Image and Vision Computing 28 (2010) 1130–1145

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

# Online learning of task-driven object-based visual attention control

Ali Borji [a,b,*], Majid Nili Ahmadabadi [a,c], Babak Nadjar Araabi [a,c], Mandana Hamidi [d]

[a] School of Cognitive Sciences, Institute for Research in Fundamental Sciences, Niavaran Bldg., P.O. Box 19395-5746, Tehran, Iran
[b] Dept. of Computer Science III, University of Bonn, Bonn, Germany
[c] Control and Intelligent Processing Centre of Excellence, Dept. of Electrical and Computer Eng., University of Tehran, Tehran, Iran
[d] Italian Institute of Technology (IIT), Via Morego 30, 16163, Genova, Italy

ARTICLE INFO

*Article history:*
Received 5 October 2008
Received in revised form 4 August 2009
Accepted 9 October 2009

*Keywords:*
Task-driven attention
Object-based attention
Top-down attention
Saliency-based model
Reinforcement learning
State space discretization

ABSTRACT

We propose a biologically-motivated computational model for learning task-driven and object-based visual attention control in interactive environments. In this model, top-down attention is learned interactively and is used to search for a desired object in the scene through biasing the bottom-up attention in order to form a need-based and object-driven state representation of the environment. 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 specific 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.



## 1. Introduction

Both biological and machine vision systems have to process enormous amount of visual information they receive at any given time. Attentional selection provides an efficient solution to this information overload problem by proposing a small set of scene regions to higher level and more computationally intensive processes; like scene interpretation, object recognition, decision making, etc. In this regard visual attention acts as a front-end to a more complex vision system. Instead of processing all incoming visual information in parallel, the brain has evolved a serial strategy which explains its near real time performance in visual interactive environments.

Visual attention selects and gates visual information based on the saliency in the image itself (bottom-up) [1,2] and on the prior knowledge about the scene (top-down) [3,4]. While bottom-up attention is solely determined by the image-based low-level cues – such as luminance and color contrasts, edge orientation and motion – top-down attention on the other hand is influenced by task demands, prior knowledge of the target and the scene, emotions, expectations, etc. Bottom-up component of the visual attention is mainly examined by the early visual areas of the brain like LGN and V1 [6]. Top-down attentional signals are largely derived from a network of areas in parietal and frontal cortex [5]. Some of the involved areas include the superior parietal lobule (SPL), the frontal eye fields (FEF), the supplementary eye field (SEF) and the lateral prefrontal cortex in the region of the middle frontal gyrus (MFG). In daily life, these two mechanisms interact to control our attentional behaviors [3,7]. Besides acting in spatial domain by selecting spatial locations [8], visual attention can also be directed to particular features such as color, orientation and direction of motion [9]. It is also believed that attention selects objects rather than spatial locations [10,11].

Like humans and primates, artificial creatures (e.g. robots) are limited in terms of allocation of their resources to huge sensory and perceptual information. That is mainly because of the serial processing mechanisms used in the design of such creatures which allows processing of only a small amount of incoming sensory information. Since they are usually supposed to guarantee a short response time, attention is an efficient solution in robotics as in biological systems. In order to gain the maximum cumulative reward in the minimum time, agents should be able to perform perceptual and physical actions simultaneously. These perceptual

* Corresponding author. Tel.: +98 21 22294035; fax: +98 21 22280352.
*E-mail addresses:* borji@iai.uni-bonn.de, borji@ipm.ir (A. Borji), mnili@ut.ac.ir (M.N. Ahmadabadi), araabi@ut.ac.ir (B.N. Araabi), Mandana.hamidi@iit.it (M. Hamidi).

0262-8856/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2009.10.006

actions are available in several forms like where and what to look in the visual modality. However, the main concern is how to select the relevant information, since relevancy depends on the tasks and the goals. In this study, we consider task relevancy of visual information and aim to extract objects which help the agent to discover its state faster for decision making.

It is important that a solution for learning task-based visual attention control to take into account other relevant and dependent cognitive processes like learning, decision making, action selection, etc. Some evidences in this regard exist in both biology and engineering. It has been previously shown that attention and eye movements are context-based and task-dependent [12]. Previous experiences also influence attention behaviors which indicate that attentional control mechanisms can be learned [13]. Some neuropsychological evidences suggest that human beings learn to extract useful information from visual scenes in an interactive fashion without the aid of any external supervisor [14]. In [15], it has been shown that attention is also affected by decision behaviors. These findings are in accordance with a new and pragmatic view in Artificial Intelligence (AI) known as embodied and situated intelligence [16]. It states that intellectual behaviors, representations, decisions, etc. are the product of interactions among brain, body and environment. In [52], authors have provided their views for architecture of situated vision systems, how to tackle the design and analysis of perceptual systems and promising future directions. In particular they have focused on inspiring from complex vision systems, like human vision, to build synthetic vision systems and integrating them with action and learning modules. There are also other supporting evidences in psychology claiming that human mind and intelligence have been formed interactively through an evolutionary process [17]. Instead of attempting to segment, identify, represent and maintain detailed memory of all objects in a scene, there are evidences that claim our brain may adopt a need-based approach [18], where only desired objects are quickly detected in a scene, identified and represented. Considering the above evidences, in this work, we introduce a model to consider the influences of task, action, learning and decision making to control top-down visual attention of an agent.

In many real-world situations, the environment is unfamiliar or not clearly defined. Moreover, required information and the optimal responses are not known at the design time. Therefore, fixed and predefined design of attention control strategies in such situations is less useful. Some complicated behaviors of humans like reading, writing, driving, etc. which need complex physical actions and attentions witness that such behaviors have been developed based on humans interaction with the surrounding world. Thus, interactive and semi-supervised approaches, e.g. Reinforcement Learning (RL) [19], seem to be the most suitable techniques for learning top-down visual attention control and action selection strategies. Such learning mechanisms have the benefit of adapting the agent to dynamic, complicated and non-deterministic environments. In RL, agents learn action-values in each state by receiving a reinforcement signal from the critic. Another characteristic of RL methods is their ability of online learning which is required for interacting with stochastic and slowly changing environments. There are mathematical convergence proofs for these methods and they are biologically plausible [20].

Our proposed top-down visual attention model is built upon a sound and widely used bottom-up visual attention model proposed in [5,21]. This model is based on the idea of saliency map, an explicit 2D topographical map that encodes stimulus conspicuity or saliency at every scene location [22]. Bottom-up model in its original form is solely data-driven and simply selects some spatial locations without using any feedback mechanism or top-down gains. Some researchers have tried to add top-down capabilities to this basic model [23–25], for instance by biasing it toward selecting specific objects. While such models are interesting, they have been partially successful to handle a limited category of tasks. Modeling top-down task-based influences on visual attention is difficult, since a general definition for a task does not yet exist. In this study, RL is used by the agent to interactively learn to search for relevant objects in a scene through biasing the bottom-up attention and the object recognition part in order to find its state and to choose physical actions accordingly. In particular, we use the U-TREE algorithm [26] to dynamically discretize the visual state space when perceptual aliasing occurs. Aliasing means that two perceptions demanding different actions are classified under the same state. That way an object-based binary tree is generated which is used for controlling top-down object-based visual attention. Our model is inspired by the abstract findings from neuroscience and psychology.

In Section 2 of this paper, related researches are reviewed. Our proposed approach is explained in Section 3. Experiments and results are shown in Section 4 and finally, Section 5 summarizes and concludes the paper.

## 2. Related researches

In this section, we review studies which are directly related to ours, especially those which have considered learning aspects of visual attention in concert with decision making. First we review some existing hypotheses and viewpoints on visual attention mainly derived from behavioral studies and then bring some successful approaches from AI for learning attention control.

An important evidence from biology reported in [13], states that attention could be learned by past experience. In a behavioral task, human subjects were supposed to answer a question about a quality of a specific visual item in a synthetic visual search scene. Subjects had lower reaction times when the quality of the object stayed the same during successive trials. This study shows that subjects developed a memory during the task. A modeling work trying to explain such behavioral data is done in [27]. They have proposed an optimization framework to minimize an objective function which is a sum over the reaction time in each state weighted by the probability of that state to occur. Then using a Bayesian Belief Network (BBN), they solved that minimization problem. These results encourage using a learning approach for attention control in AI.

Some RL studies have previously been proposed for modeling top-down visual attention control in humans. Since eye movements have high correlation with overt visual attention, these studies have tried to explain eye movement data. In [28], RL is used for modeling the behavior of an expert reader by predicting where eyes should look and how long they should stay there for achieving best comprehension from the text. Another model of human eye movements is proposed in [29] that directly ties eye movements to the ongoing demands of behavior.

RL has also been used for deriving visual attention policies for mobile robots. In [30], a 3-step architecture is proposed for an object recognition task. First, it extracts potential focuses of interest (FOI) according to an information theoretic saliency measure. Then it generates some weak object hypotheses by matching the information at the FOIs with codebooks. The final step is done using $Q$-learning with the goal of finding the best perceptual action according to the search task. In [31], two approaches are proposed in a robotic platform with neck, eyes and arms for attention control. The first approach is a simple feedforward method which uses back-propagation learning algorithm while the second one uses reinforcement learning and a finite state machine for state space representation. In [32], another robotic platform containing articulated stereo-head with 4 degrees of freedom is presented which