Online Learning of Task-driven Object-based Visual Attention Control

Ali Borji[1,3] (borji@iai.uni-bonn.de), Majid N. Ahmadabadi[1,2] (mnili@ut.ac.ir), Babak N. Araabi[1,2] (araabi@ut.ac.ir)

[1] School of Cognitive Sciences, IPM, Tehran, IRAN, [2] School of Electrical and Computer Engineering, University of Tehran, Tehran, IRAN

[3] Dept. of Computer Science III, University of Bonn, Germany

Abstract

A biologically-motivated computational model for learning task-driven and object-based visual attention control in interactive environments is proposed.

Our model consists of three layers. First, in the ***early visual processing layer****, most salient location of a scene is 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 like object recognition at the focus of attention. From this information, a state is derived in the* ***decision making and learning layer****.*

Top-down attention is learned by the U-TREE algorithm which successively grows an object-based binary tree. Internal nodes in this tree check the existence of a certain object in the scene by biasing the early vision and the object recognition parts. Its leaves point to states in the action value table. Motor actions are associated with the 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.

The proposed model is evaluated on visual navigation tasks, where obtained results lend support to the applicability and usefulness of the developed method for robotics.

Proposed Model

An agent working in an environment receives information momentarily through its visual sensor. It should determine what to look for. For this we use *RL* to teach the agent simply look for the most task relevant and rewarding entity in the visual scene (Fig.1).

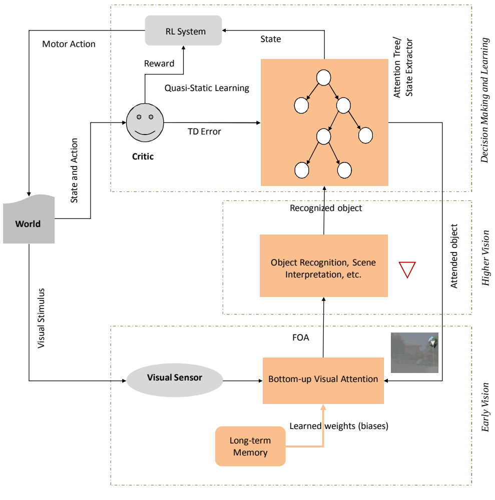


Fig. 1. Proposed model for learning task-driven object-based visual attention control

Example scenario: captured scene through the agents' visual sensor undergoes a biased bottom-up saliency detection operation and focus of attention (FOA) is determined. Object at the FOA is recognized (i.e. is either present or not in the scene), then the agent moves in its binary tree in the decision making and learning layer. This is done repetitively until it reaches a leaf node which determines its state.

The best motor action is this state is performed. Outcome of this action over the world is evaluated by a critic 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. Following subsections discuss each layer of the model in detail.

Early Visual Processing Layer

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

Aim is to find a weight vector which maximizes the object detection rate over a set *T* of *M* training images :

$$T = \{(Im_1, t_1), (Im_2, t_2), \ldots, (Im_M, t_M)\}$$

This is done by minimizing following fitness function by CLPSO [2]:

$$f(\overline{\omega}) = \frac{1}{M}\left(\sum_{i=1}^{M} \mathrm{norm}(\mathrm{Saliency}(Im_i, \overline{\omega}) - t_i)\right)$$

where *norm(.)* is the Euclidean distance between two points in an image. Saliency is the function which takes as input an image and a weight vector and returns the most salient location. t_i is the location of target object in the *i*-th image.

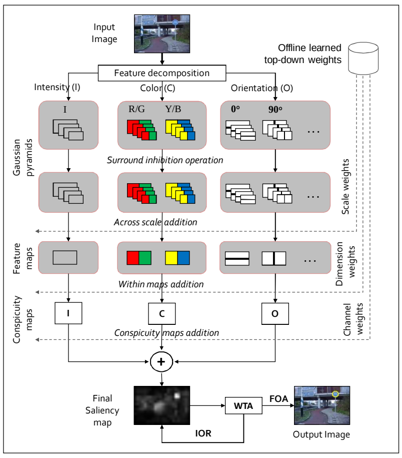


Fig. 2. Biased saliency-based attention model

Fig. 3. Sample objects in natural scenes. Best individual derived after minimization was applied to a test set.

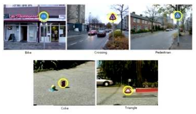
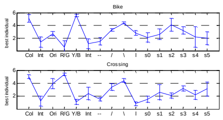
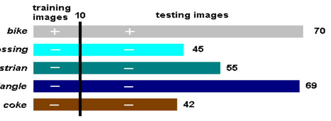


Fig. 4. Learned weights after CLPSO convergence over first two traffic signs averaged over five runs. s_0 to s_5 are scales in the image pyramid.



Higher Visual Processing Layer

The object at the attended location is recognized by the hierarchical model of object recognition (HMAX) [3, 4]. A binary SVM classifier [5], is trained with positive samples of a class and negative samples from other classes. Offline learned classifier in this way is later used for online object recognition.



OR rates over test sets:

91.28% (± 2.8),
93% (± 2.75)
87% (± 3.2)
83.4% (± 4.2)
83% (± 4.2%)

Fig. 5. Object recognition results using C2 features

Decision Making Layer

This layer controls both top-down visual attention and motor actions. The learning approach is an extension of the U-TREE algorithm [6] to the visual domain. 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 *Tree-fixed* phase, *RL* algorithm is executed for some episodes by following ε-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:

$$Q(s_t, a_t) = \alpha\left(r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right) + Q(s_t, a_t)$$

Estimation of aliasing Δ_t*:*

$$= \alpha\Delta_t + Q(s_t, a_t)$$

State discretization occurs in the *RL-fixed* phase where gathered experiences are used to refine aliased states. An object which minimizes aliasing the most is selected for braking an aliased leaf.

Selection of the object o which reduces aliasing the most:*

$$o^* = argmin_o \left(var\,\{y\} - \left(\tfrac{|L_+|}{|L|} var\{y_a|+\} + \tfrac{|L_-|}{|L|} var\{y_a|-\}\right)\right)$$

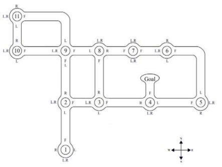
$$= argmax_o \left(\frac{|L_+|}{|L|} var\{y_a|+\} + \frac{|L_-|}{|L|} var\{y_a|-\}\right)$$

Results

Task: simulated visual navigation.

Map of the route, consisting of 11 positions, is shown in Fig. 6. The agent captures 360 × 270 RGB color images. There are 44 states. Natural scenes containing a subset of the objects are presented to the agent (5 for each combination). 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 are present in random locations of scenes. Best actions are shown besides each state. In some states two actions are optimal.



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.7)

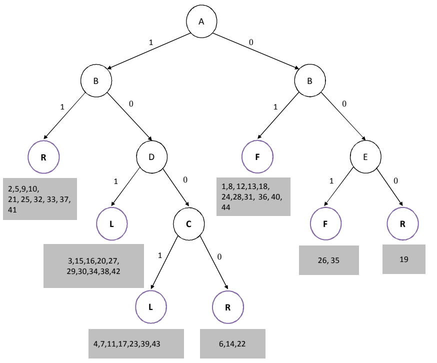


Fig. 7. Learned attention tree for the map of Fig. 6 with pruning. Forty four states were clustered into 7 leaves. 100% correct policy was achieved.

Uncertainty analysis experiment:

Since both saliency model and the *Hmax* have uncertainties, this problem also applies to our model. In this experiment, we analyze how uncertainty in the perception of the agent affects its behavior. Each observation of the agent is incorrect by probability P_u. For instance, P_u = 0.03 means that in 3 percent of observations the agent is not sure that an object is really present in the scene or not. When the agent traverses its attention tree and has to attend to an object, it gets an incorrect result with probability P_u. When observations of the agent are noisy, then the agent develops a probabilistic action selection strategy. The agent could compensate low magnitude of uncertainty in its perceptions (Fig. 8).

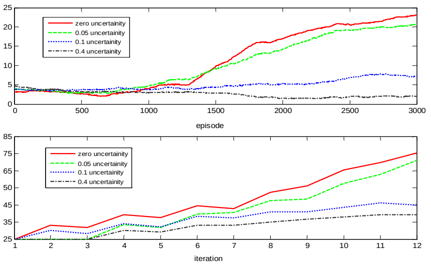


Fig. 8 Top: Cumulative average reward of the agent for different noise levels. Bottom: Cumulative percentage of correct policy during learning. Results are averaged over 7 runs.

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.*
- *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. Serre, 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 was funded by the school of cognitive sciences, IPM, Tehran, IRAN.

Presented at "Bonn Vision Workshop", Bonn, Germany, Oct. 8, 2009.