



# Interactive Learning of Top-down Visual Attention Control and Motor Actions

Ali Borji<sup>1</sup> (borji@ipm.ir), Majid N. Ahmadabadi<sup>1,2</sup> (mnili@ut.ac.ir), Babak N. Araabi<sup>1,2</sup> (araabi@ut.ac.ir)

<sup>1</sup> School of Cognitive Sciences, Institute for Studies in Theoretical Physics and Mathematics (IPM), Tehran, IRAN

<sup>2</sup> School of Electrical and Computer Engineering, University of Tehran, Tehran, IRAN



## Overview

A biologically-motivated computational model for learning task-driven and object-based visual attention control in interactive environments is proposed. It consists of three layers:

First, in the early visual processing layer, basic layout and gist of the scene are extracted. The most salient location of the scene is simultaneously derived using the biased saliency-based bottom-up model of visual attention.

Then a cognitive component in the higher visual processing layer performs an application specific operation such as object recognition and scene understanding at the focus of attention. From this information, a state is derived in the decision making layer.

Top-down attention in our model is learned by the U-TREE algorithm which successively grows a tree whenever perceptual aliasing occurs. Internal nodes in this tree check the existence of a specific object in the scene and its leaves point to states in the Q-table. Motor actions are associated with leaves. After performing a motor action, the agent receives a reinforcement signal from the critic. This signal is alternately used for modifying the tree or updating the action selection policy.

A long-term memory component holds the bias signals of important task-relevant objects of the environment. Basic saliency-based model of visual attention is devised to consider processing cost of feature channels and image resolutions. For object recognition, a recent and successful object recognition method inspired by the hierarchical organization of the visual ventral stream is used.

The proposed model is evaluated over a visual navigation tasks for learning image to action mappings.

Both the most salient location and the overall sketch of the image are then transferred to the higher vision unit. For reducing the computational complexity, higher visual processes (like object recognition) are only targeted at the focus of attention (FOA). Next state of the agent is determined by the attention tree. Motor actions are associated with the leaves. Outcome of this action over the world is evaluated by a critic who is aware of the model of the environment and a reinforcement signal is fed back to the agent to update its internal representations (attention tree) and action selection strategy in a quasi-static manner.

## Early Visual Processing Layer

Basic saliency-based model of visual attention [1] is revised for the purpose of salient region selection at this layer (fig.2)

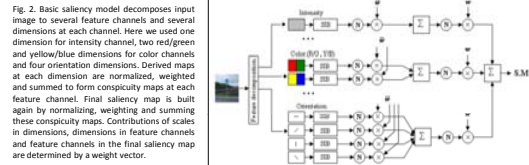
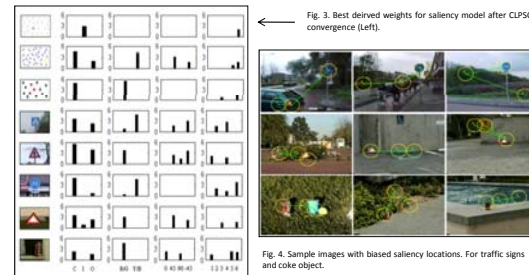


Fig. 2. Basic saliency model decomposes input image to several feature channels and several dimensions at each channel. Here we used one dimension for intensity channel, two red/green and yellow/blue dimensions for color channels and four orientation dimensions. Derived maps at each dimension are normalized, weighted and summed to form conspicuity maps at each feature channel. Final saliency map is built again by normalizing, weighting and summing these conspicuity maps. Contributions of scales in dimensions, dimensions in feature channels and feature channels in the final saliency map are determined by a weight vector.

Both feature channels and scales are associated with weights and processing costs. Then comprehensive learning particle swarm optimization (CLPSO) [2], is used in order to find weights which lead to maximum detection rate and minimum cost defined in the fitness function of equation 1.

$$Fitness(\vec{w}) = \frac{1}{m \times n} \sum_{j=1}^m \sum_{i=1}^n dist_j(Saliency(I_{img_i}, \vec{w}), t_i) \sum_{k=1}^n u_k(w_k, c_k) \quad (1)$$

m is the number of images in the train image set and n is the dimensionality of cost or feature vector.  $dist_j(\cdot)$  is the Euclidian distance between the j-th salient point generated by the saliency model and the target location  $t_i$  in the j-th image.  $u_k(\cdot)$  is the step function which is 1 when a feature channel or resolution is used. Saliency is a function which takes as input an image and a weight vector and outputs a vector of p salient points.



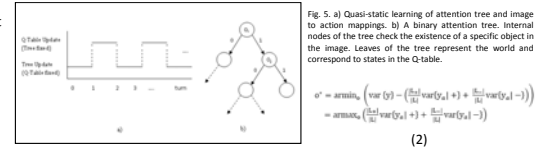
## Higher Visual Processing Layer

The object at the attended location is recognized by the hierarchical model of object recognition (HMAX) in the cortex [3, 4]. A binary SVM classifier [5], is trained with positive samples of a class and negative samples from other classes. Offline classifiers in this way are later used for online object recognition at the focus of attention. Using C2 features, 87% ( $\pm 2.2$ ), 93% ( $\pm 2.6$ ), 91.28% ( $\pm 2.8\%$ ), 94.6% ( $\pm 1.4\%$ ) and 83% ( $\pm 4.2\%$ ) recognition rates were achieved for traffic signs and the coke object respectively.

## Decision Making Layer

The core of our model is the decision making layer where visual attentions and representations are learned. This layer controls both top-down visual attention and motor actions. The learning approach has the same basics as the U-TREE algorithm [6].

Attention tree is incrementally built in a quasi-static manner in two phases (iterations). 1) RL-fixed phase and 2) Tree-fixed phase. In each phase of the algorithm external feedback of the critic (a scalar reward) is used to alternately update the policy or refine the leaves with aliasing. This process is illustrated in figure 5.



Algorithm starts with a single node in the RL-fixed phase and then moves to the Tree-Fix phase and so on. In each Tree-fixed phase, RL algorithm is executed for a number of episodes according to the learned policy from the previous phase by following  $\epsilon$ -greedy action selection strategy. In this phase, tree is hold fixed and the derived quadruples ( $S_t, a_t, r_{t+1}, S_{t+1}$ ) are only used for updating the Q-table. State discretization occurs in the RL-fixed (Tree-update) phase. In this phase, gathered experiences by the agent are used to refine leaves of the attention tree with aliasing which is measured by TD error. An object which minimizes variance the most is selected for braking an aliased leaf according to equation 2 (above).

## Results

The agent is supposed to learn how to navigate safely in a simulated navigation environment. It uses its offline learned knowledge interactively. Map of the route, consisting of 11 spots, is shown in figure 6. The agent captures 360 x 270 RGB color images. It is assumed that the agent knows which objects are in a scene when it observes an image. This information is only used for refining the aliased nodes and not for state determinations. The agent has three possible motor actions: Forward (F), Turn Left (L) and Turn Right (R) and can attend to one of n objects each time (n=5).

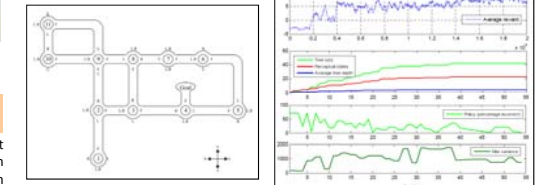
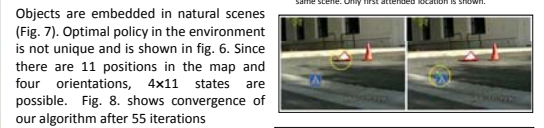


Fig. 6. Navigation map in the experiment. A subset of 5 objects is randomly embedded in each scene. Best actions are shown besides each state. In some states two actions are optimal.

There is no error in detection and recognition. Objects in the scenes (4 traffic signs plus coke object) are randomly speared in space and are not bound to specific spatial locations. Algorithm generated 7 states with average depth of 3. It means that instead of attending to five objects simultaneously, serial attention to 3 objects in average could solve the problem. (Fig.9)

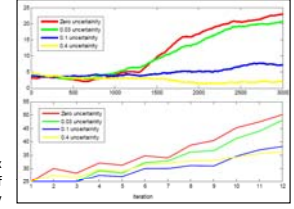
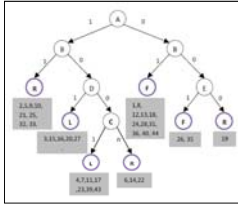


Fig. 9. Learned attention tree for the map of fig 6 with pruning. Forty four states were clustered into 7 leaves. 100% correct policy was achieved.

## Discussions and Conclusions

- A biologically inspired model for top-down object-based visual attention control was designed and partially implemented.
- Our results support the idea that the nature of the bottom-up attention is low-level mechanisms, while top-down attention is more like a control or a decision making problem.
- Rather than scanning the image from top-left to bottom-right, to detect an object in the scene, or using global representations (which usually need many computations), our model just looks at a small number of spatial locations.
- Our main contributions were proposing a method to find the low-cost weights of the saliency model to bias it for object detection and a top-down mechanism for controlling the bottom-up saliency model for doing a task.
- It was also shown that training RL with noisy data could compensate low-magnitude noises, but larger values of noise significantly degrade the RL convergence.

## References

- [1] L. Itti, C. Koch, E. Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(1998) 1254-1259.
- [2] J. J. Liang, A. K. Qin, P. N. Suganthan, and S. Baskar, Comprehensive learning particle swarm optimizer for global optimization of multimodal functions, IEEE trans. evolutionary computation, 9(2006), 3.
- [3] M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex. Nature Neuroscience, 2(1999),11, 1019-1025.
- [4] T. Serrn, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, Object recognition with cortex like mechanisms, IEEE Trans. Pattern Anal. Machine Intell. 29(2007), 3, 411-426.
- [5] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, NewYork, 1995.
- [6] A. K. McCallum, Reinforcement learning with selective perception and hidden state. Doctoral dissertation, Department of Computer Science, University of Rochester, 1995.

## Acknowledgement

This work is funded by the school of cognitive sciences, IPM, Tehran, IRAN