Attentive Stereoscopic Object Recognition

Frederik Beuth, Jan Wiltschut, and Fred H. Hamker

Chemnitz University of Technology,
Strasse der Nationen 62, 09107 Chemnitz, Germany
frederik.beuth@cs.tu-chemnitz.de,wiltschj@uni-muenster.de,
fred.hamker@cs.tu-chemnitz.de
http://www.tu-chemnitz.de/cs/KI/

Abstract. Object recognition in general is still a challenging task today. Problems arise for example from parallel segmentation and localization or the problem to detect objects invariant of position, scale and rotation. Humans can solve all these problems easily and thus neuro-computational and psychological data could be used to develop similar algorithms. In our model, attention reinforces the relevant features of the object allowing to detect it in parallel. Human vision also uses stereoscopic views to extract depth of a scene. Here, we will demonstrate the concept of attention for object recognition for stereo vision in a virtual reality, which could be applied in the future to practical use in robots.

Keywords: Object Recognition, Attention, Stereo Vision, Learning

1 Introduction

Object recognition is the task to recognize and additionally localize a searched object in an image or a scene. Many neuro-computational models, like Neocognitron [5] and HMAX [15, 19] filter the image over different stages to reduce the complexity of the filter operations. These systems are purely forward driven and do not consider the concept of attention.

Object recognition combined with attention can solve the dilemma of parallel segmentation and localization. We will first explain the concept of attention and how it solves this problem. We use a stereoscopic edge and depth detection model to achieve stereo object recognition. The object detectors are learned unsupervised and use a neuro-computational model which capture the basic principles of primate 3D perception. We will first focus on position invariant recognition and then demonstrate the ability to discriminate different objects. The model and the results are compared to neuro-computational findings.

1.1 Concept of Attention

Early concepts of visual attention define attention as to focus processing on a spatially determined part of the image, namely the spotlight of attention. The location of interest is typically determined from conspicuous or salient image features forming the saliency map [8, 10].

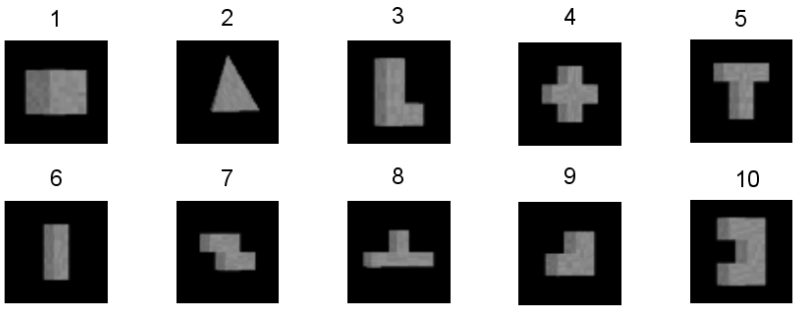


Fig. 1. The stimuli consist of 10 different 3D objects.

Recently, the "spotlight of attention" concept has been expanded to a feature-based approach [6] in which attention emerges from interactions between different brain areas. High level areas hold a template to specify the searched object and this information is propagated backwards to lower level areas. The parallel computation modifies the conspicuity of each descriptor in the system in such a way that the value represents the accumulated evidence. We implement the concept of attention as a modulating of the feed forward signals (called gain-control) dependent on the feed back from higher cortical areas. To perceive an object, a combination of several distributed visual features is required. Such binding processes can be well described by concepts of visual attention, illustrated by two continuous sub processes. The first one operates in parallel over all features and increases the conspicuity of those that are relevant for the searched object, independent of their location in the visual scene. The other subprocess is linked to action plans, e.g. eye movement plans, and combines those fragments which are consistent with the action plan, typically by their spatial location in the visual scene.

Object Selection and Segmentation For recognizing a searched object in a scene, the object must first be located and segmented, which however is only possible if the object has been recognized as such. Attention can solve this "Chicken-egg-problem" due to its parallel computation approach.[6]

2 Object Recognition System

2.1 Neuronal network architecture

We extend the concept of a population-based object representations [6] by learnable object representation based on local edge detectors. This allows to detect objects depending on their shape or texture. Additionally, we demonstrate the approach on stereoscopic images.

In our neuronal model (Fig. 2), we do not consider all the complexity of the visual stream. Rather we simulate an earlier area (V1) and a high level area

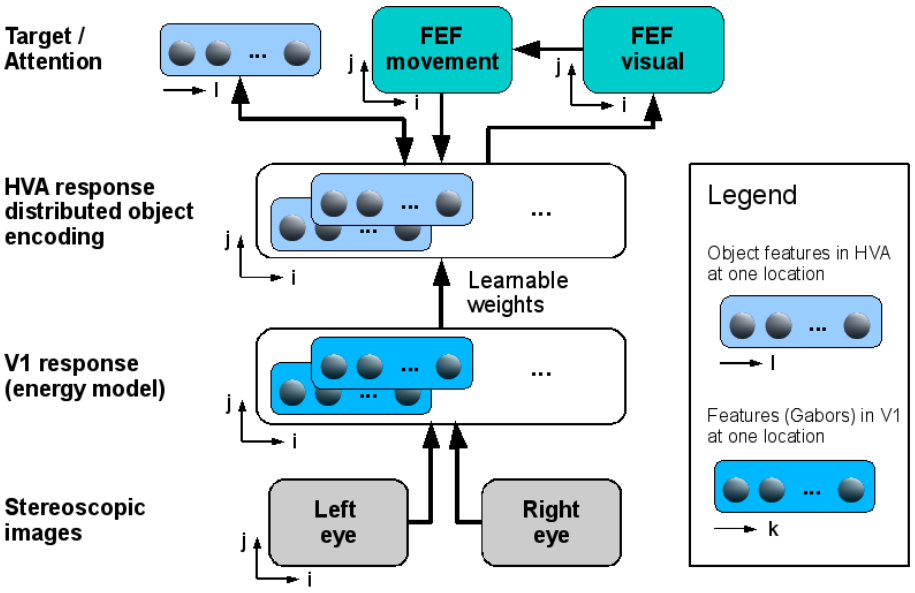


Fig. 2. Neuronal network of the stereoscopic object recognition model. The i and j indices correspondent to the spatial x and y axis of the images. The index k refers to different Gabor responses and l to different learned features in HVA.

(HVA) whose object selective cells can be mapped to area V2/V4/IT. As input stimuli we use the left and right eye view of 10 different 3D objects (Fig. 1), produced by a raytracer engine [1]. The objects are to some degree similar in their edges and thus the difficulty of the problem is comparable to cluttered scenes. The first area detects stereoscopic edges and disparities via an energy model (see [11, 13, 14]) and is comparable to area V1. This particular filter bank [17] uses 56 Gabors with 8 orientations (with a $\frac{\pi}{8}$ step size) and 7 different phase disparity shifts (with $\frac{\pi}{4}$ step size). This area builds a representation of the scene encoding edge informations, independent of the right or left view and therefore enables stereo object recognition. Overlapping receptive fields serve as input for the object selective cells of the HVA. We achieve the object selectivity by learning the feed forward weights (V1$\rightarrow$ HVA) with a biological motivated learning algorithm and a trace rule (see 2.3). The attention signal stores the features relevant for the current task. The Frontal Eye Field (FEF) consists of two areas, the saliency map (called FEFvisual) and the target of the next eye movement (called FEFmovement). One of the binding processes operates over all locations in HVA and reinforces the features of the searched object. The other is achieved by the loop over FEFvisual and FEFmovement and reinforces adjacent locations. Both processes use competition to decrease the activity of irrelevant features and location in HVA.

2.2 Neuron model

We use a rate coded neuron model which describes the firing rate r of a cell as its average spike frequency. Every cell represents a certain feature (V1: k, HVA: l) at a certain location (i, j). In the following we will omit the location indices for clarity. Consider one location in HVA, each cell in HVA gains excitation (as a weighted sum) from cells of V1 within the receptive field (here a 14x14 patch) and each cell is inhibited by all other HVA cells via Anti-Hebbian inhibition (similar as in [23]).

$$\tau_R \frac{\partial r_l}{\partial t} = \sum_i w_{kl} \cdot r_k^{\text{Input}} - \sum_{l',l' \neq l} f\left(c_{l,l'} \cdot r_{l'}\right) - r_l \quad \text{with: } f(x) = d_{nl} \cdot \log\left(\frac{1+x}{1-x}\right) \quad (1)$$

$f(x)$ gives the non-linear processing. τ_R is the time constant of the cells. The connection $w_{k,l}$ denotes the strength of the feed forward weight from input cell k to the output cell l. Lateral inhibition is given by the connection weight $c_{l,l'}$ and can differ across the cells due to the Anti-Hebbian learning.

2.3 Learning of the object descriptors

Changes in the connection strength between neurons in response to appropriate stimulation are thought to be the physiological basis for learning and memory formation [21]. In the visual system the connections between neurons (synapses) are modified according to a simple principle of joint firing, the Hebbian law [7]. According to this law synapses are strengthened if the corresponding cells are activated at the same time. Thus, over time cells "learn" to respond to and in connection with specific other cells. In our model object recognition is achieved by learning the connection weights ($w_{kl}^{\text{V1-HVA}}$) between V1 and HVA. Using a general learning algorithm, that has been shown to capture the features of early visual learning [23], cells from HVA tune themselves to specific features from the set of presented stimuli.

It has been hypothesized that the ventral pathway uses temporal continuity for the development of view-invariant representations of objects ([4, 16, 22]). This temporal continuity can be applied using a trace learning rule. The idea is that on the short time scale of stimuli presentation, the visual input is more likely to originate from different views of the same object, rather than from a different object. To combine stimuli that are presented in succession to one another, activation of a pre-synaptic cell is combined with the post-synaptic activation of the previous stimulus using the Hebbian principle. We simulate an appropriate input presentation protocol and the responses of successive stimuli are combined together to achieve a more invariant representation of an object.

During learning the connection weights $w_{k,l}^{\text{V1-HVA}}$ are changed over time according the Hebbian principle:

$$\tau_L \frac{\partial w_{kl}}{\partial t} = [r_l^{\text{HVA}} - \tilde{r}^{\text{HVA}}]^+ \left((r_k^{\text{V1}} - \tilde{r}^{\text{V1}}) - \alpha_w \left[r_l^{\text{HVA}} - \tilde{r}^{\text{HVA}}\right]^+ w_{kl}\right) \quad (2)$$

$\tilde{r}$ is the mean of the activation over the particular features $\left(e.g., \tilde{r} = \frac{1}{N}\sum_{l=1}^{N} r_l\right)$ and $[x]^+ = \max\{x, 0\}$. α_w constrains the weights analogous to the Oja learning rule [12] and τ_L is the time constant for learning. The V1-HVA weights are learned only at a single receptive field (a 14x14 patch of V1) and their values are shared with all other locations in the HVA (weight sharing approach). The learning was performed on small images containing a single stimulus before processing entire scenes (offline-learning).

Lateral connections between cells were learned by Anti-Hebbian learning. The name Anti-Hebbian implies that this strategy is the opposite of the Hebbian learning rule. Similar to the learning of the synaptic connection weights, where the connection between two cells is increased when both fire simultaneously, in the Anti-Hebbian case the inhibition between two cells is strengthened. The more frequent two cells are activated at the same time, the stronger they inhibit each other, increasing the competition among those two cells (l and l'):

$$\tau_C \frac{\partial c_{l,l'}}{\partial t} = r_{l'} \cdot r_l - \alpha_c r_{l'} \cdot c_{l,l'} \tag{3}$$

where τ_C is the learning rate of the Anti-Hebbian weights. Anti-Hebbian learning leads to decorrelated responses and a sparse code of the cell population [3].

3 Results

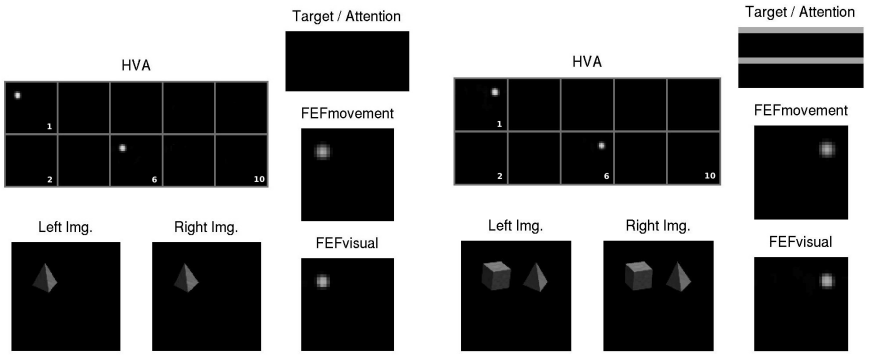
We show the ability of object recognition independent of its position within a scene containing also a distractor object. We measure the performance to recognize all objects with a discriminating value.

3.1 Object recognition independent of its position

An object must be recognized independent of its position in the image, its rotation or its relative size (for an overview see [19]). Position invariance is achieved in the cortex by pooling over a certain spatial area, which is also part of our model.

We now show an object location experiment:

1. We present an object alone in a scene without an attention signal (Fig. 3(a)). The model selects the most conspicuous region (the object) and binds the HVA activation to the working memory (which stores in our example the attention signal).
2. We present a black screen to deplete all cell activities in the system.
3. We test the ability to select the target object. We present a cluttered scene (Fig. 3(b)) (here for simplicity with only 10 features and 2 objects). The attention signal encodes the features of the object and reinforces them in HVA. By this, the system is able to locate the object again (spatial invariant recognition).



(a) (b)

Fig. 3. The figure shows the layer activities during the object location experiment. Here the stereoscopic stimuli, the responses of the feature code (with 10 features) in HVA, the attention signal (features on the y-axis) and both FEF areas are shown. Normally, the x and y axis correspondent to the spatial x and y axis of the images. **a)** The system memorizes the target object, the 'tetrahedron', and stores the HVA response as an attention signal for b). **b)** The attention signal reinforces the features which represent the 'tetrahedron' and the system detects the target object.

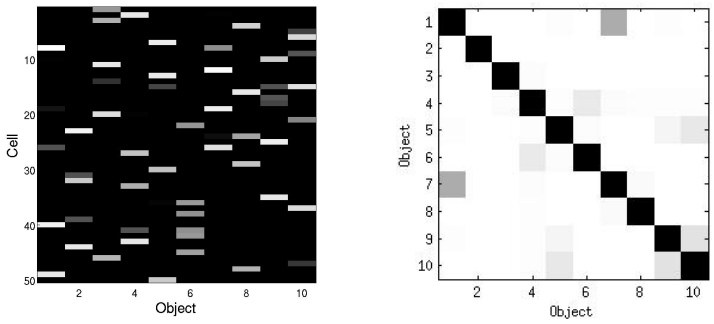
3.2 Object discrimination

To determine the similarity of two feature codes $(\mathbf{r}, \mathbf{s})$ the angle between those two vectors is considered. The lower the value of $d_{TM} \in [0; 1]$ is the more the two vectors show similar cell distributions.

$$d_{TM}(\mathbf{r}, \mathbf{s}) = 1 - \frac{\langle \mathbf{r}, \mathbf{s} \rangle}{|\mathbf{r}|\,|\mathbf{s}|} \quad \text{with: } dim(\mathbf{r}) = \dim(\mathbf{s}) \tag{4}$$

Our results show that regardless of the number of different objects and independently of the number of cells (as long as there is at least one cell per object) the model is capable to learn and discriminate all objects. It can be seen that each object is learned by several cells (Fig. 4(a)) and thus an object is characterized by a specific distributed feature code with nearly no overlap to other objects.

An analysis of whether the model is able to discriminate among the objects is shown in Figure 4(b) using the discrimination value (d_{TM}). Low values (indicated by darker areas) give clue to similar feature codes which would indicate that discrimination between those two objects is impaired. The results show that all objects are very dissimilar in their features and thus are very easy to discriminate. Only object 1 and object 7 show slightly overlapping population codes ($d_{TM} = 0.68$) but the objects can easily be discriminated (compare Figure 4(a)). Although some cells tend to code more than one object the results show that all objects can be discriminated perfectly due to the specific distributed code.



(a) Object selectivity (b) Discrimination between objects

Fig. 4. a) For each object (x-axis) the average firing response (0 dark, 1 bright) of each feature/cell (50 features on the y-axis) is plotted. The average firing response is calculated over all input stimuli that contain the same object. **b)** Using the discrimination value d_{TM}, the similarity of the average response (Fig. 4(a)) to an object is shown here (bright = dissimilar).

4 Discussion

To summarize, attention driven object recognition can solve the problem of selection and segmentation. We had successfully combined stereo vision with object recognition which requires to merge the two views of a scene. We have constructed a merged representation of the scene in low (V1) and high levels areas (HVA). Compared with the perspective of computer vision, this can be seen as a hybrid solution of two contrary approaches. One of them is to construct a merged high level scene model from both images [2], the other one is to combine both images at the level of pixels (resulting in the correspondence problem [18]), which leads to a large number of local false matches.

Our learning algorithm captures the basics of human perception, but can extend to cover complex cell dynamics like calcium traces [20]. We have shown that our system models invariances of the visual cortex. We have focused on spatial invariance and therefore we will have to extend the model and its learning algorithm to scale and rotation invariance. Most neurons in higher areas have a small rotation and scale invariance, but encode a single view to the object (called view-tuned cells [9]). In further investigations we can compare the properties of the learned cells in the HVA with the view-tuned cells.

Acknowledgments. This work has been supported by the EC Project FP7-ICT "Eyeshots: Heterogeneous 3-D Perception across Visual Fragments".

References

1. Chumerin, N.: Nikolay Chumerin's myRaytracer (2009), `OnlineResource:http://sites.google.com/site/chumerin/projects/myraytracer`

2. van Dijck, H.: Object recognition with stereo vision and geometric hashing. Ph.D. thesis, University of Twente (1999)
3. Földiàk, P.: Forming sparse representations by local anti-hebbian learning. Biol Cybern 64, 165–170 (1990)
4. Földiàk, P.: Learning invariance from transformation sequences. Neural Computation 3, 194–200 (1991)
5. Fukushima, K.: Self-organizing neural network models for visual pattern recognition. Acta Neurochir Suppl (Wien) 41, 51–67 (1987)
6. Hamker, F.H.: The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. Comp Vis Image Understand 100, 64–106 (2005)
7. Hebb, D.O.: Organization of Behavior. John Wiley and Sons (1949)
8. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res 40(10-12), 1489–506 (2000)
9. Logothetis, N.K., Pauls, J., Poggio, T.: Spatial reference frames for object recognition tuning for rotations in depth. In: AI Memo 1533, Massachusetts Institute of Technology. pp. 12–0. MIT Press (1995)
10. Milanese, R.: Detecting salient regions in an image: from biological evidence to computer implementation. Ph.D. thesis, University of Geneva (1993)
11. Ohzawa, I., DeAngelis, G.C., Freeman, R.D.: Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. Science 249(4972), 1037–41 (Aug 1990)
12. Oja, E.: A simplified neuron model as a principal component analyzer. J Math Biol 15(3), 267–273 (1982)
13. Qian, N.: Computing stereo disparity and motion with known binocular cell properties. Neural Computation 6(3), 390–404 (1994)
14. Read, J.C.A., Cumming, B.G.: Sensors for impossible stimuli may solve the stereo correspondence problem. Nat Neurosci 10(10), 1322–8 (Oct 2007)
15. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nat. Neurosci 2(11), 1019–25 (Nov 1999)
16. Rolls, E.T., Stringer, S.M.: Invariant object recognition in the visual system with error correction and temporal difference learning. Network 12(2), 111–129 (May 2001)
17. Sabatini, S., Gastaldi, G., Solari, F., Diaz, J., Ros, E., Pauwels, K., Van Hulle, M., Pugeault, N., Krüger, N.: Compact and accurate early vision processing in the harmonic space. In: International Conference on Computer Vision Theory and Applications (VISAPP), Barcelona (2007)
18. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vision pp. 7–42 (2002)
19. Serre, T.: Learning a Dictionary of Shape-Components in Visual Cortex: Comparison with Neurons, Humans and Machines. Ph.D. thesis, Massachusetts Institute of Technology (2006)
20. Shouval, H.Z., Castellani, G.C., Blais, B.S., Yeung, L.C., Cooper, L.N.: Converging evidence for a simplified biophysical model of synaptic plasticity. Biol Cybern 87(5-6), 383–91 (Dec 2002)
21. Squire, L.R., Kandel, E.R.: Memory: From Mind to Molecules. Roberts & Co Publ (2008)
22. Wallis, G., Rolls, E.T.: Invariant face and object recognition in the visual system. Prog Neurobiol 51(2), 167–194 (Feb 1997)
23. Wiltschut, J., Hamker, F.H.: Efficient coding correlates with spatial frequency tuning in a model of v1 receptive field organization. Vis Neurosci 26, 21–34 (2009)