The Performance of a Biologically Plausible Model of Visual Attention to Localize Objects in a Virtual Reality

Amirhossein Jamalian(✉), Frederik Beuth, and Fred H. Hamker

Artificial Intelligence, Chemnitz University of Technology,
Strasse der Nationen 62, 09111 Chemnitz, Germany
amirhossein.jamalian@informatik.tu-chemnitz.de

Abstract. Visual attention, as a smart mechanism to reduce the computational complexity of scene understanding, is the basis of several computational models of object detection, recognition and localization. In this paper, for the first time, the robustness of a biologically-constrained model of visual attention (with the capability of object recognition and localization) against large object variations of a visual search task in virtual reality is demonstrated. The model is based on rate coded neural networks and uses both bottom-up and top-down approaches to recognize and localize learned objects concurrently. Furthermore, the virtual reality is very similar to real-world scenes in which a human-like neuro-cognitive agent can recognize and localize 15 different objects regardless of scaling, point of view and orientation. The simulation results show the neuro-cognitive agent performs the visual search task correctly in approximately 85.4 % of scenarios.

Keywords: Computational neuroscience · Object localization · Object recognition · Virtual reality · Visual attention · Visual search

1 Introduction

Several tasks require looking for a certain object in the environment. This is known as *visual search* in the literature. In such a task, the typical human knows what exactly she/he is looking for. This predefined knowledge of the searched object stimulates a *top-down* signal in her/his brain [1], called *feature-based attention*. It originates in the *prefrontal cortex (PFC)* region (which could be considered as the object memory) and is given to the object recognition pathway of the brain which is known as *ventral stream* (Fig. 1). This pathway starts in the primary visual cortex (V1), continues through the fourth visual cortex (V4) and reaches the inferior temporal cortex (IT). The processing of the ventral stream is modulated by the frontal eye field (FEF).

Computational models of visual attention simulate the processing in the brain to different degrees. The simplest approaches are *bottom-up saliency models* [2], which simulate simple features as located in V1 to calculate a saliency map. This map indicates regions of the visual field containing the most information. This approach is often combined with top-down attention towards simple features to favor target relevant features (*top-down saliency models*, [2]). One step further goes through

A.E.P. Villa et al. (Eds.): ICANN 2016, Part II, LNCS 9887, pp. 447–454, 2016.
DOI: 10.1007/978-3-319-44781-0_53

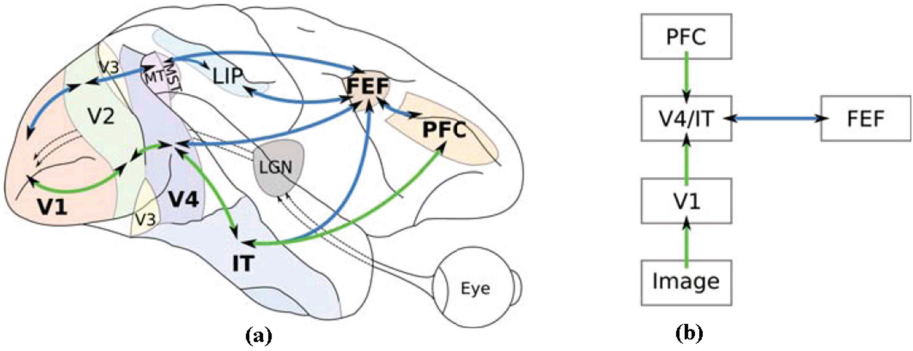

Fig. 1. (a) Primates' Visual Attention System [4]. The green arrows show the Ventral Stream and the blue arrows correspond to the Dorsal Stream, which process the type and location of an object respectively. Bottom-up processing is denoted by arrows from left to right and top-down processing (mediated by attention) by arrows from right to left. (b) Areas and connections which are simulated in this attention model. They are printed in bold in (a). (Color figure online)

proto-object-based models that define spatial regions belonging to one object (the proto-objects) in the saliency map [3]. Finally, there are *neuro-computational models* simulating the attentional processing of the brain, in which feature-based and spatial attention are closely entangled and thus operate in parallel. *Feature-based attention amplifies* the activation of neurons that encode the searched object (and suppress the others), while *spatial attention* amplifies the neurons regarding their location information. This process iterates until the location of the target would be encoded in a spatial map (FEF). Due to their parallel nature, they are called iterative [3] or holistic attention models [4].

However, such neuro-computational models have been typically developed for psychophysical experiments using very simple stimuli, and thus cannot deal with real-world objects. Hence, our aim is to further develop such models for making them applicable to real-world scenarios. A few models have been already demonstrated with real-world objects [4–9]. Yet, they have mostly used static input material. Only two studies have used 3D environments, but merely very simple ones, like three objects in a robotic setup [5] or cubic objects in a black-background virtual reality (VR) [6]. Hence, a next step towards a real-world application would be to benchmark such models in a more complex VR setup. Therefore, in this work, the performance of our model in a VR is evaluated.

In neuro-computational models, objects are encoded typically by neurons representing a specific view (view-tuned neurons); as such cells have been found in area IT [10]. In theory, this approach can encode objects under any kind of transformation. We have previously demonstrated this for objects under different rotations [4–6], different disparities [5], and small difference in the scaling [5]. Here, we will benchmark some of the remaining transformations, i.e. larger changes in the scaling and different views of the object (top versus side view). We have chosen these kinds of transformations as they occur when a virtual agent walks towards objects located on a table, which seems a plausible scenario for VR and real-world applications.

We now present the further development of our biologically-plausible model of visual attention, in particular how it is customized to work in the VR and how object representations are created by a novel learning algorithm, called *One-shot Learning*.

2 Model Structure and Functionality

2.1 Overview of the Model

This model builds upon previous integrative attention models like Hamker (2005) [8]. The present version [4] is based on a recently developed cortical microcircuit model of attention, which replicates many neurophysiological data sets of attention [11]. This and the other biological foundations of the model are explained in the original publication, while we give here only a functional overview. The model simulates the ventral stream pathway in primates' visual system (see Fig. 1(a) as well as the frontal eye field (FEF) and the prefrontal cortex (PFC). The diagram of the model is depicted in Fig. 2. Its input is an RGB image that is firstly processed by a model of the primary visual cortex (V1), whose cells encode oriented edges, red-green and blue-yellow color contrasts. Afterwards, this activity is fed to a higher visual area (HVA) encoding object views. HVA represents high-level visual areas like V4 (fourth visual cortex) and IT (inferior temporal cortex). These object-view maps are constructed via convolutions of receptive fields of V1 layer neurons (as pre-synaptic layer) by a pre-generated weight matrix, calculated offline by the one-shot learning procedure (next section).

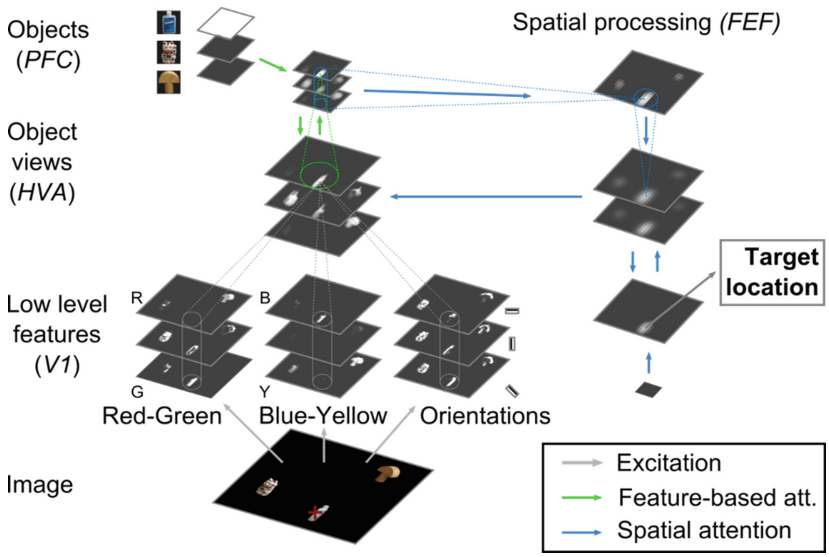

Fig. 2. The model of visual attention [4]. The processing is illustrated at the task to localize the "bottle", indicated by the red cross. See main text for details. (Color figure online)

During the search for an object, a top-down feature-based attention signal is sent from PFC to HVA. This signal contains the encoded features or views of the searched target object. Applying the attention signal, neurons encoding the corresponding object-views will be more excited in comparison to the others, while local inhibitory connections suppress views belonging to other objects. This pattern of activity will be send to the FEF for further spatial selection.

The FEF region is split into three parts, according to its neurophysiological cell properties: FEF-Visual (FEF-v), FEF-visiomovement (FEF-vm) and FEF-Movement (FEF-m). FEF-v is a kind of saliency map and contains the places where the target is probably located. FEF-vm is responsible for focusing neuronal activity at the target location. Additionally, this map projects back to visual areas (HVA), forming a recurrent loop from which spatial attention emerges. This kind of attention does not only excite the activity of the neurons in HVA around the target location, but also suppress the activity of the other neurons to decrease the effect of distractors. This cycle iterates until a saccade plan is completed, indicated by the fact that the neurons in the FEF-m layer reach a threshold. This activity blob indicates the location of the target in the image.

The model has been benchmarked in a task with up to 100 objects (COIL-100 database, [12]) and three background classes [4] and achieved an object localization accuracy of 92 % on black, of 71 % on noisy, and of 42 % on real-world backgrounds.

2.2 Model Customization for VR

The VR used in this project has been developed within the European Union project "Spatial Cognition" [13]. It is part of a framework to simulate neuro-cognitive agents in a virtual environment [14]. This framework consists of the VR (based on the game engine Unity [15]) in which the agent is placed, and a neuro-simulator [16] to simulate the "brain" of an agent. The VR provides all sensory data to the agent like stereoscopic images (from which we use the left eye here), collisions, etc., while the agent can execute actions in the VR like rotating eyes, walking, etc.

For the VR, we developed a simple and straight-forward algorithm to learn the object representation. Our *One-shot Learning algorithm* creates an object view representation in HVA directly from a stimulus patch showing the object under this view. For this, the stimulus is firstly processed by V1 and then the algorithm calculated from this V1 activity pattern (cell index i) directly the weight matrix of a HVA cell (index j). The method creates negative weights from weak V1 activities as these represent V1 features which do not appear in the object view, and positive weights from high V1 activities as these V1 features represent the object view (Eq. 1). The amount of negative weights is calibrated per view independently via the parameter v_j. A higher amount of negative weights tunes the HVA neuron more specifically to its preferred view.

Normally, the method would learn the background in the patch along with the object. Yet, this is a problem when the objects appear very small at the patch, i.e. for farer distances, as the resulting object presentations would mainly encode the background. To solve this problem, we introduce a spatial selection mask S. The mask

contains binary elements and allows selecting only a spatial part of the patch for learning. We choose five different circular masks, each one for one learned distance. Finally, the weights are normalized so that if the same stimulus appears again, the HVA neuron reacts for it maximally (here chosen as 1) (Eq. 2).

$$A_{\{i,j\}} = f\left(r_i^{V1}, v_j\right) \cdot S, \qquad with: f(r,v) = \begin{cases} -(v-r)^2, r < v \\ (r-v)^2, r \geq v \end{cases} \tag{1}$$

$$w_{\{i,j\}} = A_{\{i,j\}} / \sum\nolimits_{\{i'\}} A_{\{i'j\}} \cdot r_{\{i'\}}^{V1} \tag{2}$$

Customizing the model for this VR, we learned firstly 15 different objects with the One-shot Learning algorithm (Fig. 3(a). We used 12 differently rotated views of each object (every 30°), and 5 different distances between the agent and the target object. Hence, for each object, the model has 12 × 5 = 60 views in HVA. Selected views and distances of a typical object (the green racing car) are depicted in Fig. 3b. Since the total number of learned objects is 15, HVA has 15 × 60 = 900 view-related and PFC 15 object neurons. Besides, the VR stimuli have finer structures than the previously used COIL stimuli. To recognize them, we reduced the receptive field size in V1 (from 19 to 9 pixels) and the spatial pooling factor (from 10:1 to 6:1).

3 Simulation Results in VR

We tested the model in the VR at 3000 test scenes, whereby each scene contains at least 3 different objects under a variation of transformations: arbitrary rotations, arbitrary agent-object distances, and nine randomly-chosen positions on the table. In relation to this test set, the training set contains the objects under 12 fixed rotations (every 30°), 5 fixed distances (every 0.5 units in the VR), and at the center position. This training set was used for learning the object representation and was completely separated from the test data. In each of the 3000 test scenes, every of the three objects were considered one time as target (searched object) while the others would be the distractors, resulting in over 9000 localization tasks. On average, the model can localize the searched object in

Fig. 3. (a) The 15 target objects in Virtual Reality. (b) Six selected views of a typical object (the green racing car) at the nearest, middle, and farthest distance. (Color figure online)

approximately 180–200 steps (each step is the simulation of one millisecond in primates' brain). In Fig. 4, we depicted the result for one typical case.

The performance of the model at the 3000 test scenes is illustrated in Fig. 5a as a confusion matrix. Such a matrix illustrates for each presented target object (Y-axis) the object that has been selected by the model (X-axis). As it can be seen in this figure, in 85.4 % of the cases the model can localize the target correctly (the saccade landed within the object borders or not more than a half object away). Low accuracy values under 50 % are illustrated in red to show mislocalizations. In the matrix, the horizontal axis denotes the localized object or two special cases: B and No. The case B (Background) indicates that one non-sense point in the background was selected instead of the target, and the case No (No-localization) indicates that no location was selected because the model did not converge on any object location including background.

Figure 5b illustrates the object localization accuracy of the model for five different distances between the agent and the objects on the table. Due to the varying distances, the target object (and all others) appears under different scaling and viewpoints, thus we evaluate the robustness of the object localization against these object transformations. Besides the scaling, the viewpoint also changes as the agent looks from half-top to the objects at the closest distance and from the side at the farthest distances. It can be seen

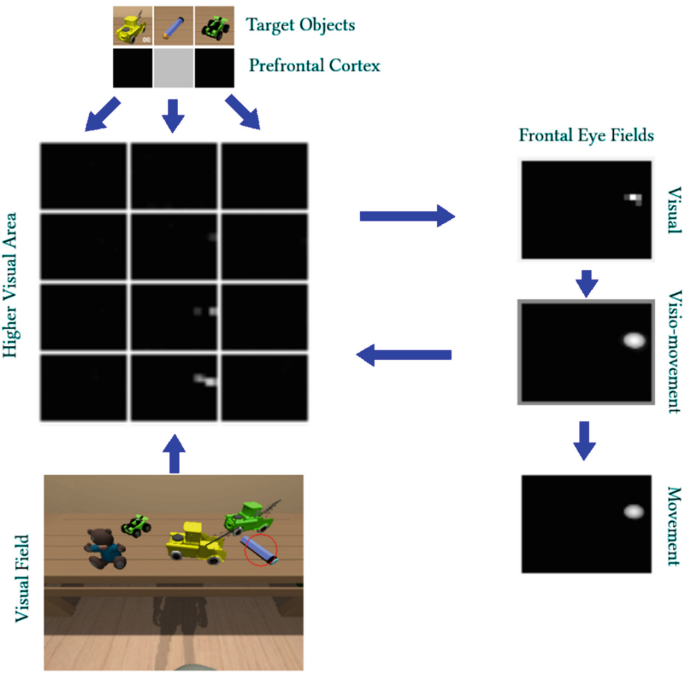

Fig. 4. The searched target (blue pencil) has been recognized and localized at the end of the simulation (indicated by a red circle). Every column of the higher visual area (HVA) is regarded to one object (as depicted in prefrontal cortex), showing four exemplary views. (Color figure online)

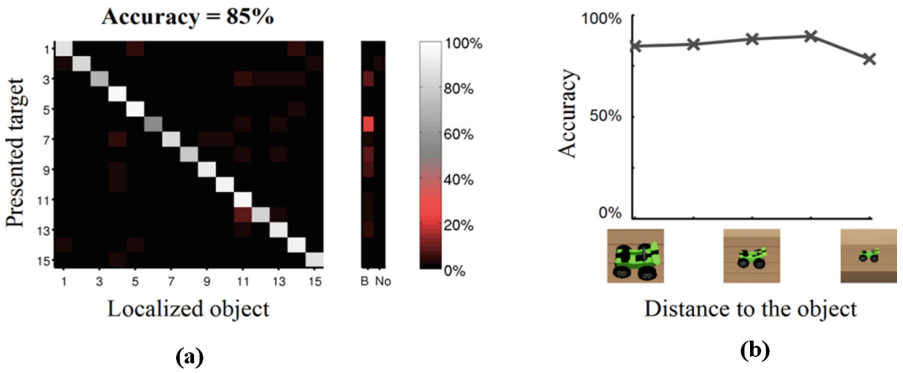

Fig. 5. (a) Performance of the model at 9000 localization tasks. (b) Performance dependent on the distance between agent and the object on the table, whereby the left side denotes the closest distance and the right side the farthest.

that the accuracy is quite stable for the fourth closest distances, showing the robustness of the object localization against changing viewpoints and scaling. The only exception is the farthest distance as the object appears very small (about 15 × 15 pixels). In this case, the accuracy of the model drops. However, it is still 78 % which is quite acceptable regarding its scale.

4 Conclusion, Limitations and Future Works

Here, the performance of a biologically-plausible model of visual attention with the capability of object recognition and localization in a VR has been demonstrated. The model used both bottom-up and top-down approaches in an iterative fashion to perform its visual search task based on an offline supervised learning phase called one-shot learning. In principle, the one-shot learning is easier to use, faster and has a better performance in comparison with previous attempt [5], but produces object representations with more cells. Thus, it is suitable to use in scenarios with a limited number of objects. In the VR, the human-like neuro-cognitive agent can perform the visual search task for 15 different objects in various rotations, places, viewpoints, and scales. The simulation results show that the agent is able to recognize and localize the objects correctly in 85.4 % of cases out of 3000 different scenes as visual search scenarios. It is the first time that such performance evaluation, in presence of multi scaling and viewpoints in VR, is performed. The performance of this model in comparison with similar model [5] which has been evaluated on COIL-100 dataset with real-world background is remarkably better (85.4 % versus 42 %). Although current approach and the old one [5] has been evaluated based on different datasets, we expect the new one would have a better performance on COIL-100 as well, since in COIL-100 all scenes are constructed by objects with same scales. In other words, the new approach performs the task better even in more complicated scenarios. Hence, its performance should be better on COIL-100 dataset as well. However, we would compare them on same datasets, COIL-100 as well

as VR, as one of our future works. Besides, we will attempt to implement the model on iCub robot and evaluate its performance in real-world scenarios.

Acknowledgement. This work has been supported by the European Union project "Spatial Cognition" under grant agreement no 600785.

References

1. Miller, E.K., Buschman, T.J.: Cortical circuits for the control of attention. Curr. Opin. Neurobiol. **23**(2), 216–222 (2013)
2. Borji, A., Itti, L.: State-of-the-art in visual attention modelling. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **35**(1), 185–207 (2013)
3. Jamalian, A., Hamker, F.H.: Biologically-inspired models for attentive robot vision: a review. In: Innovative Research in Attention Modeling and Computer Vision Applications, pp. 69–98. Information Science Reference, Hershey (2016)
4. Beuth, F., Hamker, F.H.: Attention as cognitive, holistic control of the visual system. In: Workshop of New Challenges in Neural Computation (NCNC 2015), Aachen (2015a)
5. Antonelli, M., Gibaldi, A., Beuth, F., Duran, A.J., Canessa, A., Chessa, M., Solari, F., del Pobil, A.P., Hamker, F., Chinellato, E., Sabatini, S.P.: A hierarchical system for a distributed representation of the peripersonal space of a humanoid robot. IEEE Trans. Auton. Ment. Dev. **6**(4), 259–273 (2014)
6. Beuth, F., Wiltschut, J., Hamker, F.H.: Attentive stereoscopic object recognition. In: Workshop of New Challenges in Neural Computation (NCNC 2010) (2010)
7. Chikkerur, S., Serre, T., Tan, C., Poggio, T.: What and where: a bayesian inference theory of attention. Vision Res. **50**(22), 2233–2247 (2010)
8. Hamker, F.H.: The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. Comput. Vis. Image Underst. **100** (1-2), 64–106 (2005)
9. Walther, D.B., Koch, C.: Attention in hierarchical models of object recognition. Prog. Brain Res. **165**, 57–78 (2007)
10. Logothetis, N.K., Pauls, J., Poggio, T.: Shape representation in the inferior temporal cortex of monkeys. Curr. Biol. **5**(5), 552–563 (1995)
11. Beuth, F., Hamker, F.H.: A mechanistic cortical microcircuit of attention for amplification, normalization and suppression. Vision Res. **116**(B), 241–257 (2015b)
12. Nene, S.A., Nayar, S. K., Murase, H.: Columbia Object Image Library (COIL-100), CUCS-006-96. Technical report (1996)
13. Hamker, F.H.: Spatial Cognition of humans and brain-like artificial agents. Künstliche Intelligenz **29**, 83–88 (2015)
14. http://www.tu-chemnitz.de/informatik/KI/projects/agents-vr/index.php
15. https://unity3d.com/
16. Vitay, J., Dinkelbach, H.Ü., Hamker, F.H.: ANNarchy: a code generation approach to neural simulations on parallel hardware. Front. Neuroinformatics **9**, 1–20 (2015)