In: ICANN2014, 24th International Conference on Artificial Neural Networks in Hamburg, Germany at 15-19 September 2014. In press.

How Visual Attention and Suppression Facilitate Object Recognition?

Frederik Beuth, Amirhossein Jamalian, and Fred H. Hamker

Chemnitz University of Technology, Artificial Intelligence,
Strasse der Nationen 62, D - 09111 Chemnitz, Germany,
beuth@cs.tu-chemnitz.de

Abstract. Visual attention can support object recognition by selecting the relevant target information in the huge amount of sensory data, especially important in scenes composed of multiple objects. Here we demonstrate how attention in a biologically plausible and neuro-computational model of visual perception facilitates object recognition in a robotic real world scenario. We will point out that it is not only important to select the target information, but rather to explicitly suppress the distracting sensory data. We found that suppressing the features of each distractor is not sufficient to achieve robust recognition. Instead, we also have to suppress the location of each distractor. To demonstrate the effect of this spatial suppression, we disable this property and show that the recognition accuracy drops. By this, we show the interplay between attention and suppression in a real world object recognition task.

Keywords: Object Recognition, Neurorobotics, Real World, Computational Neuroscience, Visual Attention, Suppression

1 Introduction

Object recognition in real world scenarios is a very challenging task. Usually, it involves problems like cluttered scenes analysis, existence of many distracting objects, different scaling, spatial positions, rotations of the objects and etc. The concept of attention can deal with the first two problems, as it can be used to select the relevant target information among the huge amount of sensory data. A vast volume of literature could be found in the field of attention-based object recognition in real-world scenarios or robotics. Many of them are based on bottom-up approaches and assume that the objects of interest are sufficiently salient by themselves. For example Miau et al. [8] combined an attentional front-end with the well-known object recognition system HMAX [12] to recognize either real-world scenes or simple artificial objects like circles and rectangles. Other remarkable real-world applications are the object recognition systems of Walter and Koch [18] and Frintrop and Jensfelt [3]. Since non-salient objects are not detected in bottom-up approaches, other researches used combinations of top-down and bottom-up methods like Hamker [4], Mitri and Frintrop [9], Rasolzadeh and Björkman [10], and Wischnewski et al. [19] (all are real world applications). In this paper, based on terms and concepts of visual attention

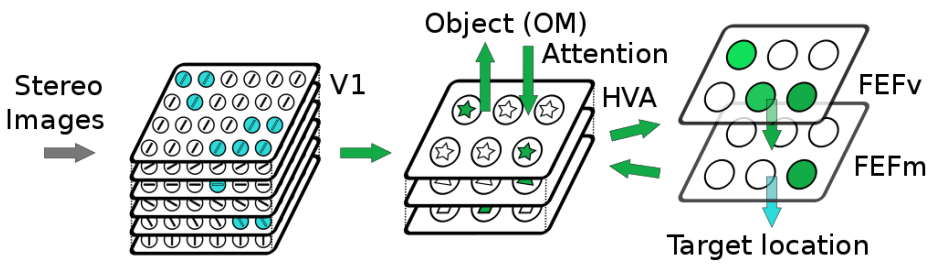

Fig. 1. The object recognition architecture that simulates the brain's visual cortex.

mentioned in [2, 5, 4], we demonstrate the impact of spatial suppression on the robustness of attention-based object recognition. The proposed object recognition system and the learning of invariant object representations are summarized in section 2. Then, section 3 explains the interplay of visual attention with suppression and shows how a new task-specific spatial suppression can be modelled. The accuracy of recognition and localization of the proposed system in presence and absence of the new spatial suppression mechanism is compared in section 4 and finally section 5 concludes the work.

2 Object Recognition System

The object recognition system (Fig. 1) has been developed for a humanoid robot within the European project "Eyeshots" [1]. The goal was to develop a cognitive and biologically plausible object recognition module, so a previously published anatomically and physiologically motivated model of attention was scaled up to allow the processing of real world scenes. Biological background can be found there [5] whereby implementation details reside in [1]. To facilitate reading, its functionality will be explained in the following:

Real world stereo images are fed into the first stage *V1* (primary visual cortex) which encodes simple visual features like the orientation of edges, local contrast differences and retinal disparity [13]. The neurons are organized feature-wise in planes and each plane has the same spatial arrangement as an image (retinotopic organization). Therefore, a particular V1 neuron will be activated if the preferred feature is located at the retinal locations of both eyes underlying its receptive field. The next stage *HVA* (High Visual Area) encodes features representing a single view of an object, similar to cells in the brain areas V4 and IT [7]. HVA is again organized plane-wise and retinotopic. Each view is encoded by the connection weights between V1 and HVA, so each HVA neuron reacts for a specific pattern of V1 neurons (Fig. 2b). These weights were determined in an off-line training phase using unsupervised learning. As this learning should lead to largely depth and scale invariant representation of an object view, our method relies on temporal continuity [16]. The idea is that on the short time scale of stimuli presentations, the visual input is more likely to originate from the same object under the same view, rather than from different objects or views.

Spatial information is encoded in the Frontal Eye Field (*FEF*), simulated by two maps: *FEFv* indicates all possible retinal locations of the searched object

(green dots in Fig. 1) whereby *FEFm* indicates only the final location (single green dot in FEFm in Fig. 1). The FEFv is computed by taking the maximum activity over all the features in HVA. The FEFm is calculated from FEFv by applying a Gaussian filter to reinforce adjacent locations and use competition to suppress others. The resulting target signal is projected back to HVA to select the target location in HVA, too. Over time, a single area of activation emerges in FEFm. If this activity reaches a threshold, a saccade will be triggered towards this target location. Physiologically, FEFv and FEFm represent the visual and movement cell types of the FEF [4, 14].

Visual attention is used to search for a particular object. The objects are encoded in a separate stage, the object memory (*OM*), like in the prefrontal cortex [5]. The bidirectional binding of HVA neurons to an object neuron was manually designed. In general, attention is defined as selecting a certain feature or object over the whole scene (*feature-based attention*) or attending a certain location (*spatial attention*). At neuronal level, this process enhances the firing rates multiplicatively by the amount of received feedback (called gain control [2, 4, 5]). For searching an object, the signal OM→HVA is used to implement feature-based attention, i.e. to enhance all HVA neurons that encode a view of the target object, and to implement feature-based suppression, i.e. to suppress distractors (section 3). Additionally, spatial attention is used to localize the target and to segment it from the background. It is implemented by the feedback projection from FEFm to HVA which enhances all HVA neurons at a certain location and suppresses all other locations.

This processing searches an object by its object-identifying features and segment it at the same time. It is executed in parallel via the loop HVA→FEFv→ FEFm→HVA to avoid the chicken-egg problem of segmentation and localization, i.e. that object segmentation depends on localization, which, in turn, requires the segmentation itself.

3 Visual Attention and Suppression

3.1 Interplay of Attention and Suppression

Attending a certain object means to select a set of object-related features or its location. At the neural level, selection usually involves the enhancement of some neurons and the suppression of others. For the latter, it is crucial to accompany visual attention with a suppression mechanism, typically either through an inhibitory network structure [5] or via a generic suppressive drive [11]. We propose to achieve selection via four modulation mechanisms, similar as in other attention models [2, 4, 5, 11]:

- *Feature-based attention* which enhances the neuronal activity of certain features in HVA over the whole scene. This is used to select the target objects via their visual features. In previous work [2, 5] and this model, it is implemented via top-down connections to HVA (Fig. 3, signal 1a).
- *Feature-based suppression* to suppress distractors over the whole scene (next section).

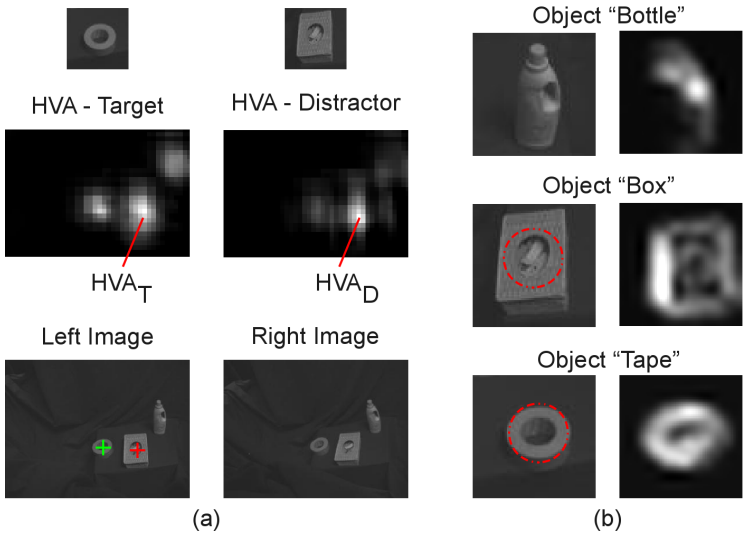

Fig. 2. a) Misclassified example without spatial suppression (C2): the tape (target, green cross) was incorrectly recognized as the box (red cross). b) HVA encodes views of objects. For each object (left), the weights V1→HVA (right) of one exemplary HVA neuron are illustrated.

- *Spatial attention* which enhances the neuronal activity of all features present at a certain location. This mechanism is used to focus attention to a single target location. In previous work [2, 5] and this model, it is implemented via the HVA-FEF loop (signals 1a, 2a, 3a and 4 in Fig. 3).
- *Spatial suppression* which decreases the neuronal activity at certain locations. This mechanism is used to move attention away from the location of distractors (next section).

Despite its function in object recognition, this concept of visual attention together with suppression is justified by neurobiological theories such as biased-competition [17]. According to the biased-competition framework, competition takes place when two different stimuli are presented inside a receptive field of a neuron. In the unattended condition, both stimuli suppress each other slightly which can be measured as recorded neurons fire less in comparison with a condition where only a single stimulus is shown. However, if attention is directed to one of the stimuli, the neuron encoding the preferred object fires more strongly whereby a neuron preferring the other stimulus is strongly suppressed.

3.2 New Task-Specific Spatial Suppression

Here we propose an additional mechanism, which actively suppresses the location of distractors. We found that the existing suppression mechanism in models of visual attention [5, 11] were not sufficient to suppress distractors under all conditions in a real world scenario. A closer examination of the misclassified cases (Fig. 2a) reveals that a distractor was incorrectly recognized as the target

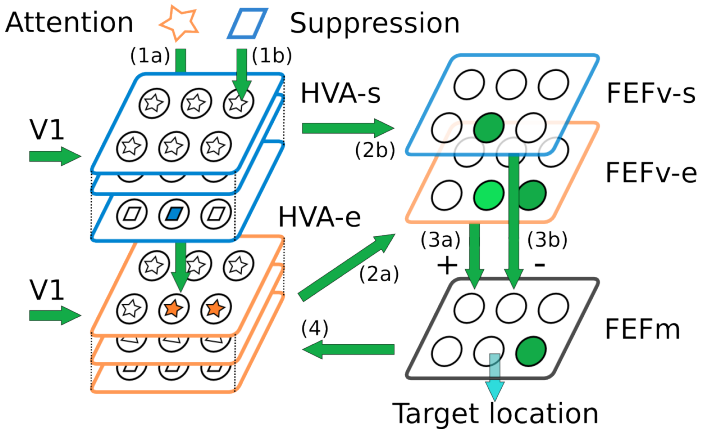

Fig. 3. Object recognition system. It consists of an excitatory component (orange, HVA-e, FEF-e) and a suppressive component (blue, HVA-s, FEF-s).

under the condition that some parts of the distractor are visually similar to some parts of the target, i.e., when the feedforward weights (V1→HVA) of two view neurons, belonging to different objects, contain a similar pattern. In our setup, the box and the tape share the inner ring as such a similar pattern (red circle in Fig. 2b). At the distractor location, obviously HVA neurons encoding the distractor view (denoted HVA_D, Fig. 2a) will react strongly, but also the neurons encoding the target view (denoted HVA_T, Fig. 2a) will respond. This HVA_T response is problematic as it will induce an incorrect FEFm activity indicating the wrong target location (red cross in Fig. 2a) instead of the correct one (green cross).

Therefore, a new mechanism was necessary to suppress the HVA_T response at the distractor location. We modeled a task specific *spatial suppression* in which at first, each distractor is seperately to the targets encoded in HVA and FEFv. Secondly, this information is projected inhibitory to FEFm to suppress the distractor location. Thus, HVA and FEFv are split in an excitatory part ($HVA-e$, $FEFv-e$, orange components in Fig. 3) representing targets, as described in section 2, and a suppressive part ($HVA-s$, $FEFv-s$, blue components in Fig. 3) representing distractors. As targets and distractors are typically defined within a specific task, there exists an excitatory feedback projection from higher cortical areas to HVA-s providing task-specificity. This signal serves as *feature-based suppression* (Fig. 3, signal 1b) which enhances the firing rates of the distractors in HVA-s. This information is subsequently projected (Fig. 3, signal 2b) to a separate suppressive FEFv layer ($FEFv-s$, blue) encoding the locations of distractors. Hence, the system can suppress them in FEFm via inhibitory connections (Fig. 3, signal 3b). The FEFm now contains only the location of the target object and projects this back (Fig. 3, signal 4) to HVA-e. In Fig. 3, the spatial suppression effect is visible on the location of the distractor "trapezoid" (single green dot in FEFv-s): it is incorrectly encoded in FEFv-e (lower middle green dot), but is successfully filtered out in FEFm (white circle).

Concerning the biological foundation, we assume that a signal originating from higher cortical areas, e.g. prefrontal cortex [5], is projected back to V4/IT representing instruction like "ignore these objects". We expected that suppression occurs rather rarely in the cortex as it requires a similar encoding of two different objects (Fig. 2b) which is typically avoided [15]. However, we use always the suppression mechanism. As the neurons in FEFv-s will respond to visual stimuli, we denote them despite their suppressive function as visual neurons [14]. However, as their activity can preserve a fixation, they may be interpreted as fixation cells [14]. Physiologically, the fixation cells suppresses either globally a saccade or a single one in a specific direction [6], whereby our implementation results in a suppression of specific locations in the field of view. On the other hand, we do not model the coordinate system transformation in the cortex, so a suppression of specific directions [6] could be functionally the same as the suppression of specific location as in our model. But to verify this, more physiological data is required. Concerning the relation to the standard attention paradigm, e.g. biased-competition, tuning-curve modulation and surrounds suppression [11], the new mechanism shares with them the task-specificity, but serves as a really different function. Hence, it is beyond the scope of this paper to investigate this relation more closely.

In summary, the task-specific spatial suppression facilitates object selection especially if the objects are very similar and challenging to discriminate.

4 Experimental Results

The system was tested under two different conditions (C1 and C2) to evaluate the effect of the new task specific spatial and feature-based suppression:

C1 The system was used with its full capabilities. This is the reference condition.
C2 Spatial suppression was disabled to investigate its influence by cutting the connection from FEFv-s to FEFm (Fig. 3, signal 3b). As this effectively disables both spatial and feature-based suppression, both mechanisms are evaluated together.

The discriminative ability of the object recognition was evaluated on a test set consisting of 27 real world scenes. Each scene was captured as a grayscale, stereo image by the robotic cameras (Fig. 2a shows one example). This test data was separately recorded from the trainings data[1]. Each test scene contains three objects and each object was recognized separately, resulting in 81 object discrimination and localization tests. The system's object discrimination rate drops from 100% in condition C1 to 95% in condition 2 (see Tab. 1 left) illustrating the effect of the proposed spatial suppression. The perfect discriminative accuracy in condition C1 is likely due to the fact that we benchmarked three objects, only. As the focus of the original project [1] was on the overall interplay of the modules, and not on the development of a novel object recognition approach, the number of recognizable objects was kept low. Nevertheless, the system is able to represent other views or objects due to its temporal continuity learning and thus, the approach can successfully be used with a larger number of objects [2].

Table 1. Left) The discrimination abilities (in %) are illustrated by a confusion matrix for each of the conditions C1 and C2. The ordinate denotes the target object and the abcise the detected object. In comparison to C1, the new spatial suppression was disabled in C2. Right) Localization rates in % and maximal mislocalizations in pixel are denoted for each object under the same conditions C1 and C2.

	C1: Full			C2: Disabled		
Object	**Box**	**Bot.**	**Tape**	**Box**	**Bot.**	**Tape**
Box	100	0	0	100	0	0
Bot.	0	100	0	0	96	4
Tape	0	0	100	11	0	89

	C1: Full		C2: Disabled	
Object	**Rate**	**Mis.**	**Rate**	**Mis.**
Box	96	7	96	8
Bottle	96	20	92	16
Tape	96	6	96	3

The mislocalization of the target was measured to evaluate the spatial precision of the system. Localization was rated as correct if the saccadic target point was located within an object border. The amount was measured as the Euclidian distance from the saccadic target point to the closest object border. As this evaluation should ignore recognition errors and should only measure spatially inaccurate target coordinates, the distance is always measured to the closest object, even if it was an incorrectly selected one. The localization rate was around 96% in condition C1 and 94% in C2 (see Tab. 1 right). Therefore, disabling the spatial suppression has only little influence on the localization accuracy. That result is not surprising as the spatial precision of the FEF depends mostly on the spatial arrangement of the scene which is identical in both conditions. Therefore, the evaluations shows that this task specific spatial suppression reduces false recognition in the case of similar views belonging to different objects and so improve the overall performance.

5 Conclusion

Visual attention can facilitate object recognition by selecting the relevant target information in the huge amount of sensory data. As such a selection process requires the enhancement of some part of the data and the suppression of the rest, it is always crucial to combine the enhancement effect of attention with an appropriate suppression mechanism.

We proposed a new and biological plausible spatial distractor suppression for existing models of visual attention and showed that this mechanism improves object recognition, especially if parts of the target and the distractor are visually similar. This induces an increase of the recognition rate from 95% to 100%. The increase of only 5% arises from achieving perfect performance on a benchmark setup containing only three objects. The low number of objects was requiered by the orginal project [1], but as the object recognition system is able to learn invariant representations of arbitrary objects, it can be used in setups with a greater number of objects [2]. In such a case, it is more likely that views will contain visually similar parts, so we expect that the new spatial distractor suppression will become more important in such scenes.

Acknowledgments This work has been supported by the EC Project FP7-ICT "Eyeshots: Heterogeneous 3-D Perception across Visual Fragments" (no. 217077) and in part by the EC Project FP7-NBIS "Spatial Cognition" (no. 600785).

References

1. M. Antonelli, A. Gibaldi, F. Beuth, A. J. Duran, A. Canessa, M. Chessa, F. Hamker, E. Chinellato, and S. P. Sabatini. A hierarchical system for a distributed representation of the peripersonal space of a humanoid robot. *Accepted for IEEE Trans. Auton. Mental Develop.*, pages 1–15, 2014.
2. F. Beuth, J. Wiltschut, and F. Hamker. Attentive Stereoscopic Object Recognition. In T. Villmann and F.-M. Schleif, editors, *Workshop NCNC2010*, page 41, 2010.
3. S. Frintrop and A. Nuchter. Saliency-based object recognition in 3D data. In *IROS2004*, pages 2167–2172, 2004.
4. F. H. Hamker. The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *J Comput Vis Image Underst*, 100:64–106, 2005.
5. F. H. Hamker. The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement. *Cerebral cortex*, 15(4):431–47, 2005.
6. R. P. Hasegawa, B. W. Peterson, and M. E. Goldberg. Prefrontal neurons coding suppression of specific saccades. *Neuron*, 43(3):415–25, 2004.
7. N. Logothetis, J. Pauls, and T. Poggio. Spatial Reference Frames for Object Recognition. Tuning for Rotations in Depth, 1995.
8. F. Miau, C. Papageorgiou, and L. Itti. Neuromorphic algorithms for computer vision and attention. In *ISOST2001*, volume 4479, pages 12–23, 2001.
9. S. Mitri and S. Frintrop. Robust object detection at regions of interest with an application in ball recognition. In *ICRA2005*, number April, pages 126–131, 2005.
10. B. Rasolzadeh, M. Bjorkman, K. Huebner, and D. Kragic. An Active Vision System for Detecting, Fixating and Manipulating Objects in the Real World. *Int J Robot Res*, 29(2-3):133–154, 2009.
11. J. H. Reynolds and D. J. Heeger. The normalization model of attention. *Neuron*, 61(2):168–85, 2009.
12. M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat Neurosci*, 2:1019–1025, 1999.
13. S. P. Sabatini, G. Gastaldi, F. Solari, K. Pauwels, M. M. Van Hulle, J. Diaz, E. Ros, N. Pugeault, and N. Krüger. A compact harmonic code for early vision based on anisotropic frequency channels. *J Comput Vis and Image Underst*, 114(6):681–699, 2010.
14. J. D. Schall. Neuronal activity related to visually guided saccades in the frontal eye fields of rhesus monkeys: comparison with supplementary eye fields. *J Neurophysiol*, 66(2):559–79, 1991.
15. N. Sigala, F. Gabbiani, and N. K. Logothetis. Visual categorization and object representation in monkeys and humans. *J Cognitive Neurosci*, 14(2):187–98, 2002.
16. M. Teichmann, J. Wiltschut, and F. H. Hamker. Learning invariance from natural images inspired by observations in the primary visual cortex. *Neural computation*, 24(5):1271–96, 2012.
17. S. Treue and J. Trujillo. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736):575–579, 1999.
18. D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural networks*, 19(9):1395–407, 2006.
19. M. Wischnewski, A. Belardinelli, W. X. Schneider, and J. J. Steil. Where to Look Next? Combining Static and Dynamic Proto-objects in a TVA-based Model of Visual Attention. *Cognitive Computation*, 2(4):326–343, 2010.