

ORASSYLL: Object Recognition with Autonomously Learned and Sparse Symbolic Representations Based on Local Line Detectors ^{*}

Norbert Krüger, Niklas Lüdtko
Ruhr-Universität Bochum,
Institut für Neuroinformatik,
Germany

Abstract

We introduce an object recognition system in which objects are represented as a sparse and spatially organized set of local (bent) line segments. The line segments correspond to binarized Gabor wavelets or banana wavelets, which are bent and stretched Gabor wavelets. These features can be metrically organized, the metric enables an efficient learning of object representations. Learning can be performed autonomously by utilizing motor-controlled feedback. The learned representations are used for fast and efficient localization and discrimination of objects in complex scenes.

1 Introduction

In this paper we describe a novel object recognition system called ORASSYLL (Object Recognition with Autonomously learned and Sparse SYmbolic representations based on Local Line detectors). In ORASSYLL representations of object classes can be learned autonomously. The learned representations are used for a fast and efficient location and identification of objects in complicated scenes.

Learning is inherently faced with the bias/variance dilemma [3]: If the starting configuration of the system is very general it will have to pay for this advantage by having many internal degrees of freedom resulting in bad generalization abilities —the “variance” problem. On the other hand, if the initial system has few degrees of freedom it may be able to learn efficiently, but there is great danger that the structural domain spanned by those degrees of freedom does not cover the given domain of application —the “bias” problem. We show here that appropriately structured *a priori* knowledge can help to cope with the bias–variance dilemma. Formulating a number of *a priori* principles to reduce the dimension of the search space and to guide learning we handle the variance–problem. We expect to avoid the bias–problem because of the general applicability of those principles. Important constraints are:

PF1 Significant features of a local area of the two–dimensional projection of the visual world are localized (bent) lines.

^{*}Supported by grants from the German Ministry for Science and Technology 01IN504E9 (NEUROS), 01M3021A4 (Electronic Eye), and the EC Human Capital and Mobility Program (ERBCHRX CT930097).

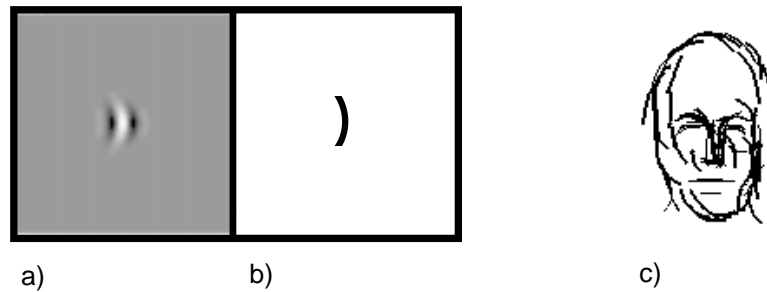


Figure 1: Path corresponding to a banana wavelet. a: Arbitrary wavelet. b: Corresponding path. c: Visualization of a representation of an object class. Gabor or Banana wavelets with lower frequencies are represented by line segments with larger width.

PF2 Metric organization of the feature space indicating their differences in the properties orientation, curvature and position.

PF3 Hierarchical processing of features.

PF4 Sparse coding.

Other constraints, discussed in detail in [5], are concerned with the division of the feature space in independent subspaces (PL1: Independence), its temporal organization (PL2: Correspondence) and statistical criteria for the evaluation of significant features for an object class (Invariance Maximization (PE1) and Redundancy Reduction (PE2)).

Our representation of a certain view of an object class comprises only important features, learned from different examples (see figure 8c and 2). In section 2 we formalize PF1 by assigning a local line segment to Gabor wavelets or banana wavelets respectively (see figure 1a,b). In addition to the parameters frequency and orientation banana wavelets possess the properties curvature and elongation (see figure 3). The space of banana wavelet responses is very large. An object can be represented as a configuration of a few of these features, therefore it can be coded sparsely (PF4). The feature space can be understood as a metric space (PF2), its metric representing the similarity of features. This metric is essential for feature extraction and the learning algorithm (section 3). The banana wavelet responses can be derived from Gabor wavelet responses by hierarchical processing (PF3) to gain speed and reduce memory requirements. The sparse representation combined with the hierarchical feature processing allows a fast and effective locating (section 4).

In order to avoid the necessity of manual intervention for the generation of ground truth we equip the system with a mechanism which can produce controlled training data by moving an object with a robot arm and following the object by fixating the robot hand. The robot produces training data on which a certain view of an object is shown with varying background and illumination but with corresponding landmarks having the same pixel position in the image (see figure 2). We apply the learning algorithm to this data to extract an object representation (see figure 2v). Another way to avoid manual intervention is one-shot learning (see figure 6), which already allows for the extraction of representations successfully applicable to difficult discrimination tasks.

Recently additional *a priori* knowledge has been introduced in terms of Gestalt principles "collinearity" and "parallelism" [10]. This is motivated by the discovery of sec-

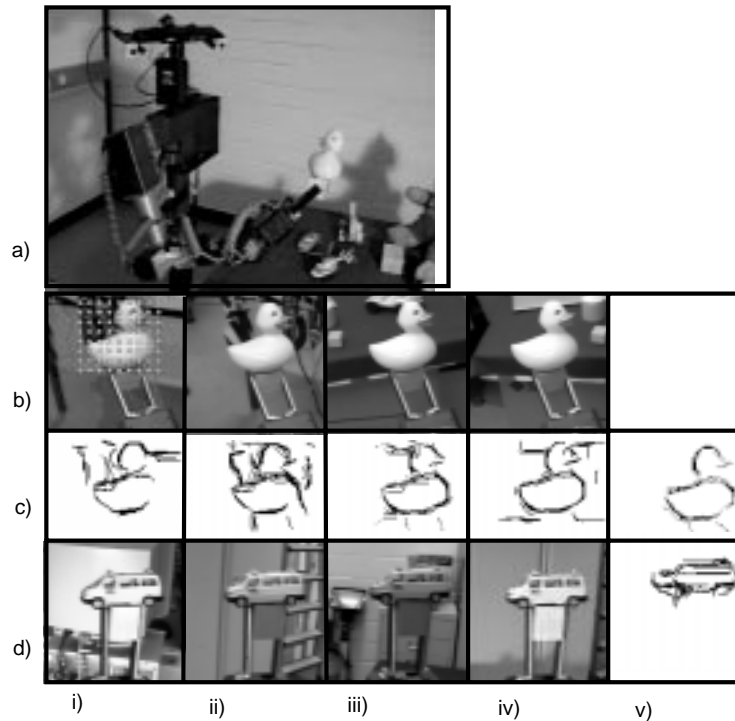


Figure 2: **a)** The robot arm with the camera. **b)** The “retinal” images produced by following the robot arm holding a toy-duck. **c,i-iv)** Significant Features per Instance extracted in an rectangular region (shown in b,i). **c,v)** Learned representation. **d)** Training data and learned representation for a toy car.

ond order correlations between collinear and parallel line segments in natural images [4]. Supposing that at least some parts of perceptual grouping in biological vision can be considered as a consequence of preattentive low level mechanisms [12], Gestalt relations are integrated by establishing “hard-wired” connections among local features.

Our system has certain analogies to the visual system of vertebrates. There is evidence for curvature sensitive features processed in a hierarchical manner [1]; sparse coding has been discussed as a coding scheme used in the visual system [2]; and metric organization of features seems to play an important role for information processing in the brain [13]. Instead of detailed modelling of brain areas we aim to apply some basic concepts inspired by brain research (such as sparse coding, hierarchical processing, metrical organisation of features, etc.) in our artificial object recognition system. For a more detailed discussion of the analogies to biology we refer to [7].

2 The Feature Space

The principle PF1 gives us a significant reduction of the search space. Instead of allowing, e.g., all linear filters as possible features, we restrict ourself to a small subset. Considering the risk of a wrong feature selection it is necessary to give good reasons for

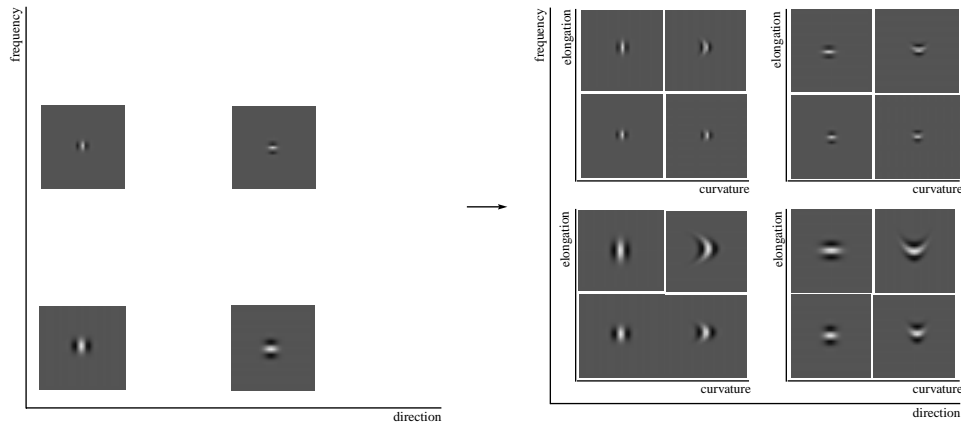


Figure 3: Relation between Gabor wavelets and banana wavelets.

$$\begin{matrix} \text{banana wavelet} \end{matrix} = \beta_1^{\vec{b}} \cdot \begin{matrix} \text{Gabor wavelet 1} \end{matrix} + \beta_2^{\vec{b}} \cdot \begin{matrix} \text{Gabor wavelet 2} \end{matrix} + \beta_3^{\vec{b}} \cdot \begin{matrix} \text{Gabor wavelet 3} \end{matrix}$$

Figure 4: The banana wavelet on the left is approximated by the weighted sum of Gabor wavelets on the right.

our decision. We argue that nearly any 2D-view of an object can be composed of localized curved lines. Furthermore, the fact that humans can easily handle line drawings of objects strengthens our assumption PF1.

Banana Wavelets: A banana wavelet $B^{\vec{b}}$ is a complex-valued function, parameterized by a vector \vec{b} of four variables $\vec{b} = (f, \alpha, c, s)$ expressing the attributes frequency (f), orientation (α), curvature (c) and size (s). It can be understood as a product of a curved and rotated complex wave function $F^{\vec{b}}$ and a stretched two-dimensional Gaussian $G^{\vec{b}}$ bent and rotated according to $F^{\vec{b}}$:

$$B^{\vec{b}}(x, y) = G^{\vec{b}}(x, y) \cdot \left(F^{\vec{b}}(x, y) - e^{\frac{\sigma_g}{2}} \right) \tag{1}$$

with

$$G^{\vec{b}}(x, y) = \exp \left(-\frac{f^2}{2} \left(\sigma_x^{-2} \left(x \cos \alpha + y \sin \alpha + c (-x \sin \alpha + y \cos \alpha)^2 \right)^2 + \sigma_y^{-2} s^{-2} (-x \sin \alpha + y \cos \alpha)^2 \right) \right)$$

and

$$F^{\vec{b}}(x, y) = \exp \left(i f \left(x \cos \alpha + y \sin \alpha + c (-x \sin \alpha + y \cos \alpha)^2 \right) \right).$$

Our basic feature is the magnitude of the filter response of a banana wavelet extracted by a convolution with an image¹. A banana wavelet $B^{\vec{b}}$ causes a strong response at pixel

¹The subtraction by the DC-part $e^{\frac{\sigma_g}{2}}$ in equation (1) insures the independence of the filter response from the mean grey-value.

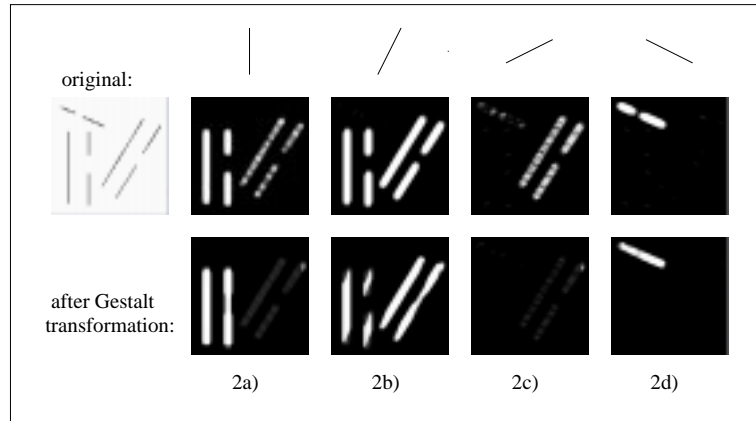


Figure 5: **upper row:** A line drawing and its Garbor transformation displayed for four different filter orientations (out of eight 0..7). Direction 0 is the vertical direction and 4 is the horizontal one. **underneath:** The Garbor transformation modified by Gestalt principles. The gaps in 2a), b) and d) have been bridged and the orientation Selectivity is improved.

position \vec{x} when the local structure of the image at that pixel position is similar to $B^{\vec{b}}$ (see [5, 6]).

The Feature Space: The six-dimensional space of vectors $\vec{c} = (\vec{x}, \vec{b})$ is called the *feature space* with \vec{c} representing the banana wavelet $B^{\vec{b}}$ with its center at pixel position \vec{x} in an image. In [5, 6] we define a metric $d(\vec{c}_1, \vec{c}_2)$. Two coordinates \vec{c}_1, \vec{c}_2 are expected to have a small distance d when their corresponding kernels are similar, i.e., they represent similar features (PF2).

Non-linear Gestalt transformation: A non-linear Gestalt transformation (see figure 5) can be performed on a Garbor-transformed image by setting up contextual relations among features. In case of collinearity and parallelism *straight* filters of equal orientation and frequency are combined. For each point in the Garbor-transform a confidence value of collinear and parallel context is computed from responses of surrounding filters within a well defined area (for details see [10]). An important aspect of ORASSYLL is a criteria of the presence of a local line segment. By applying the Gestalt principles collinearity and parallelism we can make use of *global relations* for this criterion. Globalized edge detection is expected to be more robust with respect to local distortions due to a complex background or poor conditions of illumination.

Approximation of Banana Wavelets by Gabor Wavelets: To reduce computational requirements for the extraction of the large feature space we have defined an algorithm to approximate banana wavelets from Gabor wavelets and banana wavelet responses from Gabor wavelet responses (see [5, 6]). By this hierarchical processing (PF3) we achieve a speed up to a factor 5. Figure 4 gives the idea of the approximation algorithm.

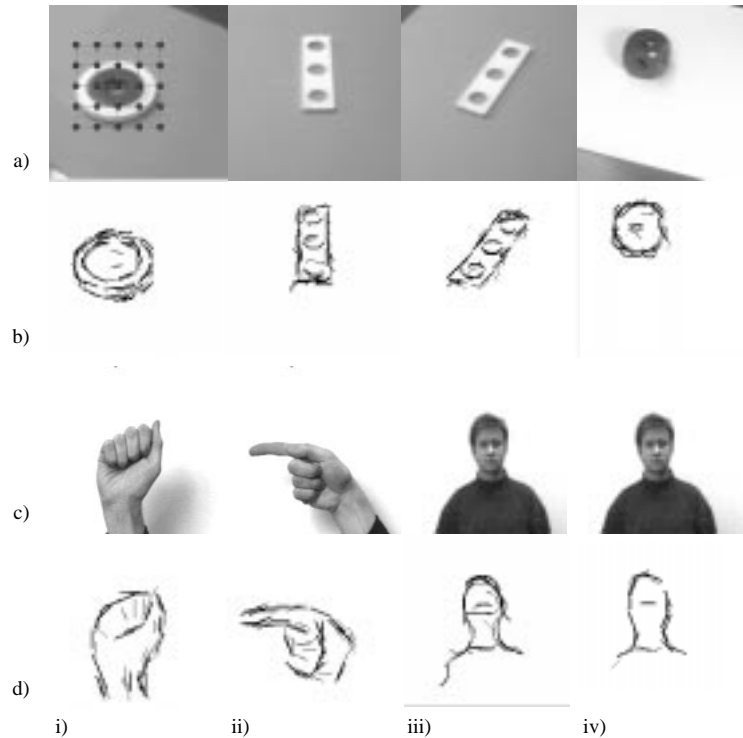


Figure 6: One-shot learning: Row a) and c) show the objects to be learned in front of homogeneous background. Row b) and d) show the extracted representations. For all objects a rectangular grid was roughly positioned on the object as in the first image a,i).

3 Learning

Extracting Significant Features Per Instance: Our aim is to extract the local structure in an image I in terms of curved lines expressed by banana wavelets. We define a *significant feature per instance* of an object by two qualities. Firstly it has to cause a strong response (**C1**), secondly it has to represent a maximum within a local area of the feature space (**C2**). Figure 8b,i-iv), 6b,c) and 2c,i-iv) show the significant features per instance for some objects (each banana wavelet is described by a curve with same orientation, curvature and size). In terms of analogy to the processing in area V1 in the mammalian visual system C1 may be interpreted as the response of a certain column which indicates the general presence of a feature, whereas C2 represents the intercolumnar competition giving a more specific coding of this feature [11].

One-shot learning: By positioning a rectangular grid on a roughly segmented object (see figure 6a,i) in front of homogeneous background and extracting significant features per instance as described above suitable representations of objects can already be extracted. These representations are successfully applied to difficult discrimination tasks.

Clustering: After extracting the significant features per instance in different pictures we apply an algorithm to extract invariant local features for a *class of objects*. Here the task is the selection of the *relevant features* for the object class from the noisy features

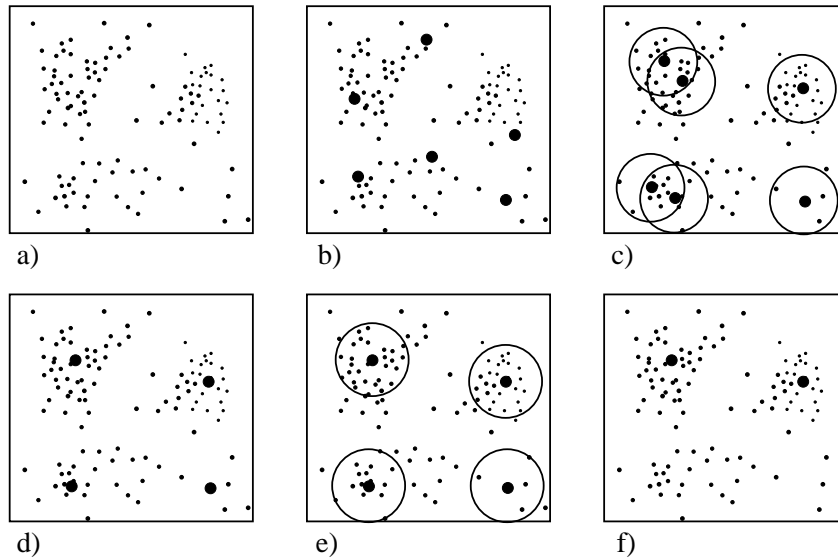


Figure 7: Clustering: a) Distribution of the significant features per instance extracted at a certain landmark. b) Codebook initialization. c) Codebook vectors after learning. d) Substituting sets of codebook vectors with small distance by their center of gravity. e) Counting the number of elements within a certain radius. f) Deleting codebook vectors representing insignificant features.

extracted from our training examples (see figure 8b,i-iv) and 2c,i-iv). We assume the correspondence problem to be solved, i.e., we assume the position of certain landmarks of an object to be known on pictures of different examples of these objects. In some of our simulations we determined corresponding landmarks manually, for the rest we replaced this manual intervention by motor controlled feedback (see section 5).

In a nutshell the learning algorithm works as follows (illustrated for two dimensions in figure 7): Fig.7a-c) For each landmark we express the significant features per instance of all training examples by six dimensional codebook vector (\vec{x}, \vec{b}) , representing the pixel position and the parameter frequency, orientation, curvature and elongation. We optimize the codebook vectors by the LBG vector quantization algorithm [9]. Fig.7d) Codebook vectors with small distances are substituted by their center of gravity (PE2: reduction of redundancy). Fig.7e,f) A significant feature for an object class is defined as a codebook vector expressing many data points. That means the feature corresponding to the code book vector or a similar feature (according to our metric d) often occurs in our training set, i.e., has high invariance (PE1). We end up with a graph with its nodes labeled with banana wavelets representing the learned significant features (see figure 8bv, and figure 2dv, ev). The edges of the graph labeled with metric relations of the landmarks.

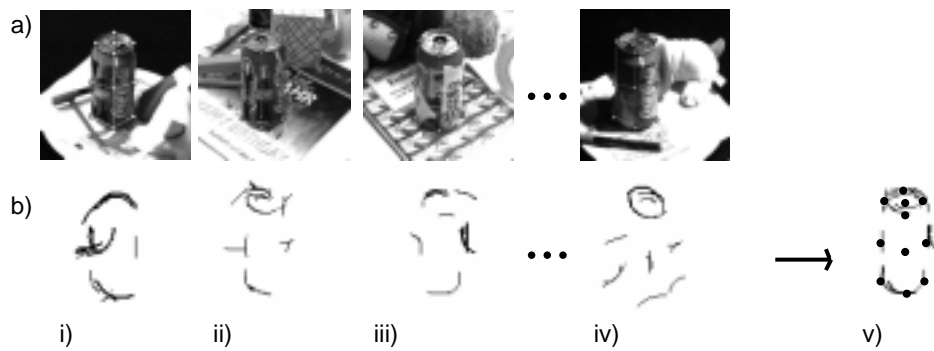


Figure 8: **a)** Pictures for training. **b,i-iv)**: Significant features per instance describing beside relevant information also accidental features such as background, shadow or surface textures. **b,v)** The learned Representation.

4 Matching

To use our learned representation for location and classification of objects we define a similarity function between a graph labeled with the learned banana wavelets and a certain position in the image. A *total similarity* averages *local similarities*. The local similarity expresses the system's confidence whether a pixel in the image represents a certain landmark. The graph is adapted in position and scale by optimizing the total similarity. The graph with the highest similarity determines the size and position of the objects within the image.

In a nutshell the local similarity is defined as follows (for details see [5]): For each learned feature and pixel position in the image we simply check whether the corresponding banana response is high or low, i.e., the corresponding feature is present or absent. Because of the sparseness (PF4) of our representation only a few of these checks have to be made, therefore the matching is fast. Because we make use only of the important features, the matching is efficient.

5 Simulations

Learning of Representation: Firstly we apply the learning algorithm to data consisting of manually provided landmarks. Our training sets consist of a set of approximately 60 examples of an object viewed in a certain pose. As objects we use cans, faces, and hand postures. Corresponding landmarks are defined manually on the different representatives of a class of objects (figure 8).

To avoid the manual generation of ground truth we can either apply one-shot learning (see section 3) or make use of motor controlled feedback: By moving an object with a robot arm and following the object by keeping fixation relative to the robot hand using its known 3D position, we produce training data in which a certain view of an object is shown with varying background and illumination but with corresponding landmarks in the same pixel position within the image (see fig 2b,d). Then we can apply our learning algorithm with a rectangular grid roughly positioned on the object (see figure 2b,i). For



Figure 9: Face finding with learned representations for three scales. The mismatch (right) is caused by the person's unusual arm position.

the generation of ground truth for frontal faces we recorded a sequence of pictures in which a person is sitting fixed on a chair. Illumination and background is changed as for cans. To extract representations for different scales we apply the learning algorithm to the very same pictures of the different sequences scaled accordingly (see figure 9).

Matching: For the problem of face finding in complex scenes with large size variation a significant improvement in terms of performance and speed compared to the older system [8, 14] could be achieved. Figure 9 shows some examples of matches and mismatches. The object finding in one picture approximately requires 1.5 seconds on a Sparc Ultra. We also performed successfully matching with cans, hand postures, and other objects, as well as various discrimination tasks (most of them described in [5]). A detailed comparison of ORASSYLL and the object recognition system [8, 14] can be found in [5].

6 Conclusion

We described an object recognition system which is able to learn autonomously efficient representations of objects. Learning is guided by a careful selection of powerful and general *a priori* constraints. The learned representations have been successfully applied to difficult matching tasks.

References

- [1] A. Dobbins and S. Zucker. Endstopped neurons in the visual cortex as a substrate for calculating curvature. *Nature*, 329:438–441, 1987.
- [2] D. Field. What is the goal of sensory coding? *Neural Computation*, 6(4):561–601, 1994.
- [3] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1995.
- [4] N. Krüger. Collinearity and parallelism are statistically significant second order relations of complex cell responses. *accepted for Neural Processing Letters*.
- [5] N. Krüger. *Visual Learning with a priori Constraints (phd thesis)*. 1998.
- [6] N. Krüger, G. Peters, and C. v.d. Malsburg. Object recognition with a sparse and autonomously learned representation based on banana wavelets. Technical report, Institut für Neuroinformatik, Bochum, 1996.
- [7] N. Krüger, M. Pöttsch, and G. Peters. Principles of cortical processing applied to and motivated by artificial object recognition. In R. Baddeley, P. Hancock, and P. Foldiak, editors, *accepted for "Information Theory and the Brain"*. Cambridge University Press, 1998.

- [8] M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R.P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamik link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1992.
- [9] Y. Linde, A. Buzo, and R.M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on communication*, vol. COM-28:84–95, 1980.
- [10] N. Lüdtkke. Integrating of gestalt principles by non-linear feature linking (studiennarbeit). Technical report, Institut für Neuroinformatik, Bochum, 1998.
- [11] M.W. Oram and D.I. Perrett. Modeling visual recognition from neurobiological constraints. *Neural Networks*, 7:945–972, 1994.
- [12] U. Polat, K. Mizobe, M.W. Pettet, T. Kasamatsu, and A.M. Norcia. Collinear stimuli regulate visual responses depending on cell's contrast threshold. *Nature*, 391:580–584, 1998.
- [13] K. Tanaka. Neuronal mechanisms of object recognition. *Science*, 262:685–688, 1993.
- [14] L. Wiskott, J.M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 775–780, 1997.