomic bulletin & review*, vol. 1, no. 2, pp. 202–238, 1994.

[9] J. K. Tsotsos, S. M. Culhane, W. Y. Kei Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial intelligence*, vol. 78, no. 1, pp.

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

507–545, 1995.

[10] F. H. Hamker, “Modeling attention: From computational neuroscience to computer vision,” *Attention and Performance in Computational Vision*, pp. 118–132, 2005.

[11] V. Navalpakkam and L. Itti, “Modeling the influence of task on attention,” *Vision Research*, vol. 45, no. 2, pp. 205–231, 2005.

[12] E. Palmer, “The effects of contextual scenes on the identification of objects,” *Memory & Cognition*, vol. 3, no. 5, pp. 519–526, 1975.

[13] M. Bar, “Visual objects in context,” *Nature Reviews Neuroscience*, vol. 5, no. 8, pp. 617–629, 2004.

[14] K. Murphy, A. Torralba, and W. Freeman, “Using the forest to see the trees: a graphical model relating features, objects and scenes,” *Advances in neural information processing systems*, vol. 16, 2003.

[15] T. Poggio, S. S. Chikkerur, and others, “What and where: a Bayesian inference theory of visual attention,” Massachusetts Institute of Technology, 2010.

[16] M. Behrmann, R. S. Zemel, and M. C. Mozer, “Object-based attention and occlusion: Evidence from normal participants and a computational model.,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 24, no. 4, p. 1011, 1998.

[17] Y. Sun, R. Fisher, F. Wang, and H. M. Gomes, “A computer vision model for visual-object-based attention and eye movements,” *Computer Vision and Image Understanding*, vol. 112, no. 2, pp. 126–142, 2008.

[18] R. C. O'Reilly and Y. Munakata, *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. The MIT Press, 2000.

[19] M. Saifullah and R. Kovordányi, “Emergence of Attention Focus in a Biologically-Based Bidirectionally-Connected Hierarchical Network,” *Adaptive and Natural Computing Algorithms*, pp. 200–209, 2011.

[20] M. Saifullah, “Exploring Biologically-Inspired Interactive Networks for Object Recognition,” 2011.

[21] B. Aisa, B. Mingus, and R. O'Reilly, “The emergent neural modeling system,” *Neural networks*, vol. 21, no. 8, pp. 1146–1152, 2008.

[22] L. Fei-Fei, R. Fergus, and P. Perona, “Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories,” *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)