

# Emergence of Attention Focus in a Biologically-Based Bidirectionally-Connected Hierarchical Network

Mohammad Saifullah, Rita Kovordányi

Department of Information and Computer Science, Linköping University  
Linköping, Sweden  
{mohammad.saifullah, rita.kovordanyi}@liu.se

**Abstract:** We present a computational model for visual processing where attentional focus emerges fundamental mechanisms inherent to human vision. Through detailed analysis of activation development in the network we demonstrate how normal interaction between top-down and bottom-up processing and intrinsic mutual competition within processing units can give rise to attentional focus. The model includes both spatial and object-based attention, which are computed simultaneously, and can mutually reinforce each other. We show how a non-salient location and a corresponding non-salient feature set that are at first weakly activated by visual input can be reinforced by top-down feedback signals (centrally controlled attention), and instigate a change in attentional focus to the weak object. One application of this model is highlight a task-relevant object in a cluttered visual environment, even when this object is non-salient (non-conspicuous).

**Keywords:** Spatial attention, Object-based attention, Biased competition, Recurrent bidirectionally connected networks

## 1 Introduction

Image processing techniques for object recognition can be made to learn relatively easily to recognize or classify a single object in the field of input. However, if more than one object is present simultaneously, it would be difficult to separate information about the target object from information about other objects in the field of input. In such cases, simpler techniques will fail to produce an output. Artificial neural techniques for image processing will on the other hand tend to produce an output that reflects a mixture of the objects, recognizing neither the target object, nor the clutter in the background. Neither solution is satisfactory.

In contrast, the human visual system can function without problem in the face of potentially irrelevant objects cluttering up the visual field. Multiple visual inputs are disambiguated via top-down, knowledge-driven signals that reflect previous memory for an object, present task requirements, or goals and intentions. Hence, on the basis of top-down signals, the human visual system can focus on those objects that are relevant to the task, while inhibiting irrelevant information. As the relevancy of information changes from environment to environment and task to task, attention has to work as an adaptive filter.

In order to adapt to visual saliency as well as task requirements, attention is controlled by bottom-up sensory cues as well as by top-down, task dependent information. In image processing, this process is often divided into a two-stage process. At the first stage, a saliency map is created based on the visual conspicuity or intensity of the input. At the second stage, attention is focused on the most salient region, and an attempt is made to recognize the relevant object in the focused region. As a further development, top-down image processing approaches allow for direct, simultaneous top-down influence of the saliency map through a feedback-loop, so that the position of the most salient region is determined not only on the basis of saliency, but also whether the region contains features that (could) belong to the sought object. For example, Lee and colleagues use a feedback loop, where object-information is injected into the early stage of spatial processing stream [1]. Other notable examples include [2, 3].

Truly mutual interaction between bottom-up and top-down driven computation cannot be achieved unless computation is naturally implemented in small, incremental update steps, for example, using artificial neural networks (ANN) where unit activations are calculated incrementally.

In addition to this, biologically-based ANN-approaches capitalize on the fact that 1. The human visual system is intrinsically bidirectionally connected, which allows for mutual interaction between bottom-up and top-down information, and 2. There is inherent competition within groups of processing units (corresponding to ANN-layers or groups of artificial units), which indirectly creates an inhibitory effect where unwanted information outside the region of interest is suppressed [4-6].

Among the early biologically-based ANN-approaches is MAGIC, developed by Behrmann and colleagues, which uses bidirectional connections to model grouping of features into objects within the ventral what-pathway (cf. Fig. 1). MAGIC models feature-based object-driven selection, but does not combine this with selection of a spatial region of interest [7]. Hence, it is not possible to focus on a particular location in the image. Phaf and coworkers [8] present SLAM, which include spatial attention and models the earliest stages of visual processing. Subsequently, these biologically-based models were extended to eye movement control. For example, Hamker uses top-down connections in a recurrent (bidirectional) network to mediate task information and thereby influence the control eye movements [9]. Likewise, Tsotsos and coworkers present a partially recurrent network for eye movement control [10]. Sun and Fisher present a system where object-based and space-based attention work together to control eye movements [11].

## 2 Our approach

None of the above models utilize the potential of a fully recurrent network, where top-down feedback can naturally permeate bottom-up-driven computation. In the approach we have chosen, attentional focus arises as an emergent side-effect of normal visual computation. No extra mechanisms are required (for example, special feedback loops) to achieve an attentional focus with inhibitory fringe, as we use the normal visual feedback connections and the intrinsic competition mechanisms that are

present in the visual system and are required for normal learning and processing. Similar ideas of integrated or biased competition that requires no extra mechanisms were presented by Duncan and O'Reilly [4, 5].

We use a network (Fig. 1) where we model the interaction between the dorsal and the ventral pathway in the primate visual system using a bidirectionally-connected hierarchically-organized artificial neural network. We use this network to study the interaction between top-down and bottom-up information flow during the emergence of attention focus. The network architecture is based on O'Reilly's model of attention [5], where we re-modeled the dorsal pathway for spatial attention, allowed for object based top-down information flow along the ventral pathway, and let this information interact with the normal bottom-up flow of visual information.

In contrast to many image processing approaches, the proposed model does not consider spatial attention and the object recognition as two separate processing stages, rather they work in parallel and attention emerges as a natural consequence of interactions between top-down and bottom-up influences.

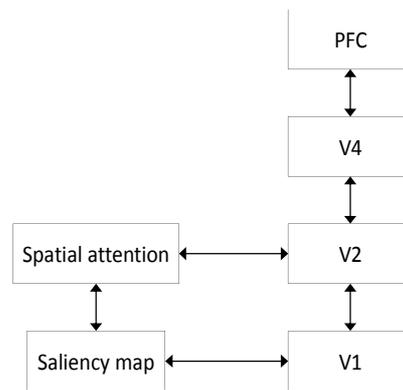


Fig. 1. A schematic diagram showing the interactions in the ventral what-pathway and the dorsal where-pathway, as well as cross-communication between the pathways.

As a result of mutual interaction between top-down and bottom-up signals, the region of interest and the object in focus emerges in parallel, mutually reinforcing each other. The bottom-up saliency of visual cues in the input image is computed in terms of the strongest activated area in the early stages of processing (V1, V2). The input-driven feedforward flow of information across the hierarchy of the network is modulated by top-down task information at each incremental update of all activations across the network. In this way, which object should be focused on modulates the bottom-up driven saliency map and helps to activate the most relevant region (in the where-pathway) and the relevant features of the task specific object (in the what-pathway). The interaction between top-down and bottom-up signals give rise to a focus which activates the most salient and task relevant location, as well as the corresponding object and its features (constituent line segments) at that location in the input image.

### 2.1. A closer look at the network used in the simulations

The network used in our simulations (Fig. 2) is a combination of two sub-networks: ‘what’ and ‘where’. The ‘what’ network models the ventral or ‘what’ pathway and composed of five layers: Input, V1, V2, V4 and Object\_cat, with layer sizes 60x60, 60x30, 33x33, 14x14 and 5x1 respectively.

The units in layers V1 and V2 were divided into groups of 2x4 and 8x8 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\_cat is an output layer and its size depends on the number of categories used for simulations. Layers V2, V4 and Object\_cat are bidirectionally connected in hierarchy, while Input, V1 and V2 are connected in bottom up fashion.

The ‘where’ network, which simulates the functionality of the dorsal pathway and mediates spatial information, is a two layer network. The network layers, Saliency\_map and Attention, 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. For the simulations, we combined the object recognition and spatial attention networks by bidirectionally connecting the Saliency\_map and Attention layers to V1 and V2 layers respectively.

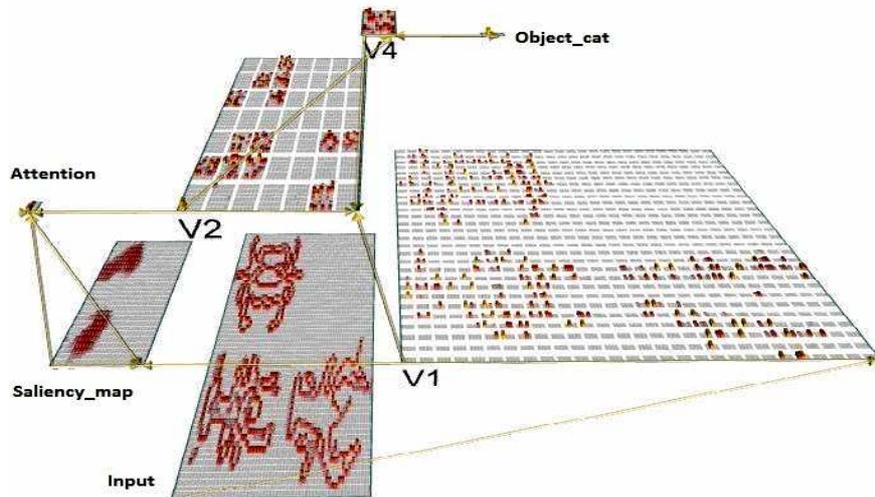


Fig. 2. Network used in the simulations. All connections were bidirectional except for connections going from Input to V1, and V1 to V2 and Saliency\_map). The connections are schematically displayed: Connections were either all-to-all, or were arranged in a tiled fashion, where each connection mediated information from a sending unit group to a receiving group).

The network was developed in Emergent [12], using the biological plausible algorithm Leabra [5]. Each unit of the network had a sigmoid-like activation function:

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

$\gamma$  = gain  
 $V_m$  = membrane potential  
 $\Theta$  = 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 [5]:

$$\begin{aligned} \text{CPCA: } \Delta_{\text{hebb}} &= \varepsilon y_j (x_i - w_{ij}) \\ \varepsilon &= \text{learning rate} \\ x_i &= \text{activation of sending unit } i \\ y_j &= \text{activation of receiving unit } j \\ w_{ij} &= \text{weight from unit } i \text{ to unit } j \in [0, 1] \\ \\ \text{CHL: } \Delta_{\text{err}} &= \varepsilon (x_i^+ y_j^+ - x_i^- y_j^-) \\ x_i^-, y_j^- &= \text{act when only input is clamped} \\ x_i^+, y_j^+ &= \text{act when also output is clamped} \\ \\ \text{L\_mix: } \Delta w_{ij} &= \varepsilon [c_{\text{hebb}} \Delta_{\text{hebb}} + (1 - c_{\text{hebb}}) \Delta_{\text{err}}] \\ c_{\text{hebb}} &= \text{proportion of Hebbian learning} \end{aligned} \quad (2)$$

We used a relatively low amount of Hebbian learning, 0.1% of the total L\_mix (Eq. 2), for all connections in the two networks except for the connections from Input to V1. For Input to V1, Gabor filters were used to extract oriented bar like feature for four orientations. These values for the Hebbian learning, as well as other parameters, are based on previous work [13, 14].

## 2.2. Data set

For this study we took five object categories from the Caltech-101 data set [15]. For each object category, three images were selected. Each object image was converted to gray scale before detecting the edges in the image. Each image was resized to 30 x 30 pixels (Fig. 3). The size of the input to the network is 60 x 60. This implies that object size is approximately one fourth of the network input size, so that each object could appear in one of four locations in the input (Fig. 4). During training each object was presented to the network at all four locations, one location at a time, so that network could learn the appearance of the object in a position invariant manner at all locations.

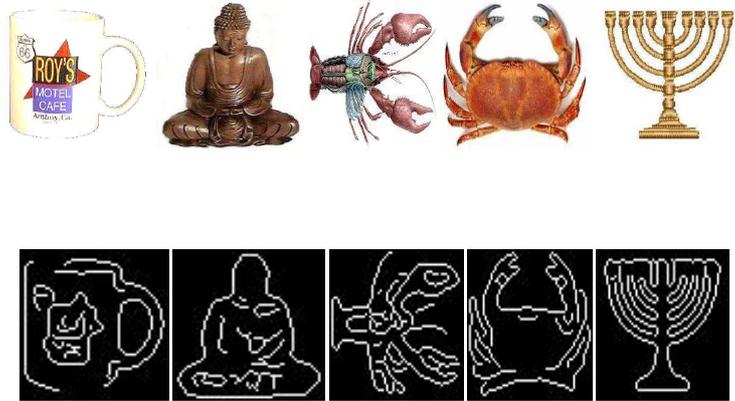


Fig. 4. Top row: Examples of the five object categories that were selected for training. Bottom row: Edge representation of the images.

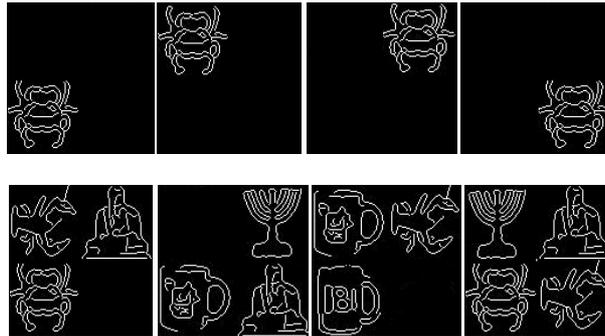


Fig. 3. Top: A few input images as illustration of the four positions where data were presented during training of the network. Object size was almost one fourth of the input size so that the objects could appear at one of four locations within the input image. Bottom: sample test images containing multiple objects in various locations.

### 2.3. Procedure

First of all, we trained the network on all five object categories. After training, we made sure by testing that the network has learnt all the object at all four locations. After that, we evaluated the network in three steps. In the first step, to get a baseline for how the network performs on several simultaneous objects, we fed to the network test images containing more than one objects at different locations. The network response was quite arbitrary and most of the time results in error.

In the next step, we connected the ‘where’ network with object recognition network to evaluate how spatial attention interacted with the object recognition pathway to

facilitate object recognition, when more than one objects were present in the input image. When input containing two objects was presented to the network, parallel processing of the input begin along two pathways. Object recognition pathway extracted features from the input and processing of features moves along its hierarchy. In the meanwhile, the Saliency\_map layer in the dorsal pathway generated a saliency map of the input and the next layer in the hierarchy, the Attention layer, selected one of the blob-shaped activation patches that were formed in the saliency map due to KWTA competition within the layer which allowed only 50% of the units to become activated. As the Attention layer was connected with the V4 layer in the 'what' network, its interaction with the V4 layer reinforced via feedback connections the activation of the corresponding V2 units at the location which was sharing connections with active unit of Attention layer. This interaction caused V2 unit at a particular location, which represent the activation of a particular object to be more active and thus reducing the activation of all other units at all other locations due to inhibition within V2 layer. Consequently, higher layers of the networks get feedback of the one object, which has higher activation due to spatial attention mechanism. The network correctly recognizes the object which got focus of spatial attention.

In the third step, we investigated that how interaction between top down, goal-directed and bottom-up image based affects lead to focusing of attention on a specific object (Fig. 5). For this purpose the role of the Object\_cat layer was changed from output to input, and Intention layer was set as output layer, in order to observe that at what location attention focuses on, as a result of interaction between top-down and bottom-up effects. For this simulations, input with three and four objects at a time, and at different locations within input, are presented to the network. The Saliency\_map layer indicated the salient regions in the input, and Attention layer selected the most salient region in Saliency\_map. In the meantime, top-down effects along the object recognition pathway strengthened the relevant features of the specific object category; category information was fed at Object\_cat in a top-down fashion, by interacting with the input/bottom-up activations along the same pathway. This results in the activations of the category specific units at all four locations at V2 layer. But, the location where object of the specific category was presented became more active comparative to other locations. The local inhibition also play its role and inhibit the activation of less active units, thereby most of the remaining active units belong to true location of the specific object category. As V2 layer is bidirectionally connected with the Attention layer, it interacts with the Attention layer and a sort of competition starts between the top-down effects through V2 layer and bottom-up effects through Saliency\_map. The active unit at Attention layer indicates the true location of attention of the focus of attention on the network.

### 3 Results and Analysis

#### 3.1. Multiple objects without any attention

We fed the train network with two, three and four objects at a time to observe network behavior. The network output was arbitrary, as it could not handle multiple objects. The multiple objects, at multiple locations activated units representing their features. But, due to KWTA, the most active units in all objects remain active. The representation at the higher layers contains features belonging to all objects fed at input. Now, it is the sheer chance that which category is decided by the network at output. It clearly exposes the network's inability to deal with multiple objects simultaneously.

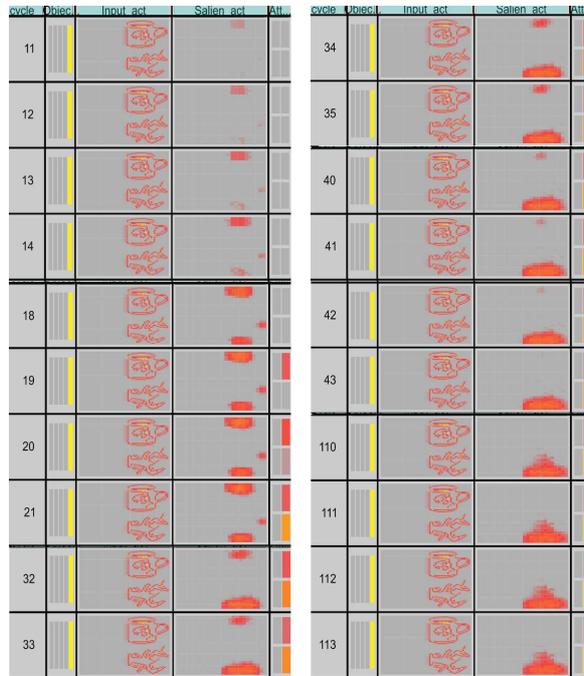


Fig. 5. Consecutive snapshots of activations in the various network layers (each layer is made up of a matrix of units, and the activation values of these matrices are shown here). The recorded changes in activation for different processing cycles illustrate how task-based focus of attention emerges as a result of top-down and bottom-up interactions. For each graph, in the order from left to right, the columns represent: Number of processing cycle (how far computation of activation has gone), activations in Object\_cat layer, Input layer, Saliency\_map layer, and Attention layer of the network. Yellow (light) colors denote high activation values, red (dark) colors low activation. Gray (neutral) color means no activation.

### 3.2. Multiple objects with bottom-up attentional effects

Bottom-up effect, along the dorsal pathway, removed the arbitrary behavior of the network in the presence of multiple objects as input. Network, begin to recognize the most salient object in the input field. This is because, the Saliency\_map generate a saliency map on the basis of input from V1. Due to KWTA, the activations of the strongest object, in terms of activation survive, while rest gets inhibited. The mutual interaction between Saliency\_map and Attention layer activates the location unit of the said object. The active unit in the Attention layer, through interaction, enhances the activations of units pertaining to location which represents the strongest object in the input. This process interpreted as the attentional focus on particular location due to dorsal pathway. If the proper, inhibition-of-return phenomenon is implemented, then the next focus of attention would move to the second strongest object and so on. But, if the objective is to look for a specific object, top-down effects are required.

### 3.3. Multiple objects with both bottom-up as well as top-down attentional effect

The interaction between the bottom-up and top-down effects, as well as between these two effects and feed forward flow of the image information along the ventral pathway lead to a more controlled network behavior. For example consider a single case, Figure. 4, an input containing two different objects, a cup and a crayfish, are presented at the input layer of the network. The activations produces by cup are stronger as compare to those of crayfish. Therefore, saliency map would select cup as an object to put attention on. But the Object\_cat layer, that has changed its role from output to input layer, is clamped with the pattern that represent crayfish. It will work as top-down effects are biasing the network for searching crayfish. Now from the figure 4, It is evident that initially Saliency\_map layer build a saliency map which shows the most salient object i.e. cup (cycle:11-14). And Attention layer, through interaction with Saliency\_map layer, activate the position of cup in Attention layer, in this case the right top unit in the Attention layer (cycle 19, 20). Meanwhile, top down effects interacts with the unit activation at the V4, V2 units and bias the layer activations towards crayfish specific features. This interaction leads to strengthening activity at location where crayfish was actually presented. This interaction in turn activates the location unit, representing the crayfish location, at Attention layer. It is the time, when a completion for winner takes place for locations, between bottom-up and top-down effects (cycles 21-42). In this case, this competition gives way to attention focus on crayfish (cycle 42-113).

## 4 Conclusions

We have presented a model where attentional focus arises as an emergent side-effect of mechanisms that are inherent to visual computation, namely interaction between top-down expectations and bottom-up visual information, and intrinsic lateral competition within processing modules (layers or groups of artificial units). In this

model, attentional focus arises from the mutual interaction between the top-down and bottom-up influences.

We have shown how these influences give rise to a step-by-step development of activations in the network that reflect attention focused on a specific location and at the same time on a specific object. We have demonstrated how top-down influence can override bottom-up saliency in some cases, so that an object that is at first only weakly activated based on visual cues, can be focused on intentionally (top-down), so that its activation is enhanced and finally emerges as a winner (gains focus). While top-down effects ensure that only task relevant objects and features get activated, their interaction with bottom-up effects helps delineate the object's features and ignore the clutter within the image.

## References

1. Lee, S., Kim, K., Kim, J., Kim, M., Yoo, H.: Familiarity based unified visual attention model for fast and robust object recognition. *Pattern Recognition*. 43, 1116-1128 (2010).
2. Poggio, T., Poggio, T., Serre, T., Tan, C., Chikkerur, S.: An integrated model of visual attention using shape-based features. MIT-CSAIL-TR-2009-029. (2009).
3. Navalpakkam, V., Itti, L.: Modeling the influence of task on attention. *Vision Research*. 45, 205-231 (2005).
4. Duncan, J.: Converging levels of analysis in the cognitive neuroscience of visual attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 353, 1307-1317 (1998).
5. O'Reilly, R.C., Munakata, Y.: *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. MIT Press (2000).
6. O'Reilly, R.C.: Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences*. 2, 455-462 (1998).
7. Behrmann, M., Zemel, R., Mozer, M.: Object-Based Attention and Occlusion: Evidence from Normal Participants and a Computational Model. *Journal of Experimental Psychology: Human Perception and Performance*. 24, 1011-1036 (1998).
8. Phaf, R., Van der Heijden, A., Hudson, P.: SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology*. 22, 273-341 (1990).
9. Hamker, F.H.: The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *Computer Vision and Image Understanding*. 100, 64-106 (2005).
10. Rothenstein, A.L., Rodríguez-Sánchez, A.J., Simine, E., Tsotsos, J.K.: Visual feature binding within the selective tuning attention framework. *International Journal of Pattern Recognition and Artificial Intelligence*. 22, 861 (2008).
11. Sun, Y., Fisher, R., Wang, F., Gomes, H.: A computer vision model for visual-object-based attention and eye movements. *Computer Vision and Image Understanding*. 112, 126-142 (2008).
12. Aisa, B., Mingus, B., O'Reilly, R.: The emergent neural modeling system. *Neural Networks: The Official Journal of the International Neural Network Society*. 21, 1146-52 (2008).
13. Kovordányi, R., Roy, C.: Cyclone track forecasting based on satellite images using artificial neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*. 64, 513-521 (2009).
14. Kovordányi, R., Saifullah, M., Roy, C.: Local feature extraction — What receptive field size should be used? Presented at the International Conference on Image Processing, Computer Vision and Pattern Recognition, Las Vegas, USA (2009).
15. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *IEEE. CVPR 2004. Workshop on Generative-Model Based Vision*. (2004).