

Integrating Inhomogeneous Processing and Proto-object Formation in a Computational Model of Visual Attention

Marco Wischnewski, Jochen J. Steil, Lothar Kehler and Werner X. Schneider

Abstract We implement a novel computational framework for attention that includes recent experimentally derived assumptions on attention which are not covered by standard computational models. To this end, we combine inhomogeneous visual processing, proto-object formation, and parts of TVA (Theory of Visual Attention [2]), a well established computational theory in experimental psychology, which explains a large range of human and monkey data on attention. The first steps of processing employ inhomogeneous processing for the basic visual feature maps. Next, we compute so-called proto-objects by means of blob detection based on these inhomogeneous maps. Our model therefore displays the well known "global-effect" of eye movement control, that is, saccade target landing objects tend to fuse with increasing eccentricity from the center of view. The proto-objects also allow for a straightforward application of TVA and its mechanism to model task-driven selectivity. The final stage of our model consists of an attentional priority map which assigns priority to the proto-objects according to the computations of TVA. This step allows to restrict sophisticated filter computation to the proto-object regions and thereby renders our model computationally efficient by avoiding a complete standard pixel-wise priority computation of bottom-up saliency models.

Marco Wischnewski
Center of Excellence - Cognitive Interaction Technology (CITEC) and Neuro-cognitive Psychology, Bielefeld University, e-mail: marco.wischnewski@cit-ec.uni-bielefeld.de

Jochen J. Steil
Research Institute for Cognition and Robotics (CoR-Lab), Bielefeld University, e-mail: jsteil@cor-lab.uni-bielefeld.de

Lothar Kehler
Neuro-cognitive Psychology, Bielefeld University, e-mail: lothar.kehrer@uni-bielefeld.de

Werner X. Schneider
Neuro-cognitive Psychology and Center of Excellence - Cognitive Interaction Technology (CITEC), Bielefeld University, e-mail: wxs@uni-bielefeld.de

1 Introduction

With the advent of increasingly cognitive robots using active vision, computational models for guidance of their visual attention become a standard component in their control architectures [1, 6, 12, 16]. Many of these models are driven by the ambition to realize human-like perception and action and rely on bottom-up computation of features. The most influential model has been proposed by Itti & Koch [10]. It computes a pixel-wise saliency map that shifts attention to the location with the highest saliency value. However, current computational models are not able to explain classical human data on covert visual attention (e.g., from "visual search" or "partial report paradigms") with the same degree of specification and predictive power as psychological models [2, 4, 15, 18, 21]. In this paper, we argue that progress is possible by relying on one of the most sophisticated theoretical frameworks of experimental psychology and cognitive neuroscience, the Theory of Visual Attention (TVA, developed by Bundesen [2]). TVA explains essential empirical findings of human visual attention [3] and has a neurophysiological interpretation fitting findings from all major single cell studies on attention [4]. TVA can serve as a pivotal element in computational modeling of visual attention because it both allows for simple weighting of low level feature channels and proposes mechanisms for object-based task-driven control of processing resources.

TVA differs from most existing bottom-up saliency models in proposing that saliency (priority) is not computed pixel-wise, but rather entity-wise based on perceptual units. These units are competing elements of an attention priority map (APM). While TVA itself does not specify how these units are formed, we will assume that the perceptual units can be described as so-called proto-objects. Proto-objects are formed within the APM and they refer to homogeneous regions in low-level feature maps, which can be detected without sophisticated object recognition. There are some other recent models which rely on proto-object formation [13, 17], but not in connection with TVA. One extension of the classical Itti & Koch model [19] forms proto-objects around the maxima of the saliency map. In contrast to our model, it assumes that saliency has been already determined before forming the proto-object postattentively.

Our model gains further biological und psychological plausibility by implementing inhomogeneous low-level feature processing which is lacking in most recent models (e.g. [19]). It is based on detailed findings about processing of retinal information in early stages of the visual cortex [20]. In contrast, most standard bottom-up attention-models operate pixel-wise in the visual field and it makes no difference whether an object (or a feature) appears foveally or in the periphery. Therefore, they cannot explain classical effects that demonstrate the inhomogeneous nature of visual processing such as the well-known "global effect" [7] of eye movement control. Saccadic eye movements to two nearby objects tend to land within the center-of-gravity of these objects given eye movements have to be made under time pressure. Given spatial proximity of the two objects, our model computes in this case one common proto-object. Importantly, this averaging effect increases spatially with increasing retinal eccentricity. Due to the inhomogeneity of the feature maps, our model shows

this effect because proto-object computation in the periphery allows fewer candidate regions (proto-objects) to survive as individuated single objects.

Our model implements the whole path from visual input up to the APM (see Fig. 1). Incipient with the input image, we compute the inhomogeneous feature maps for color and luminance of one selected filter level for determining the proto-objects. This selection is motivated by the finding in human experiments which suggest that under time pressure only filters up to a certain resolution level enter the computation [11]. At this stage, costly computation of all Gabor filters is still delayed until information about the proto-objects regions is available. Further, only for the proto-object regions all filters are computed and summed according to the TVA equations, which allow for a task specific weighting of feature channels. Based on these computations, the attentional priority map (APM) according to TVA is formed. We assume that the APM serves as linkage between the ventral ("what") and the dorsal ("where") pathway. Proto-object computation is performed by the dorsal pathway while attentional priorities for proto-objects are computed within the ventral pathway (e.g. [15]). The subsequent sections guide along the path illustrated in Fig. 1, incipient with the input image up to the APM. Examples illustrate the properties of our model in terms of the global-effect.

2 Inhomogenous Retinal/V1 Processing

To comply with the inhomogeneous density of photoreceptors in the human retina [14], the homogeneous pixel grid input image (e.g. from a robot's camera) is transformed into an inhomogeneous pixel grid which serves as input for all subsequent filter operations (see Fig. 2, left). The grid positions ("receptors") are computed in the same way as the positions of the subsequent filters, but to cope with the Nyquist-Theorem, the density is doubled in relation to the filter layer with the highest density. The inhomogeneous feature maps are based on a biological driven mathematical description of V1 bar and edge Gabor filters developed by Watson [20]. The inhomogeneity of the filter structure is defined by the following relations: With increasing angle of eccentricity (a) the filters' size and (b) the distance between adjacent filters increases, whereas (c) the filters' spatial frequency decreases. The scaling s is linear with respect to a scaling parameter k (see. Eq. (1)), where e is the angle of eccentricity in degree. In the human visual system, k is estimated around 0.4 [20].

$$s = 1 + k * e \quad (1)$$

According to Watson, all subsequent parameters are computed as follows: At the center of the visual field, with $e = 0$, there is a central filter with $s = 1$. This filter has a given central spatial frequency f_{center} . The size of the filter is assigned by its width at half height: $w_{center} = 1.324/f_{center}$. With increasing eccentricity, concentric rings of filters are generated recursively. The parameters of an inner ring (or the central filter) determine the distance to the next outer ring as well as the distance

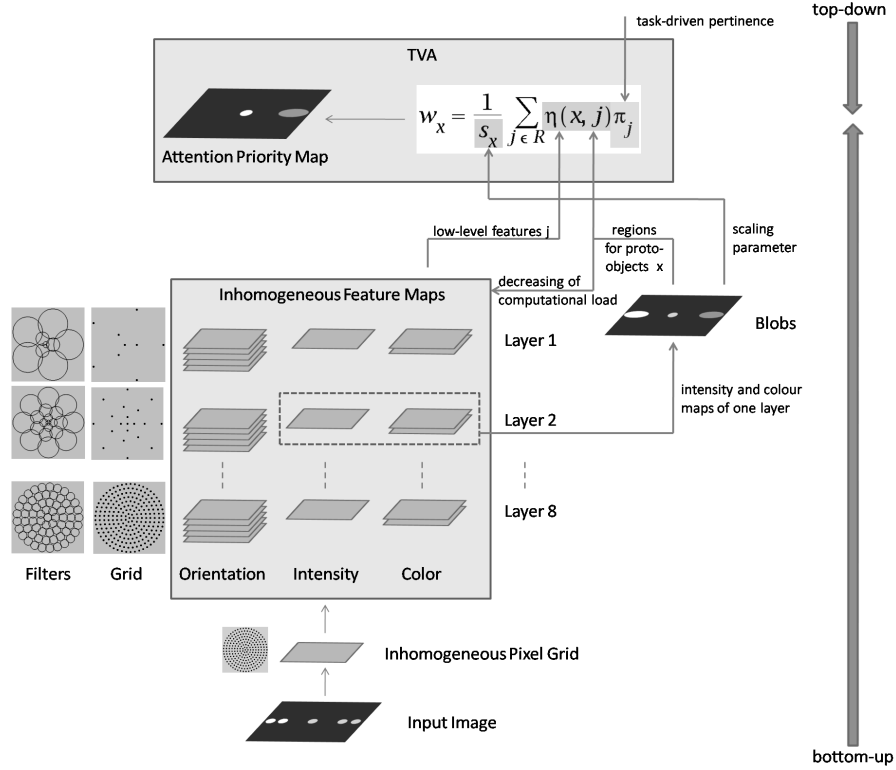


Fig. 1 The figure illustrates the structure of the model. A TVA-driven attention priority map (APM) results from top-down (pertinence) and bottom-up (proto-objects and feature maps) processes. In this example, the peripherally located proto-objects show the global effect. Furthermore, the proto-object on the left side disappears in the APM due to being task-irrelevant. Finally, the attentional weight of the proto-object on the right side is downscaled according to its high angle of eccentricity.

(and thereby the number) of the filters within this outer ring: $d = 1/(f * 1.33)$. The frequency and width for each filter are adjusted as $f = f_{center}/s$ and $w = w_{center} * s$. Consequently, there can be added as many rings as necessary to cover a desired area of the visual field (see Fig. 2, right).

Each Gabor filter consists of a cosine function overlaid by a Gaussian function where θ represents the angle and ϕ the phase (2).

$$f(x, y) = \exp\left(\frac{4 \ln(2)(x^2 + y^2)}{w^2}\right) \cos(2\pi f(x \cos \theta + y \sin \theta) + \phi) \quad (2)$$

For each filter, the orientation is varied fivefold (0° , 36° , 72° , 108° and 144°) and the phase twice (0° and 90°). For each orientation, the filters' outcome of both phases (bar and edge filter) is combined to obtain the locally shift invariant output [9].

This results in five orientation feature maps (each represents one orientation) for one filter structure given a central frequency. To cover the whole human frequency

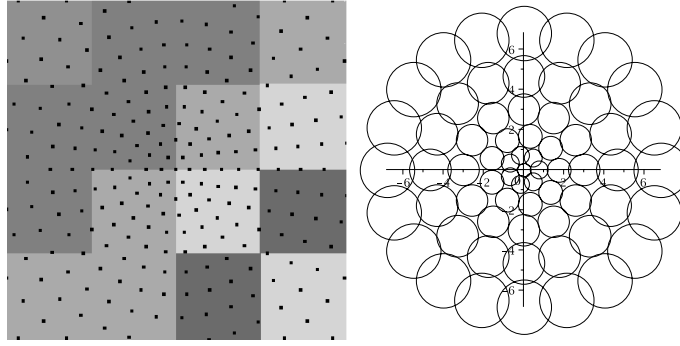


Fig. 2 Left: The figure shows an input image centric section of 4x4 pixel. The black dots match the inhomogeneous pixel grid positions. Right: The first five rings surrounding the central filter with $f_{center} = 1$ and $k = 0.4$. The axes of abscissae and ordinate reflect the angle of eccentricity. For illustration, the filter's size was reduced.

space, it is necessary to layer these structures. Thus, the model consists of 8 layers in which the first layer starts with a center frequency of 0.25 cyc/deg. From layer n to layer $n + 1$, this center frequency is doubled, so that at the end layer eight has a center frequency of 32 cyc/deg. We obtain 40 feature maps (8 layer each with 5 orientation feature maps). Note that, however, the full filter bank of Gabor filters needs to be computed only for grid position comprising proto-objects, which are determined based on the color and intensity filters for a certain selected resolution alone.

For the color and intensity feature maps, the described filter structure is adopted, but filters are restricted to the Gaussian part in (2). The color feature maps rely on the physiological RG/BY space [19]. Again, 8 layers are computed, so there are 8 intensity and 16 color feature maps (8 RG and 8 BY). In sum, we obtain 64 feature maps. Again note that they have to compute in full scale only for the proto-object regions.

3 Proto-object Formation

For blob detection we use an algorithm developed by Forssén [8]. It makes use of channel representations within the three-dimensional color/intensity space and thereby allows for spatial overlapping of resultant blobs, which are spatial homogeneous regions approximated by ellipses. The intensity and color feature maps of one layer serve as input. The choice of the layer simulates to what extent the system is under time pressure, because high-resolution layers need more time for processing. Thus, a high degree of time pressure yields a low-resolution layer as input and thereby merging of objects into one proto-object is observable (global-effect). In order to utilize the blob algorithm, the inhomogeneous pixel grid of the feature maps

is transformed back into voronoi cells on a homogeneous pixel grid (see Fig. 3, c). The size of the homogeneous pixel grid is, layer-independent and constant, as large as being necessary to avoid an overlapping of back transformed pixel even if the layer with highest center frequency was chosen.

Due to the peripherally increasing size of the voronoi cells, we included a filtering mechanism. That is, every blob has to have at least a size of $n * v$ in both of its axes where v is the size of the voronoi cell containing the blob's centroid and n is a scaling factor with $n > 1$. Thus, depending on factor n , a minimum number of n^2 cohering voronoi cells are necessary to build a blob, which works uncoupled with respect to the angle of eccentricity.

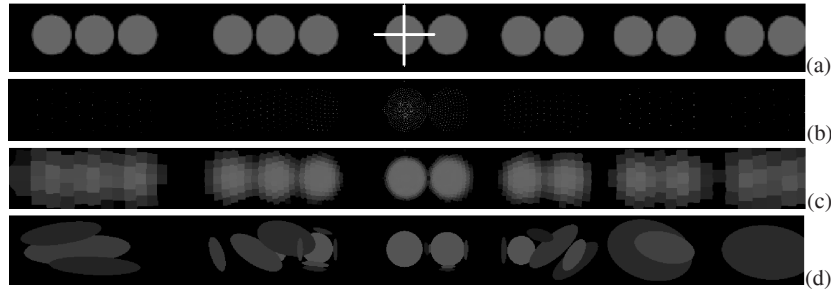


Fig. 3 The figure shows (a) the input image in which the white cross marks the foveal center, (b) the inhomogeneous pixel grid, (c) the voronoi cells, and (d) the blobs. To demonstrate the global effect, circles combined to groups of two and three are used. The blobs in (d) show the change from the fovea to the periphery: Whereas foveally positioned blobs represent rather accurate the circles of the input image, peripherally positioned blobs tend to represent more circles and to be more inaccurately. The intensity of each blob reflects the average intensity of the channel's region.

4 TVA

In TVA, each proto-object x within the visual field has a weight which is the outcome of the *weight equation* (3). A spatial structured representation of all these weights is called the *attentional priority map* (APM). Each w_x -value is computed as the sum over all features which are element of R . The $\eta(x, j)$ -value indicates the degree of proto-object x having feature j weighted by top-down task-dependent controlled pertinence π_j (e.g. search for a red object). Thus the $\eta(x, j)$ -value restricts feature computation to proto-object regions, while π_j implement a standard feature channel weighting as also present in other saliency models.

$$w_x = \sum_{j \in R} \eta(x, j) \pi_j \quad (3)$$

At this point, all needed bottom-up data are available to compute the η -values: The region of each proto-object as found by the blob detection algorithm, and the features computed within this region as determined in the inhomogeneous feature maps.

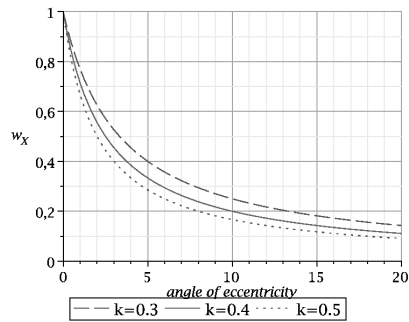
If two proto-objects p_1 and p_2 are completely equal in the size of their regions and the values of the feature maps within these regions and vary only in the angle of eccentricity, then the more foveally located proto-object gets a higher attentional weight. A computational expedient way to integrate this relation is the implementation of an *inhomogeneity parameter* s_x which represents the scaling of proto-object x depending on its angle of eccentricity (4, right). The mathematical embedding of s_x leads to a modified weight equation (4, left). Thus, if it is the case that $s_1 = 2s_2$, then p_1 's region has to be double in size to get the same attentional weight as p_2 .

$$w_x = \frac{1}{s_x} \sum_{j \in R} \eta(x, j) \pi_j \quad \text{with} \quad s_x = 1 + k * e_x \quad (4)$$

5 Results

How does the global effect emerge from our computational architecture? Fig. 5 shows the result of the blob-algorithm to determine the proto-objects of layer 2 to 8 depending on the hyperparameter k . Layer one was skipped because it produces no blobs. Layers with lower frequency, whose choice was motivated by time pressure for saccadic eye movements, produce the global effect. This means, the visual system cannot distinguish adjacent objects in consequence of the low spatial filter resolution. Therefore, these objects fuse together to a "surrounding" proto-object and a saccadic eye movement lands in the center-of-gravity of these objects within the visual field which roughly equals the center of this proto-object.

Fig. 4 The figure illustrates the influence of the inhomogeneous parameter s_x on the attentional weight w_x for different k -values. The axis of abscissae reflects the angle of eccentricity and the axis of ordinates the attentional weight w_x .



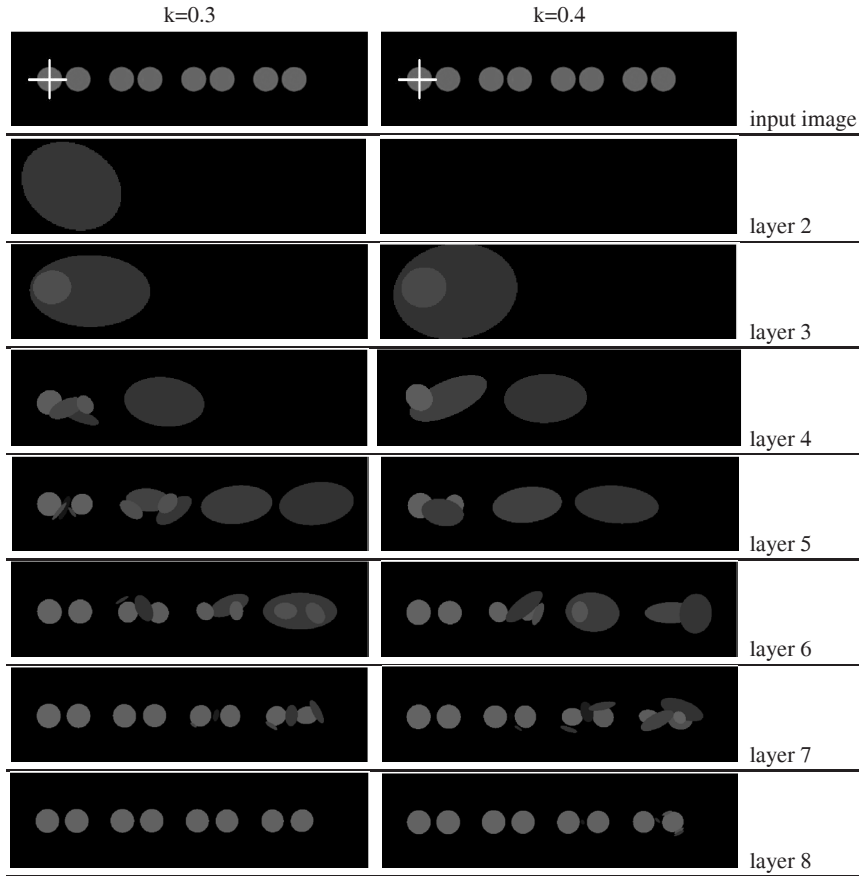


Fig. 5 The figure shows the resulting blobs depending on feature layer and parameter k . On the top, the input image is shown. The white cross marks the foveal center and the intensity of each blob reflects the average intensity of the channel's region.

Layers with lower frequency also produce, due to the lower spatial resolution, larger blobs. The lower the frequency, the more objects within the periphery of the visual field are not represented by a proto-object. They simply disappear. The k parameter is a hyperparameter for scaling these effects within a layer, because decreasing the k -value leads, according to the scaling function (1), to a slower decrease of the filter density from the foveal center to the periphery. We choose k motivated by the finding in the human visual system [20].

Applied to images consisting of natural objects our model yields psychological feasible results (see Fig. 6). Peripherally located objects tend to be represented by only one proto-object or even disappear, whereas more foveally located objects tend to produce proto-objects which represent parts of them.

Finally, Fig. 4 shows the influence of the inhomogeneous parameter s_x on the attentional weight w_x . If we assume a proto-object with $\sum \eta(x, j) * \pi_j = 1$, the figure

illustrates the outcome of the modified TVA weight equation (4) which then only depends on s_x . This is shown for different k -values. The higher the k -value, the stronger the angle of eccentricity affects the attentional weights.

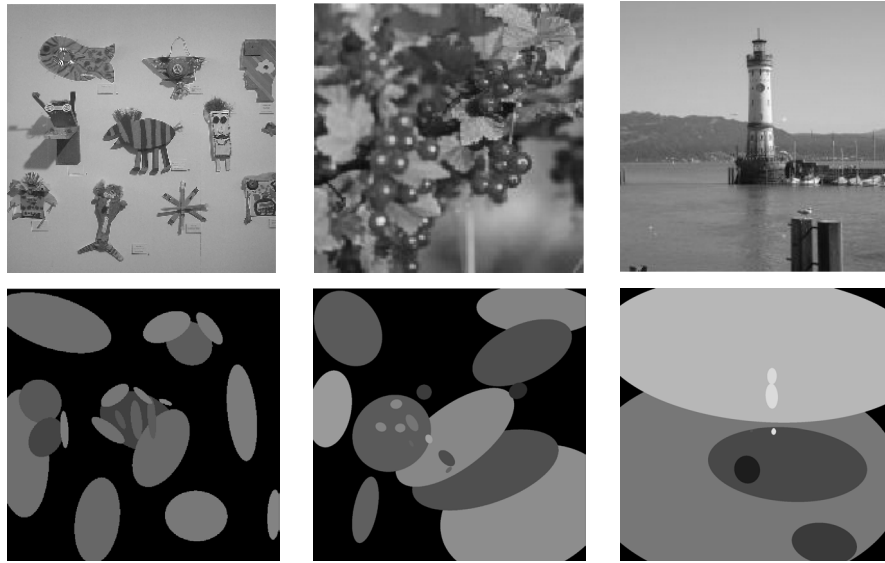


Fig. 6 Proto-objects formed on the basis of inhomogeneous processing in natural images with $f_{center} = 4$ and $k = 0.4$.

6 Outlook

We attempted to show in this paper that the combination of inhomogeneous processing, proto-object formation and TVA leads to new and interesting forms of controlling overt and covert visual attention [5]. Moreover, proto-object computation allows to include sophisticated task-driven control of visual attention according to TVA. Furthermore, this approach provides the possibility to reduce computational load. Only those features from the feature maps (orientation, intensity and color) which are located in proto-object regions have to be computed for the η -values of TVA. These bottom-up η -values are multiplicatively combined with top-down pertinence (task) values and result in proto-object based attentional weights. Future computational research is needed in order to evaluate the potential of this TVA-based task-driven form of attentional control for robotics, psychology and cognitive neuroscience.

References

1. Breazeal C, Scassellati B (1999) A context-dependent attention system for a social robot. In: Proc. 16th ICJAI, 1146-1153
2. Bundesen C (1990) A theory of visual attention. *Psych. Rev.* 97:523-547
3. Bundesen C, Habekost T (2008) Principles of visual attention. Oxford University Press, Oxford
4. Bundesen C, Habekost T, Killingsbaek S (2005) A neural theory of visual attention: bridging cognition and neurophysiology. *Psych. Rev.* 112:291-328
5. Deubel H, Schneider WX (1996) Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vis. Res.* 36:1827-1837
6. Driscoll JA, Peters II RA, Cave KR (1998) A visual attention network for a humanoid robot. In: Proc. IEEE/RSJ IROS1998, 12-16
7. Findlay JM (1982) Global processing for saccadic eye movements. *Vis. Res.* 22:1033-1045
8. Forssén PE (2004) Low and medium level vision using channel representations. Dissertation No. 858, ISBN 91-7373-876-X
9. Fogel I, Sagi D (1989) Gabor filters as texture discriminator. *Biol. Cybern.* 61:103-113
10. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. In: IEEE ICCV, 195-202
11. Kehler L (1989) Central performance drop on perceptual segregation tasks. *Spatial Vision* 4:45-62
12. Nagai Y, Hosoda K, Morita A, Asada M (2003) A constructive model for the development of joint attention. *Connection Science*, 15:211-229
13. Orabona F, Metta G, Sandini G (2007) A Proto-object based visual attention model. In: L. Paletta and E. Rome (ed) WAPCV 2007, LNAI 4840, 2007. Springer-Verlag, Berlin Heidelberg, pp. 198-215
14. Palmer SE (1999) *Vision Science*. The MIT Press, Cambridge, Massachusetts, pp. 29-31
15. Schneider WX (1995) VAM: A neuro-cognitive model for visual attention control of segmentation, object recognition, and space-based motor action, *Vis. Cog.* 2:331-375
16. Steil JJ, Heidemann G, Jockusch J, Rae R, Jungclaus N, Ritter H. (2001) Guiding attention for grasping tasks by gestural instruction: the GRAVIS-robot architecture. In: Proc. IEEE/RSJ IROS2001, 1570-1577
17. Sun Y, Fisher R, Wang F, Gomes HM (2008) A computer vision model for visual-object-based attention and eye movements. *Computer Vision and Image Understanding* 112:126-142
18. Treisman AM (1988) Features and objects: the fourteenth Bartlett memorial lecture. *Quarterly Journal of Experimental Psychology*, 40A, 201-237
19. Walther D, Koch C (2006) Modeling attention to salient proto-objects. *Neural Networks* 19:1395-1407
20. Watson AB (1983) Detection and recognition of simple spatial forms. In: Braddick OJ, Sleigh AC (eds) *Physiological and biological processing of images*. Springer, Berlin Heidelberg New York, pp. 100-114
21. Wolfe JM (1994) Guided search 2.0: a revised model of visual search. *Psychonomic Bulletin and Review*, 1:202-238