Stereo features in a hierarchical feed-forward model

Sven Eberhardt, Manfred Fahle and Christoph Zetzsche

Cognitive Neuroinformatics,
Universität Bremen
Bibliothekstraße 1, 28359 Bremen
eMail: sven2@uni-bremen.de

Abstract. Depth information is an important auxiliary component to biological and artificial visual systems alike. In an object detection context, it is usually used for object segmentation, for 3D localization of visually matched objects or for direct identification of objects without visual information. However, in all these approaches, depth annotations are handled independently of the visual object recognition process. Here, we present a novel, biologically motivated approach in which luminance and depth information are processed directly as compound features within a pattern matching hierarchy.
We apply the model to train a feature dictionary from a dataset of 3D-rendered objects and show that resulting features include both depth and shape information. We show that these features can be used to improve performance on object detection as well as localization tasks. We further show that the depth annotations in the feature dictionary can be used to produce a 3D structure estimate if only 2D shape information is present.
We hypothesize that binding of multiple submodalities into compound features may prove to be an important building block of how visual information is represented in the human brain and knowledge of this structure might helps us make artificial object recognition systems more robust.

Acknowledgments. This work was supported by DFG, SFB/TR8 Spatial Cognition, project A5-[ActionSpace].

1 Introduction

Despite decades of extensive research, visual pattern recognition in natural scenes remains a constant challenge in research of artificial intelligence [13]. Advancements in this task play a key role in development of autonomous agents - consider for example the technological progress from the first unsuccessful trials of self-driving cars in a desert in 2004 [22] to autonomous cars driving 100'000s of kilometers crash-free on public roads less than a decade later [1].

Notwithstanding the power of technological progress, artificial visual frameworks are usually built and fine-tuned to a specific problem at hand. Humans still outperform such machines in that we have a system that works reliably and robustly when performing a wide range of tasks such as object segmentation and recognition, self-localization and scene categorization, utilizing all available visual dimensions including luminance, color and depth.

How exactly the humans visual system solves these problems is still poorly understood [3]. In particular, it is still unclear how and on which level humans integrate the different visual submodalities for the purpose of object recognition.

Regarding luminance processing, a popular and very promising modeling approach for the human visual system is given by hierarchical ("deep") neural networks [8, 15]. They consist of a cascade of layers which match simple features such as edges and corners at the early stages and build increasingly complex features on

top of that. Features to be matched are commonly extracted in an unsupervised manner from a set of natural images. This structure reflects electrophysiological findings on the visual cortex of primates [20]. However, the hierarchical vision models usually ignore two important visual properties of natural scenes, which both contribute to the strong performance achieved by biological systems: Spectral reflectance of different materials and depth structure.

Depth processing through disparity calculations between retinal images of the eyes is a relatively young development in vertebrates [12]. It gives predators a crucial advantages by breaking camouflage of potential prey that mimics the reflectance of their surroundings, but is not able to conceal their depth structure [7]. Neurophysiologically, depth processing has been established to happen in visual cortex areas V3/KO [2,14]. Recently, functional magnetic resonance imaging studies have shown evidence that these areas do not just process disparity cues into a depth map, but integrate directly with texture properties as well [9].

Although depth information is commonly used in robot vision systems, the traditional approach consists of processing disparity separately from visual information to generate a depth image. This depth image is either used solely for object recognition, or within a 2.5D image map to perform object segmentation [17].

Based on the neurobiological findings from [9] and on considerations about computational efficiency, we here introduce a novel method of calculating and binding disparity directly to features. To this end, we build on a hierarchical neural network model called HMax, which has been developed to model human performance on rapid animal detection in natural scenes [18]. We integrate this with a disparity calculation not on the input image, but on recognized features on intermediate levels of the HMax hierarchy and bind the information directly into the stored pattern dictionary.

The main contributions of this paper are: We present a neurobiologically motivated image processing model that binds shape and disparity information into compound features early in the processing hierarchy. We show that such features can be used for visually driven object classification and localization tasks, and to reproduce depth information when only monocular shape cues are present.

2 Methods

HMax stereo model

We built the presented model as an extension to the HMax implementation by [18]. The standard HMax layers are only described briefly here. For an in-depth description of simple and complex cell layers of this model type, the reader shall consult [18] and [11].

The core principle of HMax is to alternate between layers of simple and complex cells, where the simple cell layers implement pattern matching and the complex cell layers implement pooling over matched patterns at several locations and scales. This alteration of different layer types achieves a trade-off between specifity to a stored pattern and invariance to variations in location and scale of the pattern. Cells in layers early in the processing hierarchy have small receptive field sizes and correspond to simple patterns. In higher stages, both receptive field size and pattern complexity increase.

In our study, we extend the simple and complex cell layer hierarchy of [18]. We duplicate the model into two parallel processing paths and insert matching layers to calculate local disparity. The resulting architecture is illustrated in figure 1.

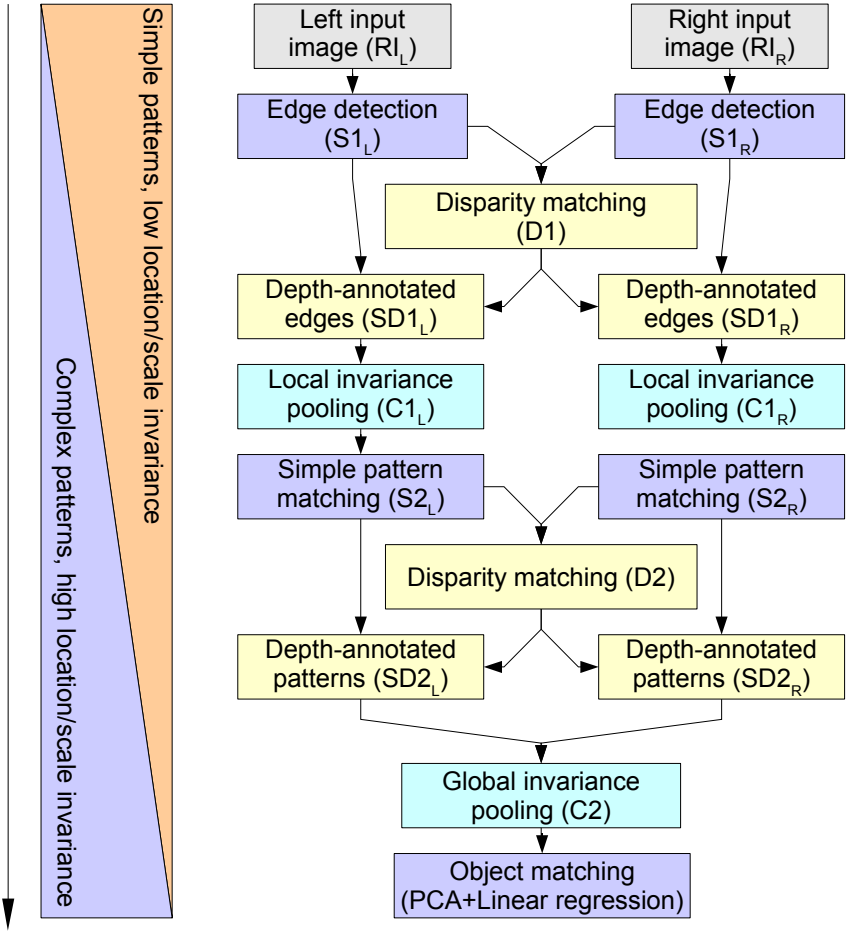

Figure 1: Schematic illustration of HMax implementation: The model alternates between matching (blue) and pooling (green) layers. Matching layers match patterns at several scales and locations, while pooling layers provide invariance over a region of the matched scales and locations. We add disparity matching layers (yellow) to the matching stage, i.e. layers that take the input data from one eye, find corresponding patterns in the other eye and store the horizontal spatial difference between the matches as a disparity map. Information from this map is then used to annotate the features with relative depth components. An additional disparity matching stage is introduced later in the hierarchy because matching of more complex patterns is less prone to errors introduced due to ambiguous matches.

Input images (RI). A pair of corresponding grayscale input images for each eye is loaded into layers RI_L and RI_R for left and right eye respectively. For each layer, ten spatial scales between 256x256 and 52x52 are created by downscaling with linear interpolation (layers SI_L and SI_R).

Simple cell layers (S1). On every resolution, Gabor filters of four different orientations are used to detect edges, leading to four feature maps for each scale on layers $S1_L$ and $S1_R$. Filters are applied as normalized dot product, i.e. the image is convolved with the filter and then divided by the average input pixel luminance.

Disparity layers (D1). Disparity matching is done separately for each feature at every location where the feature activation exceeds a threshold value α. A matching score v for a disparity d of corresponding feature activations f_L and f_R in layers $S1_L$ $S1_R$ at location x, y is calculated as the sum within a window of size w:

$$v(d, f, x, y) := \sum_{x'=-w}^{+w} \left(\frac{f_A(x + x', y)}{1 + |f_A(x + x', y) - f_B(x + x' + d, y)|} \right)$$

The maximum matching score within a matching range m is determined:

$$d_{\text{best}}(f, x, y) := \underset{d \in [-m,m]}{\arg\max}\ v(d, f, x, y)$$

This means that instead of one depth map per image, matching is done on a per-feature basis and each feature carries its own disparity information. The disparity value is probably incorrect at every location where the feature is not matched ($f_L(x, y) = 0$), but this is irrelevant as unmatched features are not stored in the dictionary of following simple cell layers.

Disparity layers are concatenated with their corresponding edge feature layer of each processing path into common feature vectors in layers $SD1_L$ and $SD1_R$.

Local invariance pooling (C1). To simulate complex cell behavior [6], SD1 responses of neighbored locations and scales are combined with a *max* function. This leads to spatial invariance of responses at the cost of spatial resolution.

Simple cell layers (S2). S2A and S2B layers simulate simple cells again. Cells in this layer match their input against a dictionary of stored patterns sampled from random locations and scales of a separate set of sample images. Matching between input and stored pattern happens as Gaussian radial basis functions. Since the layer contains both image and disparity features from layer SD1, patterns are compared based on visual as well as on depth properties. To account for the simplicity of the task, we choose a very small dictionary size of 32 templates.

Disparity layers (D2). The matching algorithm of D1 is repeated in D2 using input layers $S2_L$ and $S2_R$. Again, disparity is re-concatenated into the feature vectors resulting in mixed modality layers $SD2_L$ and $SD2_R$.

Global invariance (C2). In the final model layer, global maximum is taken over all scales and all locations of a feature. This leads to one value per feature per image, which can be fed into a classifier.

The model is implemented in CNS, a GPU-based framework by Mutch et al. [10]. Since processing in the model hierarchy is strictly feed-forward, efficient parallelization on GPU hardware is possible. One complete iteration of an image pair runs in 217ms on an nVidia GeForce8800M GTX GPU, which makes it potentially usable for real-time applications.

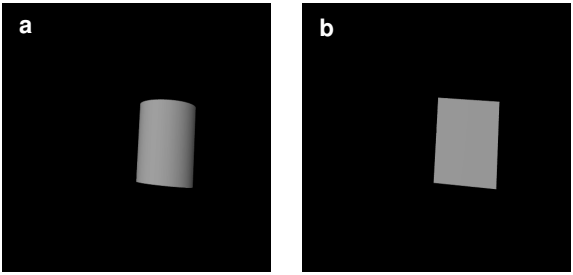

Figure 2: Examples of rendered objects. a) Cylinder class; b) Box class. Monocular dicrimination can happen on small features at the top and bottom caps of the object only. Binocular dicrimination is able to store local depth structure based on the offset of the shading between left and right image, as well as the shading itself, into a feature to identify cylindrical objects.

Datasets and classification

A dataset to evaluate the model has to be carefully chosen to be a) not completely trivial so it could be done on simple shapes and b) sufficiently simple so a classifier working on low level image features, since we only run the low levels of the HMax hierarchy. Therefore, we used POV-Ray[1] to render a dataset of relatively simple objects at varying angles and scales (see figure 2). The total dataset size is 150 images per category. In the object recognition task the two tested classes are cylinders and cubes. In the localization task, the two classes both show cylinders, where one set of cylinders is closer and one is farther from the camera.

We test the model on classification performance between the two shapes doing linear regression with the GURLS software package [19]. To normalize feature count from all input stages, we run principal component analysis and extract the first 128 components of the output vector. Then we split all outputs into a randomized training set of ten images and test performance on the remaining items. Because performance variability between different training set splits is relatively high, each classification is repeated 500 times with randomized test and training sets and the results are used to calculate test statistics.

To determine whether disparity features improve classification performance, disparity is removed from all images in a monocular control run. Significance in difference between monocular and stereo models is calculated with an unpaired t-test between these runs.

3 Results

Classification

Resulting mean performances for different conditions are plotted in figure 3. For the object recognition task (see fig. 3a), the strongest performance at 58.3% correct is achieved by classifying on C2 feature outputs. This is significantly higher than 57%

[1] "Persistence of Vision Raytracer", http://povray.org/

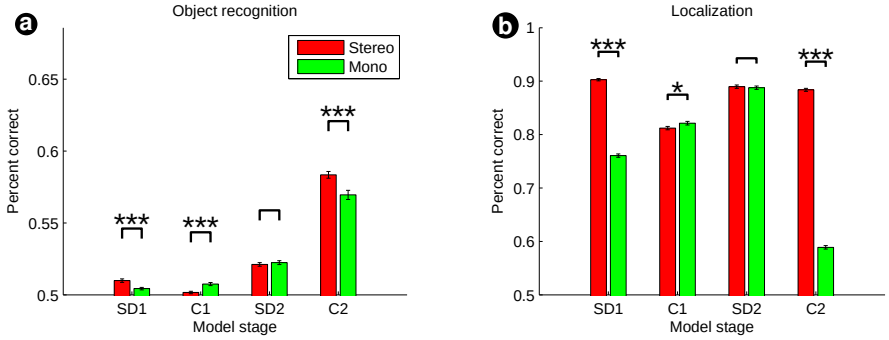

Figure 3: Results performances of classification task on a) object recognition dataset and b) localization dataset for monocular and binocular input. Labels SD1 to C2 correspond to model stages illustrated in figure 1. Asterisks mark significance level of difference between monocular and stereo run.

achieved by monocular input ($p < 0.001$). This shows that disparity bound into the higher level features contributes to the dicriminability between the classes.

Classification on features extracted from lower levels of the architecture yields low performances barely above chance level (see fig. 3a SD1 and C1). We can conclude that object recognition depends on the translation and scale invariance introduced in the final complex cell layer.

On the localization task, absolute performance is higher than on the object recognition task (see fig. 3b). Surprisingly, classification on low level features in SD1, i.e. the first pattern merging stage, leads to higher performance than on high level SD2 features. This hints that localization can be performed well on low-level features and the pooling introduced by high level layers discards information that is crucial for localization tasks.

Again, the performance gain from adding disparity into the features is significant both on low level S1 features ($p < 0.0001$) and on high level C2 features ($p < 0.0001$). This result shows that although no full disparity map is calculated by our approach, binding disparity on stored patterns provides sufficient information to encode the location of the object in depth.

Depth reconstruction

One key advantage of encoding information in multi-modal patterns is that after features have been learned, modalities can be reconstruct from each other. To illustrate this approach, we match SD2 patterns learned on binocular image to a monocular image input. We then sum up the disparity component of all matched features at every location. The resulting depth image reconstructed from a cube is shown in figure 4.

The image shows that valid depth information could be reconstructed near locations that had valid pattern matches of low level features, i.e. edges and corners. No depth information is present at the center of the cube because due to lack of textures, no features were matched there. However, using a larger dictionary which also includes large scale image features could probably remedy this issue.

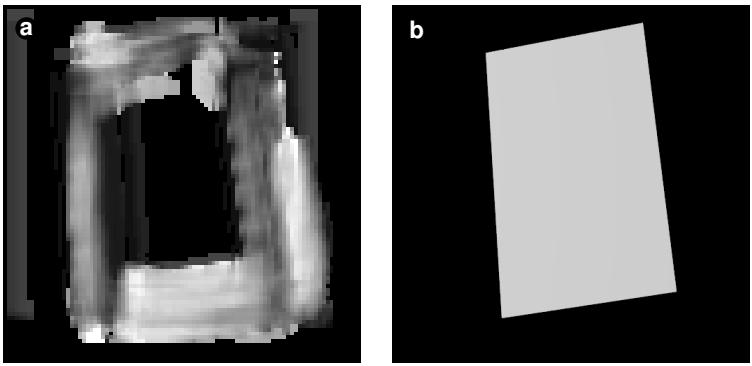

Figure 4: Depth reconstruction of a cube: Depth could be determined near matched features (corners and edges) of the cube only. The bottom right cube is slightly slanted towards the viewer and although no depth cues are present on the image, the reconstruction found the slant. No features are matched in the center of the cube, so no depth information is present there.

4 Discussion

We show that binding multiple modalities into compound features early in the processing hierarchy yields processing units that can be used to improve performance compared to mono-modal features in basic vision tasks. This result is interesting not only with respect to our understanding of how depth processing in the human brain might work, but also shows that we could learn from biology and build recognition systems that employ early binding of modalities before detection of features.

Intuitively, per-feature disparity storage seems like a much more plausible way for humans to store and process depth information than the 2.5D map traditionally often used in robotics. It allows for an important processing optimization, since disparity would only need to be evaluated for features that are actually used for visual tasks. A similar method has been used successfully by Saxena et al. [16] in a robotic grasping task of novel objects.

The approach also provides a suggestion on how to circumvent the binding problem, i.e. the question how separately extracted shape and depth features from multiple objects in a scene are later attributed to matching objects [21]. In our model, we simply bind the modalities early when location information is still contained in the features.

When spatial feature information is pooled in the final complex cell layer C2 of the model, performance on the localization task drops significantly when executed on monocular features (see fig. 3b). Here we found this effect on small-scale localization with respect to one object only. However, studies have shown that even for self-localization on larger scales (e.g. within a city), histograms over simple features provide a stronger location cue than more complex features [4, 5], suggesting the hypothesis that this is a general property inherent to the localization task. We conclude that for self-localization, it is advantageous to extract and match on simple features aggregated from multiple modalities - like shape, disparity and color - rather than building complex features from a single modality.

References

[1] The self-driving car logs more miles on new wheels., http://goo.gl/dNgvF

[2] Ban, H., Preston, T., Meeson, A., Welchman, A.: The integration of motion and disparity cues to depth in dorsal visual cortex. Nature neuroscience (2012)

[3] DiCarlo, J.J., Zoccolan, D., Rust, N.C.: How Does the Brain Solve Visual Object Recognition? Neuron 73(3), 415–434 (2012)

[4] Eberhardt, S., Kluth, T., Reineking, T., Zetzsche, C., Schill, K.: Models for invariant place recognition. In: Proceedings of KogWis. p. 10 (2012)

[5] Eberhardt, S., Zetzsche, C.: Low-level global features for vision-based localization. In: Proceedings of the KI 2013 Workshop on Visual and Spatial Cognition. pp. 5–13 (2013)

[6] Hubel, D., Wiesel, T.: Receptive fields and functional architecture of monkey striate cortex. The Journal of physiology (1968)

[7] Julesz, B.: Foundations of cyclopean perception. (1971)

[8] Krüger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., Rodríguez-Sánchez, A.J., Wiskott, L.: Deep hierarchies in the primate visual cortex: what can we learn for computer vision? IEEE transactions on pattern analysis and machine intelligence 35(8), 1847–71 (2013)

[9] Murphy, A.P., Ban, H., Welchman, A.E.: Integration of texture and disparity cues to surface slant in dorsal visual cortex. Journal of neurophysiology 110(1), 190–203 (2013)

[10] Mutch, J., Knoblich, U., Poggio, T.: CNS: a GPU-based framework for simulating cortically-organized networks. Tech. rep., Massachusetts Institute of Technology, Cambridge, MA (2010)

[11] Mutch, J., Lowe, D.G.: Object Class Recognition and Localization Using Sparse Features with Limited Receptive Fields. International Journal of Computer Vision 80(1), 45–57 (2008)

[12] Pettigrew, J.D.: Evolution of binocular vision. Visual neuroscience pp. 208–222 (1986)

[13] Pinto, N., Cox, D.D., DiCarlo, J.J.: Why is real-world visual object recognition hard? PLoS computational biology 4(1), e27 (2008)

[14] Poggio, G., Gonzalez, F., Krause, F.: Stereoscopic mechanisms in monkey visual cortex: binocular correlation and disparity selectivity. The Journal of neuroscience (1988)

[15] Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nature neuroscience (1999)

[16] Saxena, A., Driemeyer, J., Ng, A.: Robotic grasping of novel objects using vision. International Journal of Robotics (2008)

[17] Scharstein, D., Szeliski, R.: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. International Journal of Computer Vision 47(1-3), 7–42 (2002)

[18] Serre, T., Oliva, A., Poggio, T.: A feedforward architecture accounts for rapid categorization. Proceedings of the national academy of sciences 104(15), 6424–6429 (2007)

[19] Tacchetti, A., Mallapragada, P.K., Santoro, M., Rosasco, L.: GURLS: a toolbox for large scale multiclass learning. In: Big learning workshop at NIPS (2011), http://cbcl.mit.edu/gurls/

[20] Tanaka, K., Saito, H., Fukada, Y., Moriya, M.: Coding visual images of objects in the inferotemporal cortex of the macaque monkey. J Neurophysiol 66(1), 170–189 (1991)

[21] Treisman, A.: The binding problem. Current Opinion in Neurobiology 6(2), 171–178 (1996)

[22] Urmson, C., Anhalt, J., Clark, M.: High speed navigation of unrehearsed terrain: Red team technology for grand challenge 2004. Tech. Rep. CMU-RI (2004)