Spatial Attention Improves Object Localization: A Biologically Plausible Neuro-Computational Model for Use in Virtual Reality

Amirhossein Jamalian Julia Bergelt Helge Ülo Dinkelbach
Fred H. Hamker
Artificial Intelligence, Department of Computer Science, Chemnitz University of Technology
Chemnitz, Germany
amirhossein.jamalian@informatik.tu-chemnitz.de

Abstract

Visual attention is a smart mechanism performed by the brain to avoid unnecessary processing and to focus on the most relevant part of the visual scene. It can result in a remarkable reduction in the computational complexity of scene understanding. Two major kinds of top-down visual attention signals are spatial and feature-based attention. The former deals with the places in scene which are worth to attend, while the latter is more involved with the basic features of objects e.g. color, intensity, edges. In principle, there are two known sources of generating a spatial attention signal: Frontal Eye Field (FEF) in the prefrontal cortex and Lateral Intraparietal Cortex (LIP) in the parietal cortex. In this paper, first, a combined neuro-computational model of ventral and dorsal stream is introduced and then, it is shown in Virtual Reality (VR) that the spatial attention, provided by LIP, acts as a transsaccadic memory pointer which accelerates object localization.

1. Introduction

At any given time, a vast volume of input data enters our visual system. The visual system has been evolved in such a way that it relinquishes most of them and only considers the important parts of the input data. Otherwise, the complexity of data processing would be too high. The mechanism of ignoring unrelated data and attending to only important parts of the visual field for the sake of faster scene understanding is called visual attention. Visual attention can be driven by two major classes of factors: bottom-up and top-down [9]. Bottom-up factors are involved with basic and complex features and are derived from the visual scene [17] whereas top-down ones are cognitive and mostly based on prior knowledge, expectations and goals [8].

The importance of the visual field parts could be related either to specific regions (spatial attention) or specific characteristics of the objects in it (feature-based attention). Basically, these two kinds of attention appertain to the top-down class and cognitively determine where and what should be attended, respectively. The brain generates corresponding attention signals from several regions, e.g. Frontal Eye Field (FEF), Lateral Intraparietal Cortex (LIP), Prefrontal Cortex (PFC).

In the literature of the neural processing of vision, it is widely accepted that two streaming pathways exist: ventral and dorsal stream [12]. The former is mostly involved in object recognition and contains feature-based attention signals while the latter is responsible for the object's spatial location and generates a spatial attention signal.

Heretofore, many computational models of visual attention have been developed [5] to perform specific tasks e.g. visual search [19], object recognition [4, 3], robot vision [15]. A former but still interesting survey and taxonomy of the visual attention models could be found in [11].

Although object recognition/localization tasks can be performed using only bottom-up process, top-down signals i.e. feature-based as well as spatial attention can facilitate the process and are more biologically plausible. In this work, we focus on the effect of spatial attention generated by the LIP region of the brain. Inspired from biology, it would be expected that the presence of spatial attention results in faster object localization. It has been proposed that LIP or related areas may encode an attention pointer to memorize the location of a task relevant object to inform areas involved in object identity about the location of the relevant object feature [7]. Importantly, such a pointer has to be updated with every saccade. The computational analysis of this object localization after a primary saccade is the core contribution of this paper. For this purpose, two previously developed models (one for ventral [14] and the other for dorsal stream [20]) have been combined with each other to consider the effect of spatial attention signal. The combined model is capable of performing object

recognition/localization, when attention requires spatial updating due to saccades (rapid eye movements). The whole model consists of several populations of rate-coded neurons which are implemented using the neural simulator ANNarchy [18]. The performance of the model has been evaluated in a Virtual Reality (VR) in which an agent can search for certain objects.

2. Model

This section provides a condensed description of the model. As mentioned before, the combined model consists of two separately-developed models: one for ventral and the other for dorsal stream. Each of them can be run solely to perform its specific tasks. However, to demonstrate the effect of spatial attention, generated by the dorsal stream model, on the object localization task, we have connected them to each other for the first time as illustrated in Figure 1.

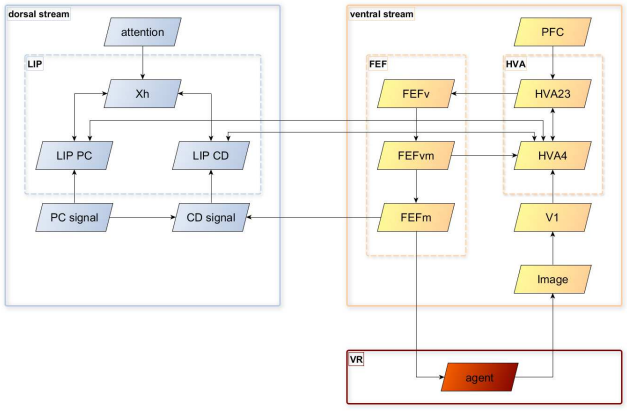

Figure 1. Structure of combined model. Left: The dorsal stream part of the model located in LIP with its four inputs eye position (PC), eye displacement (CD) and retinal image (both received from ventral stream) as well as spatial attention. Right: The ventral stream part which receives a visual image from VR and recognizes the place of searched object based on an HVA-FEF cycle with the aid of feature-based attention signal provided by PFC. HVA refers to a higher visual area, such as V4 or TEO, and HVA4 to the layer 4 and HVA23 to layer 2/3 in this area. Bottom: Virtual Reality (VR) with the agent providing the visual image for the model. Furthermore, the agent performs the eye movement to the location given by the frontal eye field.

The left part of the Figure 1, drawn by blue blocks, shows the dorsal stream model while the right part depicts the ventral model with yellow blocks. Although the FEF is not considered part of the ventral stream, it is historically strongly linked to attention in the ventral stream [13]. The red block illustrates the VR.

2.1. Ventral Stream

The ventral stream part of the combined model (the right block diagram in Figure 1) has been originally introduced in [3] based on a cortical microcircuit model developed in [2]. Its accuracy of object recognition/localization in the COIL-100 database [16] is 92% on black, 71% on noisy, and 42% on real-world backgrounds [3]. In a more recent article, it has been revised and tested in VR and the results corroborated its 85% exactitude [14]. In that revision, the model was robust against large object variations in visual search in which a human-like neuro-cognitive agent could recognize and localize 15 different objects regardless of scaling, point of view and orientation.

The input of the model is an RGB image (visual field) which is preprocessed to extract the basic features like oriented edges, red-green and blue-yellow color contrasts. The results of the preprocessing phase are assigned to the neurons in the V1 map from where they are routed to higher visual areas (HVA) like V4 or IT. The subsequent HVA4 neurons encode the object views via convolutions of receptive fields of V1 by a pre-generated weight matrix. This weight matrix can be obtained by an offline learning procedure described in [14]. The role of HVA23 is max-pooling as well as handling top-down feature-based attention signal which comes from PFC. It also propagates the feature-based signal back towards HVA4. In addition to the feature-based attention from HVA23, the HVA4 map receives spatial attention from the LIP maps of the dorsal stream.

The FEF region consists of three parts: FEF-Visual (FEFv), FEF-Visiomovement (FEFvm) and FEF-Movement (FEFm). FEFv can be considered as a saliency map because it contains the places where the target is probably located. FEFvm is responsible for focusing neuronal activity at the target location and the FEFm layer determines the final saccade target location. When the activities of the FEFm neurons reach a threshold, a saccade is triggered to this location. More details can be found in [14]. It is worth noting that FEF itself is an originate of spatial attention in the brain. However, although the FEF can also show sustained activation, the FEF may be more involved in the generation of spatial attention towards a potential saccade target. The maintenance of additional attention pointers may more likely involve the dorsal stream.

2.2. Dorsal Stream

The dorsal stream part of our combined model is based on the neuro-computational model of Ziesche and Hamker [20]. This model was primarily developed to explain the mislocalization of briefly flashed stimuli in total darkness, nonetheless they showed that the model is able to explain some further visual phenomena like saccadic suppression of displacement or masking [21, 1]. The model resides in the Lateral Intraparietal Cortex (LIP) and uses gain fields

as well as radial basis functions to perform coordinate transformation between eye- and head-centered reference frames. With the help of two extraretinal, eye position related signals, namely the proprioceptive (PC) eye position signal and the corollary discharge (CD) signal, as well as a retinal signal representing the (eye-centered) stimulus position, the model simulates the spatial position of a stimulus during eye movements in a head-centered reference frame. The structure of the model is shown in Figure 1, left side. In the combined model, the retinal signal originates from HVA4 of the ventral stream, as this map encodes the visual input from the retina in eye-centered coordinates and represents extrastriate visual areas like V4. The retinal signal is fed into two separated LIP maps where it is gain modulated by either the PC signal or the CD signal to obtain a joint representation of the stimulus position and the eye position or eye displacement, respectively. In our combined model, the CD signal is also received from ventral stream model, more precisely from the FEFm map, which triggers the saccade and can therefore be interpreted as corollary discharge signal. The two LIP maps interact with each other via an intermediate, head-centered map Xh using feedback projections. Those intermediate neurons combine the information from both LIP maps and as a result encode the perceived spatial position of a stimulus in a head-centered reference frame.

We use this model from Ziesche and Hamker with two small modifications: First, as the original model only covers a one-dimensional space, we adapt it to two dimensions by adding a second dimension representing the height to each input map. Second, we use the model in a top-down way. Originally, the model is proposed to transform the eye-centered retinotopic signal into a head-centered spatial position of the stimulus. But here, we use the model to transform a head-centered attentional pointer introduced through the intermediate neurons of Xh into an eye-centered one, which then feeds into the ventral stream. The head-centered attention pointer may encode spatial memory which is not explicitly modeled here.

As mentioned above, the dorsal stream part is not necessarily needed to generate a spatial attention signal for the ventral stream part. But using the LIP instead of FEF has several advantages. First of all, the spatial attention signal generated in LIP represents a memory pointer which only has to be updated to its new location and is thus available quickly after each saccade whereas the FEF needs significantly more steps to determine a spatial attention signal as it first requires a feature search. Therefore, HVA4 can be enhanced earlier which will lead to a faster object recognition. As the attention signal needed as input for the dorsal stream is head-centered, it can be easily received and maintained by some higher areas like Medial Temporal Lobe, where object memory takes place [6]. For instance, when entering a known room, the (head-centered) position of an object kept in memory can be used via the dorsal stream to locate the precise (eye-centered) position with the eyes.

3. Results

The performance of the combined model has been evaluated within a VR based on the game engine Unity [1]. The VR consists of an agent called Felice as well as different objects and provides images representing Felice's visual field. Once an object is localized by the model, the corresponding eye movement to this object is sent to the VR and executed by Felice.

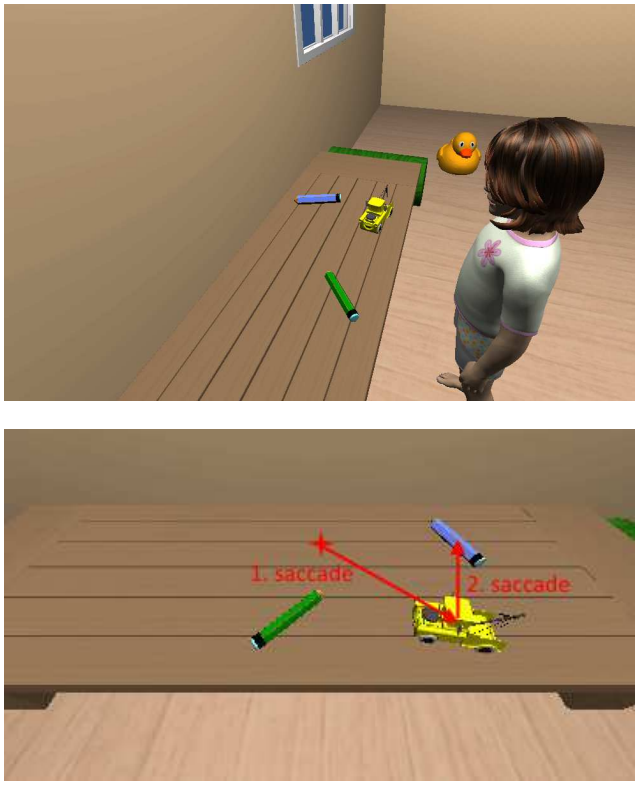

Figure 2. Layout of the simulated experiment in the VR. Top: Top view of the scenario with Felice and the three objects (yellow crane, blue and green pencil). Bottom: The visual field of Felice. The task is depicted in red: Fixating on the center of the visual field (red star), Felice has to search for the yellow crane, fixate on it and subsequently, she has to locate the blue pencil (red arrows).

Figure 2 illustrates the scenario (top) as well as the task (bottom) designed in the VR. The agent Felice is looking at the table with three different objects: a yellow crane, a blue pencil, and a green pencil. At the beginning, she fixates at the center of the visual field. Her task is to first locate the yellow crane and execute a saccade to its location. After the

[1] https://unity3d.com/

eyes landed, she has to search for a second object, namely the blue pencil, and fixate on it. In order to show the effect of spatial attention, we introduce a top-down attention signal to the blue pencil. The idea is that with spatial attention, the model will locate the second saccade object, i.e. the blue pencil, faster than without.

The simulation has been split into two phases: First, Felice has to find the blue pencil so that we can deploy top-down spatial attention to it. For this, we use the (standalone) ventral stream model performing an object localization task for the blue pencil. The position of the highest excitation of map HVA4 gives us the spatial position of the blue pencil in the visual field, which can be used as top-down spatial attention position for the dorsal stream model. The results of phase 1 are shown in Figure 3.

This spatial attention pointer is now deployed to the second phase of the simulation. Here, the combined model is used to perform a double-step object localization task. First, the yellow crane should be located and after fixating on it, the blue pencil should be found. To examine the effect of spatial attention we simulate the task once with the additional spatial attention signal and once without. We compare the activity of map HVA4, where the attentional signal operates, as well as FEFm, which is responsible for executing saccades. The simulation results are shown in Figure 4. The firing rates of HVA4 (blue) and FEFm (green) over time are plotted with spatial attention from LIP (solid lines) and without (dashed lines). The left plot shows the activity for the first localization until the saccade is execute. As can be seen, there are no differences in the activities of both maps, thus, the spatial attention has no effect on localizing the yellow crane. Table 1 contains the relevant time steps of the localization task for both simulations. For the first object, start and end of the localization process are the same. This is consistent as the spatial attention is deployed to a different object and should not enhance the localization of a non-attended object. However, for the second localization, there is a difference in the firing rates (Figure 4, right plot). Since the attended object and the searched object are the same now, the spatial attention signal increases the activity in HVA4 for this object. Consequently, the firing rates in FEFm at the position of the blue pencil increase faster and therefore, they reach the threshold for executing a saccade earlier which leads to an earlier saccade onset. As summarized in Table 1, the starting time of the second localization is equal for both simulations, but the blue pencil was found 19 ms earlier in the presence of spatial attention than its absence.

event	no spatial attention	with spatial attention
start search for yellow crane	0 ms	0 ms
localization of yellow crane	165 ms	165 ms
start search for blue pencil	262 ms	262 ms
localization of blue pencil	493 ms	474 ms

Table 1. Summary of timings for the second phase in absence and presence of spatial attention. The values indicate the time steps of the corresponding events. Spatial attention reduces the localization time for the blue pencil (attended object) by 19 ms while the localization time of the yellow crane (unattended object) remains the same.

4. Conclusion

In this paper, a combination of two biologically plausible models of ventral and dorsal stream has been introduced and evaluated in VR. The ventral stream part is responsible for object localization whereas the dorsal part performs the updating of spatial attention. In the combined model, the LIP part of the dorsal stream is connected to map HVA4 of the ventral stream model to deliver a spatial attention pointer which has a modulatory effect on activities of the HVA4 neurons. Therefore, the HVA4 neurons were excited more which resulted in faster object localization. The simulation results showed that the object can be localized 19 milliseconds faster in presence of the spatial attention signal than without it.

To further investigate the benefit of spatial attention generated by the dorsal stream model on the ventral stream model, one can conduct more experiments dealing with different questions:

- What happens, if there are more than one distractors or if the distractor is very similar to the target?
  The ventral stream model alone performs the object localization tasks, however, if the setup is too complicated due to too many or too similar objects, it might fail to find a given target object and may require a long serial search. In this case, the spatial attention from the dorsal stream model can help to locate the target. Anyway, this additional attention pointer should always improve the time needed for the object localization independent of number or type of objects.

- What happens for object localization tasks with more than two objects in a row, performing successively more than two saccades?
  Independent of the order of the target sequence, that

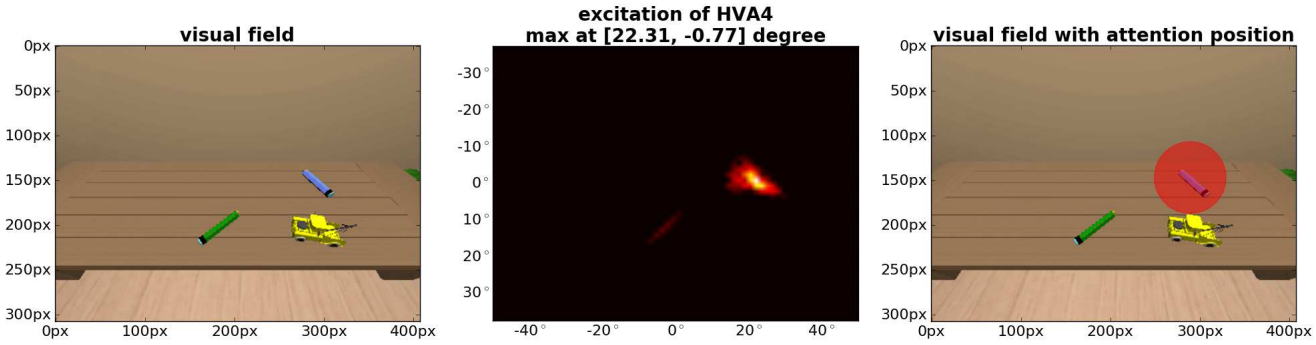

Figure 3. Simulation results of phase 1: For the given visual field (left), the ventral stream model performs an object localization task for the blue pencil. The position of highest excitation of HVA4 (middle) gives us the spatial position of the attention pointer in the visual field (red circle in the right image).

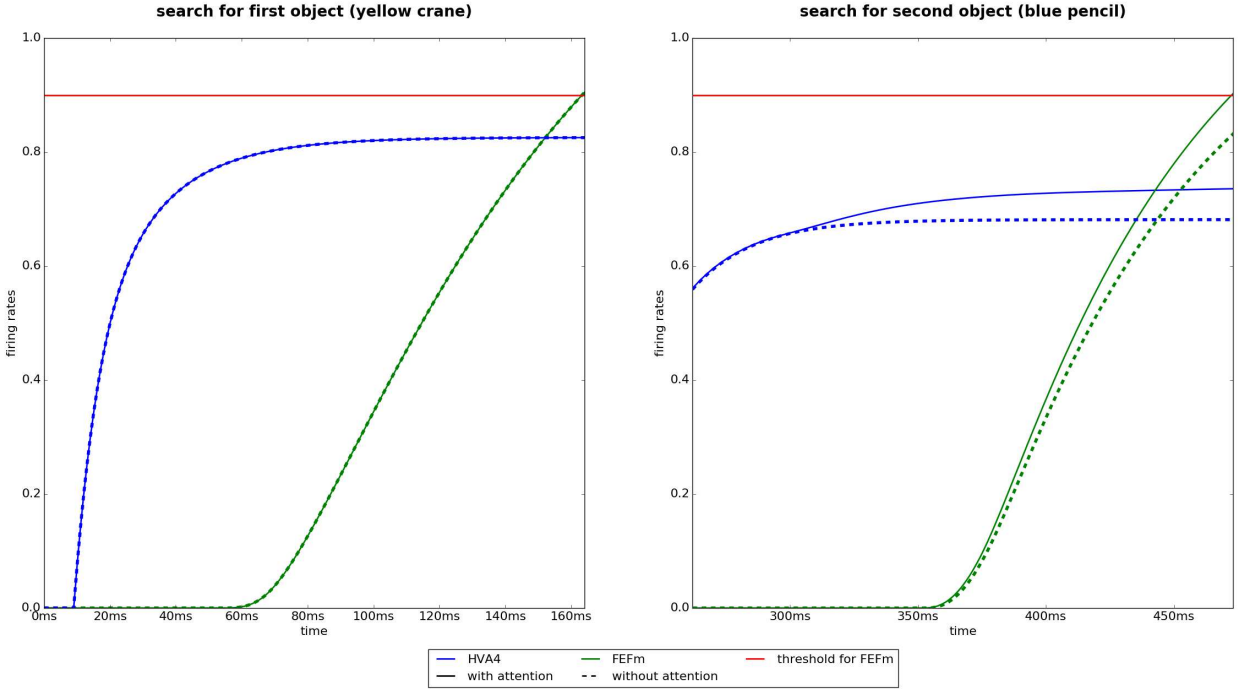

Figure 4. Simulation results of phase 2. Firing rates of HVA4 (blue) and FEFm (green) for first (left) and second (right) object localization. Firing rates modulated by spatial attention are plotted in solid lines, firing rates from simulation without spatial attention are plotted with dashed lines. Additionally, we illustrate the threshold for FEFm for executing a saccade with a red line.

means when exactly the attended object will be located, the dorsal stream model should update the attention position of this attended object until it is the target for the search task and facilitate a faster object localization. Meanwhile, the localization time for other objects should not be affected.

- What happens if more than one object should be attended? [10]

The dorsal stream model should be able to update more than one attention position either simultaneously or serially. Thus, the object localization should be faster for all attended objects independent of the number of attended objects, if the number of attended object is limited to plausible values.

Acknowledgments

This work has been supported by the European Unions Seventh Framework Programme (FET, Neuro-Bio-Inspired Systems: Spatial Cognition) under Grant Agreement No. 600785 and by the Federal Ministry of Education and Research within the grant "US-German collaboration on computational neuroscience" (BMBF 01GQ1409).

References

[1] J. Bergelt and F. H. Hamker. Suppression of displacement detection in the presence and absence of eye movements: A neuro-computational perspective. *Biological Cybernetics*, 110:81–89, 2016.

[2] F. Beuth and F. H. Hamker. A mechanistic cortical microcircuit of attention for amplification, normalization and suppression. *Vision Research*, 116:241–257, 2015.

[3] F. Beuth and F. H. Hamker. Attention as cognitive, holistic control of the visual system. *Proc Workshop New Challenges in Neural Computation 2015 - NCNC 2015, Machine Learning Reports 03/2015*, pages 133–140, 2015.

[4] F. Beuth, J. Wiltschut, and F. Hamker. Attentive Stereoscopic Object Recognition. In T. Villmann and F.-M. Schleif, editors, *Workshop NCNC2010*, page 41, 2010.

[5] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IIEEE Trans Pattern Anal Mach Intell*, 35(1):185–207, jan 2013.

[6] P. Byrne, S. Becker, and N. Burgess. Remembering the past and imagining the future: a neural model of spatial memory and imagery. *Psychological review*, 114(2):340–375, 2007.

[7] P. Cavanagh, A. R. Hunt, A. Afraz, and M. Rolfs. Visual stability based on remapping of attention pointers. *Trends in cognitive sciences*, 14(4):147–153, 2010.

[8] M. Corbetta and G. L. Shulman. Control of Goal-Directed and Stimulus-Driven Attention in the Brain. *Nature Reviews Neuroscience*, 3(3):215–229, 2002.

[9] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Reviews of Neuroscience*, 18:193–222, 1995.

[10] L. Dugué, D. McLelland, M. Lajous, and R. VanRullen. Attention searches nonuniformly in space and in time. *Proceedings of the National Academy of Sciences of the United States of America*, 112(49):15214–15219, 2015.

[11] S. Frintrop, E. Rome, and H. I. Christensen. Computational Visual Attention Systems and Their Cognitive Foundations: A Survey. *ACM Transactions on Applied Perception*, 7(1):1–39, 2010.

[12] M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992.

[13] 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):431–447, 2005.

[14] A. Jamalian, F. Beuth, and F. H. Hamker. The Performance of a Biologically Plausible Model of Visual Attention to Localize Objects in a Virtual Reality. In *Proceedings of International Conference on Artificial Neural Networks (ICANN2016)*, volume 9887 LNCS, pages 447–454. Springer, Barcelona, 2016.

[15] A. Jamalian and F. H. Hamker. Biologically-Inspired Models for Attentive Robot Vision: A Review. *Innovative Research in Attention Modeling and Computer Vision Applications*, pages 69–98, 2016.

[16] S. A. Nene, S. K. Nayar, and H. Murase. Columbia Object Image Library (COIL-100). *Technical Report CUCS-006-96*, 1996.

[17] H.-C. Nothdurft. Salience of Feature Contrast. In *Neurobiology of Attention*, chapter 38, pages 233–239. Academic Press, 2005.

[18] J. Vitay, H. Ü. Dinkelbach, and F. H. Hamker. ANNarchy: a code generation approach to neural simulations on parallel hardware. *Frontiers in Neuroinformatics*, 9(19):1–20, 2015.

[19] J. M. Wolfe. Guided Search 2 . 0 A revised model of visual search. *Psychnomic Bulletin & Review*, 1(2):202–238, 1994.

[20] A. Ziesche and F. H. Hamker. A Computational Model for the Influence of Corollary Discharge and Proprioception on the Perisaccadic Mislocalization of Briefly Presented Stimuli in Complete Darkness. *Journal of Neuroscience*, 31(48):17392–17405, 2011.

[21] A. Ziesche and F. H. Hamker. Brain circuits underlying visual stability across eye movements – converging evidence for a neuro-computational model of area {LIP}. *Frontiers in Computational Neuroscience*, 8(25):1–15, 2014.