Internal Report 97–17

Principles of Cortical Processing Applied to and Motivated by Artificial Object Recognition

by

Norbert Krüger, Michael Pötzsch, Gabriele Peters

Ruhr-Universität Bochum
Institut für Neuroinformatik
44780 Bochum

IR-INI 97–17
Oktober 1997
ISSN 0943-2752

Principles of Cortical Processing Applied to and Motivated by Artificial Object Recognition*

Norbert Krüger, Michael Pötzsch, Gabriele Peters

Abstract

In this paper we discuss the biological plausibility of the object recognition system described in detail in (Krüger, Peters and v.d. Malsburg 1996). We claim that this system realizes the following principles of cortical processing: hierarchical processing, sparse coding, and ordered arrangement of features. Furthermore, our feature selection is motivated by response properties of neurons in striate cortex and by Biederman's theory of object representation on higher stages of visual processing (Biederman 1987). Inspired by the current discussion about aspects of cortical processing, we hope to derive more efficient algorithms. By discussing the functional meaning of these aspects in our object recognition system, we hope to attain a deeper understanding of their meaning for brain processing.

1 Introduction

It is an assumption of the neural computation community that the brain as the most successful pattern recognition system is a useful model for deriving efficient algorithms on computers. But how can a useful interaction between brain research and artificial object recognition be realized? We see two questionable ways of interaction. On the one hand, a very detailed modelling of biological networks may lead to a disregard of the *task* solved in the brain area being modelled. On the other hand, the neural network community may loose credibility by a very rough simplification of functional entities of brain processing. This may result in a questionable naming of simple functional entities as neurons or layers to pretend biological plausibility. In our view, it is important to understand the brain as a tool *solving a certain task* and therefore it is important to understand the *functional meaning* of principles of cortical processing such as, hierarchical processing, sparse coding, and ordered arrangement of features. Some researchers (e.g., (Atick 1992; Barlow 1961; Földiák 1990; Olshausen and Field in press; Palm 1980) have already made important steps in this direction. They have given an interpretation of some of the above–mentioned principles in terms of information theory. Others (e.g., (Hummel and Biederman 1992; Lades et al. 1992)) have tried to initiate an interaction between brain research and artificial object recognition by building efficient and biologically motivated object recognition systems. Following these two lines of research, we suggest to look at a functional level of biological processing and to utilize abstract principles of cortical processing in an artificial object recognition system.

In this paper we discuss the biological plausibility of the object recognition system described in detail in (Krüger, Peters and v.d. Malsburg 1996). We claim that this system realizes the above–mentioned principles. Although the system's performance is not comparable to the power of the human visual system, it is already able to deal with difficult vision problems. The object recognition system is based on *banana wavelets*, which are generalized Gabor wavelets. In addition to the parameters frequency and orientation, banana wavelets have the attributes of curvature and elongation (figure 1). The space of banana wavelet responses is much larger compared to the space of Gabor wavelet responses, and an object can be represented as a configuration of a few of these features (figure 2v); therefore it can be coded sparsely. The space of banana wavelet responses can be understood as a metric space, its metric representing the similarity of features. This metric is utilized for the learning of a representation of 2D–views of objects. The banana wavelet responses can be derived from Gabor wavelet responses by hierarchical processing to gain speed and reduce memory requirements. A set of examples of a certain view of an object class (figure 2i–iv) is used to learn a sparse representation, which

*Supported by grants from the German Ministry for Science and Technology 01IN504E9 (NEUROS) and 01M3021A4 (Electronic Eye).

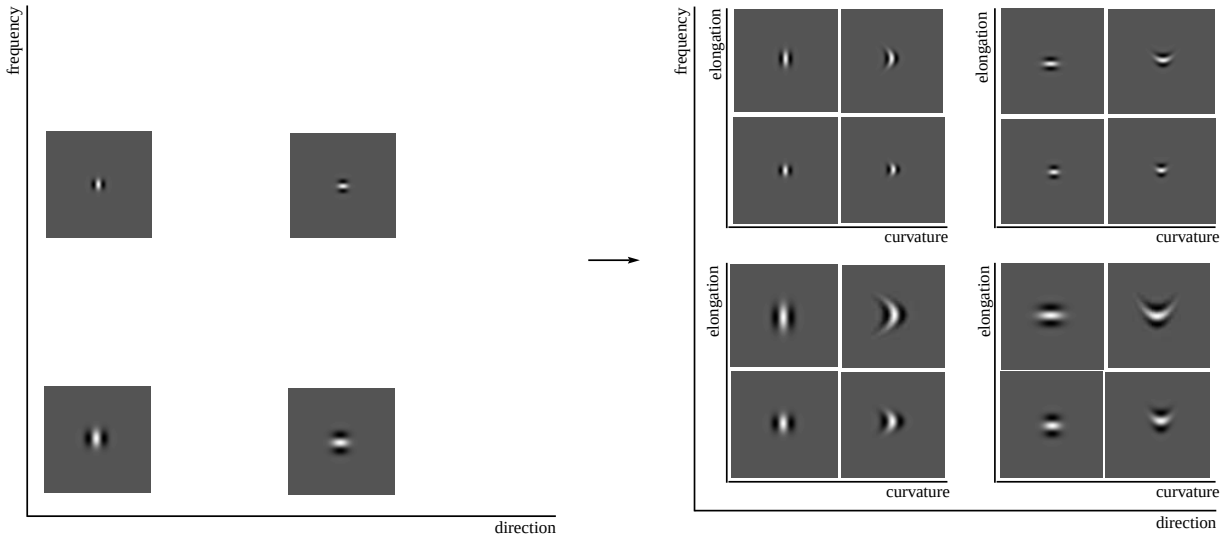

Figure 1: Relationship between Gabor wavelets and banana wavelets. Left: four examples of Gabor wavelets which differ in frequency and direction only. Right: 16 examples of banana wavelets which are related to the Gabor wavelets on the left. Banana wavelets are described by two additional parameters (curvature and elongation).

contains only the important features. This sparse representation allows for a quick and efficient localization of objects.

By discussing the functional meaning of sparse coding, hierarchical processing, and order in the arrangements of features, as well as the implication of our feature selection for our artificial object recognition system, we hope to attain a deeper understanding of their meaning for brain processing. Following the discussion of principles of cortical processing, we hope to be inspired to derive more efficient algorithms. In section 2 we give a short description of our system. In section 3 we discuss the above–mentioned principles of visual processing in their biological context, as well as their algorithmic realization in our system. We compare both aspects, and we claim that the utilization of the above–mentioned principles supports the strength of our system. We close with a conclusion and an outlook in section 4.

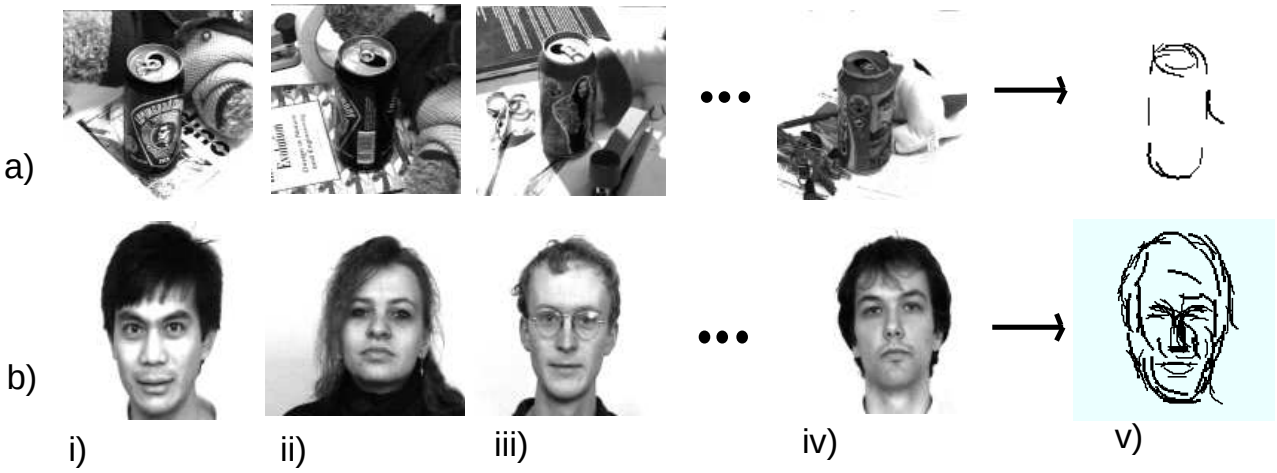

Figure 2: i–iv) Different examples of cans and faces used for learning. v) The learned representations.

2 Object Recognition with Banana Wavelets

In this section we give a description of the basic entities of our system (for details see (Krüger, Peters and v.d. Malsburg 1996). We restrict ourselves to those aspects relevant to the discussion in section 3. In our approach we limit ourselves to form processing and we ignore color, movement, texture, and binocular information. In the literature (see e.g., (Treisman 1986)) a largely independent processing of these different clues is assumed with the shape clue as the most powerful one for higher level classification tasks. The object recognition system is influenced by an older system developed in the von der Malsburg group (Lades et al. 1992; Wiskott et al. 1997) and by Biedermans criticism (Biederman and Kalocsai in press) of this system.
The system introduced here differs from the older system in two main aspects: Firstly we introduce curvature as a new feature. Secondly, and even more important, we introduce a *sparse* object representation: we describe an object by an ordered arrangemant of a few *binary features* which can be interpreted as local line segments. From this reduced representation the original image is not reconstructable but the object is represented in its essential entities. In the older system, which is based on Gabor wavelets, an object is described by a much larger amount of data representing the object as sets of local Gabor wavelet responses, called "jets", from which the original image is almost completely recoverable.
In (Biederman and Kalocsai in press) it is shown that there is a high correlation of the older system's performance and human performance for face recognition but only low correlation for object recognition tasks. As one of the main weaknesses he points to Gestalt principles not utilized by the older system but by humans. We think the object recognition system described here represents an important step towards an integration of higher perceptual grouping mechanisms.

2.1 The Banana Space

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 elongation (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 $F^{\vec{b}}$ (figure 3):

$$\begin{aligned} B^{\vec{b}}(x,y) &= G^{\vec{b}}(x,y) \cdot \left(F^{\vec{b}}(x,y) - e^{-\frac{\sigma_x}{2}} \right) \\ G^{\vec{b}}(x,y) &= \exp\left(-\frac{f^2}{2} \left(\frac{\left(x\cos\alpha + y\sin\alpha + c(-x\sin\alpha + y\cos\alpha)^2\right)^2}{\sigma_x^2} + \frac{(-x\sin\alpha + y\cos\alpha)^2}{\sigma_y^2 s^2} \right) \right) \\ F^{\vec{b}}(x,y) &= \exp\left(if \left(x\cos\alpha + y\sin\alpha + c\,(-x\sin\alpha + y\cos\alpha)^2 \right) \right). \end{aligned}$$

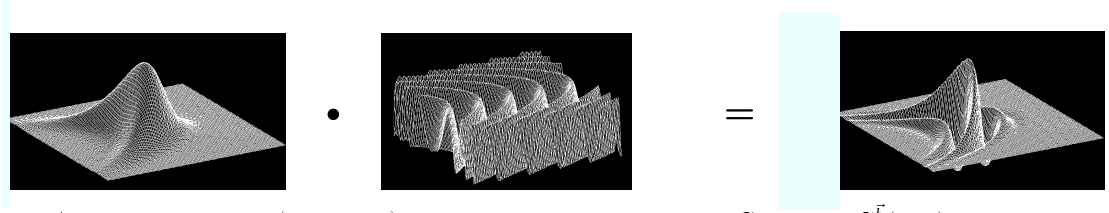

Figure 3: A banana wavelet (real part) is the product of a curved Gaussian $G^{\vec{b}}(x,y)$ and a curved wave function $F^{\vec{b}}(x,y)$ (only the real part of the kernel is shown).

Our basic feature is the magnitude of the filter response of a banana wavelet extracted by a convolution with an image:

$$\mathcal{A}I\left(\vec{x}_0, \vec{b}\right) = \left| \int B^{\vec{b}}\left(\vec{x}_0 - \vec{x}\right) I\left(\vec{x}\right) d\vec{x} \right|.$$

A banana wavelet $B^{\vec{b}}$ causes a strong response at pixel position $\vec{x}_0$ when the local structure of the image at that pixel position is similar to $B^{\vec{b}}$.
The Banana Space: The six–dimensional space of vectors $\vec{c} = (\vec{x}, \vec{b})$ is called the *banana (coordinate) space* with $\vec{c}$ representing the banana wavelet $B^{\vec{b}}(\vec{x})$ with its center at pixel position $\vec{x}$ in an image. In (Krüger,

Peters and v.d. Malsburg 1996) 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 $d(\vec{c}_1, \vec{c}_2)$ when their corresponding kernels are similar, i.e., they represent similar features.
Approximation of Banana Wavelets by Gabor Wavelets: The banana response space contains a huge amount of features, their generation requiring large computation and memory capacities. In (Krüger, Peters and v.d. Malsburg 1996) we define an algorithm to derive banana wavelets from Gabor wavelets which makes it possible to derive banana wavelet responses from Gabor wavelet responses. This approximation can be performed for all banana wavelet responses (we call it the *complete mode*) before matching (see below) starts. Alternatively, the Gabor wavelet responses can be calculated in a *virtual mode*, which means that only the much faster Gabor transformation is performed before matching, and only those banana wavelet responses are evaluated during matching which are actually required. Because of the sparseness of our representations of objects only a small subset of the banana space is actually used for matching and can therefore be evaluated very quickly. In the complete mode the hierarchical processing leads to a speed–up of a factor 5 compared to the computation of the banana wavelet responses directly from the image. In the virtual mode we can reduce memory requirements by a factor 20. Figure 4 gives the idea of the approximation algorithm the hierarchical processing is based on.

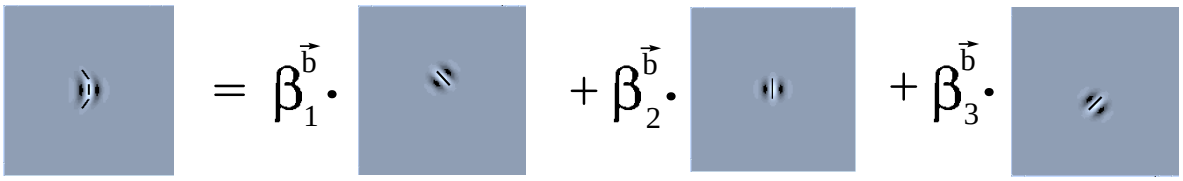

Figure 4: The banana wavelet on the left is approximated by the weighted sum of Gabor wavelets on the right.

2.2 Learning and Matching

Extracting Significant Features for one example: 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 for one example by two qualities. Firstly, it has to cause a strong response (**C1**); secondly, it has to represent a local maximum in the banana space (**C2**). Figure 5bi–iv shows the significant features for a set of cans (each banana wavelet is described by a curve with same orientation, curvature, and elongation).
C1 represents the requirement that a certain feature or similar feature is present, whereas C2 allows a more specific characterization of this feature. Banana responses vary smoothly in the coordinate space. Therefore the six–dimensional function $\mathcal{A}I(\vec{x_0}, \vec{b})$ is expected to have a properly defined set of local maxima. We can formalize C1 and C2 as follows: A banana response $\mathcal{A}I(\vec{x_0}, \vec{b_0})$ represents a significant feature for one example if

C1 $\mathcal{A}I(\vec{x}_0, \vec{b}) > T$, for a certain threshold T and

C2 $\mathcal{A}I(\vec{x}_0, \vec{b}_0) \geq \mathcal{A}I(\vec{x}_i, \vec{b}_i)$ for all neighbours of $(\vec{x}_0, \vec{b}_0)$.

Clustering: After extracting the significant features for different examples we apply an algorithm to extract important 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 extracted from our training examples. We assume the correspondence problem to be solved, i.e., we assume the position of certain landmarks (such as the tip of the nose or the middle of the right edge of a can) of an object to be known in images of different examples of these objects. In our representation each landmark is represented by a node of a graph. In some of our simulations we determined corresponding landmarks manually, for the rest we replaced this manual intervention by motor controlled feedback (see (Krüger, Peters and v.d. Malsburg 1996)). In a nutshell the learning algorithm works as follows: For each landmark we divide the significant features of all training examples into clusters. Features which are close according to our metric d are collected in the same cluster. A significant feature for an object class is defined as a representative of a *large* cluster. That means this or a similar feature (according to our metric d) occurs often in our training set. Small clusters are ignored by the learning algorithm. We end

up with a graph whose nodes are labeled with a set of banana wavelets representing the learned significant features (see figure 5bv).

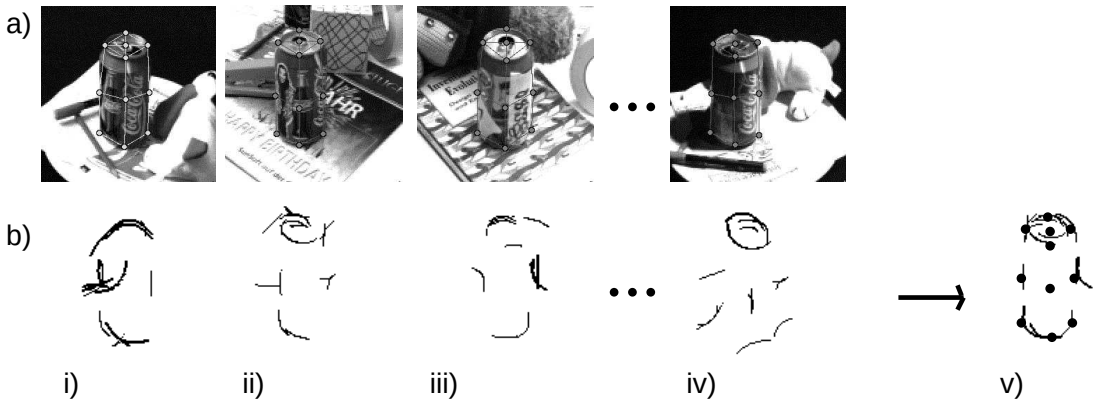

Figure 5: a: Pictures used for training. bi)–iv): Significant features for different cans describing, besides relevant information, also accidental features such as background, shadow or surface textures. c: the learned representation.

Matching: We use elastic graph matching (Lades et al. 1992) for the location and classification of objects. To apply our learned representation we define a similarity function between a graph labeled with the learned banana wavelets and a certain position in the image. A *graph similarity* simply averages *local similarities*. The local similarity expresses the system's confidence whether a pixel in the image represents a certain landmark and is defined as follows (for details see (Krüger, Peters and v.d. Malsburg 1996)): A local normalization function transforms the banana wavelet response to a value in the interval $[0,1]$ representing the system's confidence of the presence or absence of a local line segment. For each learned feature and pixel position in the image we simply check whether the corresponding normalized banana response is high or low, i.e., the corresponding feature is present or absent. During matching the graph is adapted in position and scale by optimizing the graph similarity. The graph with the highest similarity determines size and position of the object within the image.
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 (Lades et al. 1992; Wiskott et al. 1997) (which is based on Gabor wavelets) could be achieved. We also successfully performed matching with cans and other objects, as well as various discrimination tasks.
To improve the use of curvature in our approach we introduce a non–linear modification of the banana wavelet response. Using similar criteria than C1 and C2 we determine for each orientation in a local region whether the maximal banana wavelet responses represents a straight, convex, or concave line segment. The banana wavelet reponses corresponding to the detected class of curvature are inforced and the other banana wavelet response are reduced. For the problem of face finding we could achieve a further improvement of performance by this non–linear modification of the banana wavelet responses.

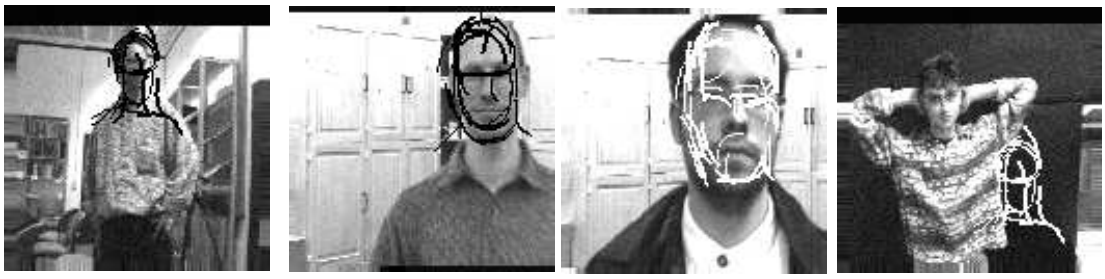

Figure 6: Face finding with banana wavelets. The mismatch (right) is caused by the person's unusual arm position.

3 Analogies to Visual Processing and Their Functional Meaning

In this section we discuss four analogies of our object recognition system to the visual system of primates concerned with

- feature selection,
- feature coding and hierarchical processing,
- sparse coding, and
- ordered arrangement of features.

For each item we first give a short overview of the current knowledge about its occurrence and functional meaning in brain processing as discussed in the literature. Then we describe the realization of the four aspects in our approach. At the end of each subsection, we discuss the relationship of the functional meaning of the aspects in the human visual system and the object recognition system. We are aware of the problem to discuss four fundamental aspects of visual processing within such a limited space. However, such a compressed description may enable the detection and characterization of important relationships between the different aspects.

3.1 Feature selection

According to (Hubel and Wiesel 1979) in the area V1 of primates there is a huge amount of local feature detectors sensitive to orientated edges, movement, or color. A data extension seems to arise from the retina to V1: for every position of the retina a large amount of features are extracted. In more recent studies also neurons maximally sensitive to more complex stimuli, such as cross–like figures (Shevelev et al. 1995) or curved lines (Dobbins and Zucker 1987) were found in striate cortex. A contributory factor for curvature as a feature computed preattentively (i.e., processed at very early stages of visual processing) arises from psychophysical experiments in (Treisman 1986) who showed that a curved line "pops out" in a set of straight lines. A question that is still open is the role of feedback in early stages of visual processing. It has been argued (Oram and Perrett 1994) that the short recognition time humans need for unknown objects (in the range of 100ms) makes computationally costly feedback loops unlikely. Others criticize this opinion, pointing to the huge amount of feedback connections between adjacent areas or to context sensitivity of cell responses (see, e.g., (Kapadia et al. 1995; Zipser, Lamme and Schiller 1976)).
For a representation of objects on a higher level of cortical processing in (Biederman 1987), psychophysical evidence is given for an internal object representation based on volumetric primitives called "geons" (see figure 7a–c). A selection of geons combined by basic relations such as "on top" or "left of" and the relative sizes of geons are used to specify objects.

In our object recognition system banana wavelets, i.e., "curved local lines detectors", are used as basic features which are given *a priori*. The restriction to banana wavelets gives us a significant reduction of the search space. Instead of allowing, e.g., all linear filters as possible features, we restrict ourselves to a small subset. We propose that a good feature has to have a certain complexity but an extreme increase of complexity resulting in a specialization to a very narrow class of objects has to be avoided. Banana wavelets fit this characterization well. They represent more complex features than Gabor wavelets but they are not restricted to a certain class of objects. Considering the risk of a wrong feature selection it is necessary to give good reasons for our decision. Firstly, the use of curvature improves matching because it is an important feature for discrimination. Secondly, and even more important, our application of banana wavelets mediates between a representation of objects in a grey level image and a more abstract, binary, and sparse representation. The representation of the gray level image by the full set of *continuous* banana wavelet responses allows for an almost complete reconstruction of the image, even a much smaller set of filters would be sufficient for this purpose. A representation of an object by *binarized* banana wavelets, e.g., by the corresponding curves, allows for a sparse representation of an object by its essential features. We think banana wavelets are a good feature choice because they enable an efficient representation of almost any object, because almost any object can be composed of localized curved lines. Aiming at a more abstract representation of objects embedded in

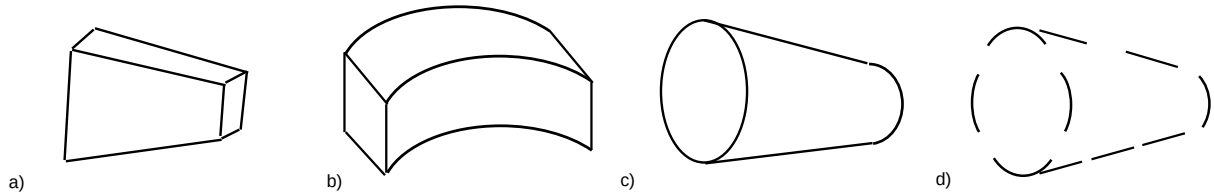

Figure 7: a–c) A subset of geons. d) A sparse representation of the geon in c) by banana wavelets.

Biedermans geon theory we argue that with an ordered set of curved lines we are able to represent the full set of geons in a sparse and efficient way (see figure 7d). Therefore we suggest that banana wavelets represent a suitable intermediate stage between lower and higher level representations of objects and we aim to define a framework in which geon–like constellations of line segments can be *learned* by visual experience.

Our feature selection is motivated by the functional advantages described above and the implicit necessity to decrease the dimension of the learning task, i.e., to face the bias/variance dilemma (see, e.g., (Geman et al. 1995)): If the starting configuration of the system is very general, it can learn from and specialize to a wide variety of domains, but it will in general have to pay for this advantage by having many internal degrees of freedom. This is a serious problem since the number of examples needed to train a system scales very poorly with the system's dimension, quickly leading to totally unrealistic learning time — the "variance" problem. 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 at all — the "bias" problem. Like any other learning system the brain has to deal with the bias/variance dilemma. As a consequence, it has to have a certain amount of *a priori* structure adapted to its specific input (the idea to overcome the bias/variance dilemma by appropriate *a priori* knowlege is elaborated in more detail in (Krüger 1997a; Krüger, Pötzsch and v.d. Malsburg 1997)). Being aware of the risk of a wrong feature selection leading to a system unable to detect important aspects of the input data — a wrong bias — and also being aware of the necessity to make such feature selection in order to restrict the dimension of the learning task — to decrease variance — we have chosen banana wavelets as a basic feature. We have justified this choice by the functional reasons given in the preceding paragraph and by the performance of our system.

In our system, feedback in the sense of local interaction of banana wavelet responses is used by the criteria C1 and C2 for sparsification and non–linear curvature enhancement. We also think that Gestalt principles can be coded within our approach by a similar kind of interaction. In a recent work (Krüger 1997b) we could give a mathematical characterization of the Gestalt principles collinearity and parallelism in natural images within the framework of our object recognition system.

Gabor wavelets as a subset of banana wavelets can be learned from visual data by utilizing the abstract principles sparse coding and information preservation (Olshausen and Field 1996). But does this kind of learning happen on the time scale of a human life? At least the experiment in (Wiesel and Hubel 1974), who have shown that an ordered arrangement of orientation columns develop in the visual cortex of monkeys with no visual experience, contradicts this assumption. The fact that Gabor wavelets result from an application of sparseness to natural image and not banana wavelets may support the objection that the suitability of curvature as basic feature does not necessarily follow from the statistics of natural images. However, the number of possible filters is restricted in (Olshausen and Field 1996). A similar learning algorithm enabling the coding with a larger number of filter may lead to additional attributes (such as, e.g., curvature) in the set of resulting filters. The more recent results about single cell responses in V1 (Shevelev et al. 1995) suggest that a larger set of features than Gabor Wavelet responses may be computed in V1.

To sum up this subsection, we have given references to biological and psychophysical findings which support the view that local curved lines are an important feature in early stages of visual processing. Furthermore, we have justified our feature choice by functional advantages of these features, such as discriminative power and the ability of an efficient representation of geons, for vision tasks. We have given reasons for the necessity of such a feature choice by pointing to the bias/variance dilemma which has to be faced by any learning system.

3.2 Feature Coding and Hierarchical Processing

In the visual cortex of primates hierarchical processing of features of increasing complexity and increasing receptive field size occurs. As a functional reason for processing of this type the advantages of capacity sharing, minimization of wiring length, and speed–up have been mentioned (see e.g., (Oram and Perrett 1994)). Different coding schemes for features are discussed in the literature. The concept of "local coding" in which one neuron is responsible for one feature (Barlow 1972) leads to problems: Because for each possible feature a separate neuron has to be used, a large amount of neurons is required. Another concept is called "assembly coding" (Georgopoulus 1990; Sparks, Lee and Rohrer 1990) in which a feature is coded in the activity distribution of a set of neurons. Assembly coding allows the coding of a larger amount of features for a given set of neurons, but the labeling of the set of active neurons with a certain feature remains a problem (see, e.g., (Singer 1995)).
In our object recognition system the main advantage of hierarchical processing is speed–up and reduction of memory requirements. We utilize hierarchical processing in two modes (see section 2.1): In the "complete mode" we gain a speed–up but no memory reduction and in the "virtual mode" we additionally gain a reduction of memory requirements.
In the virtual mode we utilize "local coding" and "assembly coding" for features of different complexity. In the virtual transformation we may interpret the Gabor transformation as a first description of the data (in a local coding). However, a response of a banana wavelet is coded in the distribution of Gabor wavelet responses and is only calculated if requested by the actual task of the system, i.e., if the matched representation comprises this specific feature. For frequently used low level features (such as Gabor wavelets) the advantage of fast data access outweighs the disadvantage of increase of memory requirements in the "local coding" concept. But for less frequently and more complex features (such as banana wavelet responses) the decrease of memory requirements may outweigh the increase of costs for a dynamical extraction of these features from lower level features (i.e., the costs of the interpretation of the actual activity distribution in the assembly coding). We do not claim that curvature is processed in the brain by assembly coding. Maybe curvature is such a frequently used feature that it is more likely to be computed in a local coding. Nevertheless, the general trade off between the two coding schemes can be exemplified.
To sum up, we have successfully applied hierarchical processing in our object recognition system resulting in speed–up and reduction of memory requirements. Furthermore, in our algorithm we have demonstrated the application of coding schemes which have analogies to coding schemes currently discussed in brain research and we have described their advantages and disadvantages within our object recognition system.

3.3 Sparse Coding

Sparse coding is discussed as a coding scheme of the first stages of visual processing of vertebrates (Field 1994; Olshausen and Field 1996). An important quality of this coding scheme is that "only a few cells respond to any given input" (Field 1994). If objects are considered as input this means that a certain feature is only useful for coding a small subset of objects and is not applicable for most of the other objects. Sparse coding has the biologically motivated advantage to minimize the wiring length for forming associations. (Baum, Moody and Wilczek 1988; Palm 1980) point to the increase of associative memory capacity provided by a sparse code. In (Olshausen and Field in press) it is argued that the retinal projection of the three–dimensional world has a sparse structure and therefore a sparse code meets the principle of redundancy reduction (Barlow 1961) by reducing higher–order statistical correlations of the input.

In our object recognition system a certain view of an object is represented by a small number of banana filters. The total amount of filters in the standard setting (2 levels of frequency, 12 orientations, 5 curvatures, and 3 elongations) is 360 filters per pixel (which can be reduced without loss of information in the banana domain because especially the low frequency filters are oversampled). Sparse coding is achieved by determining local maxima in the space of all filter responses (criterion C2), which only leads to about 60 responses remaining in an image of size 128×128, which means only 0,004 responses per pixel are needed (i.e., about 10^{-6} of all available features are required). Only these 60 responses are needed for the representation of an object in a 2D–view which means 0 up to 3 responses are needed for every node of the graph. In a setting without different curvatures and elongations, that is if Gabor filters are applied, the total number of responses is of same order compared to the above mentioned standard setting.

The aim of our system is to solve a certain task, namely the recognition of a class of objects, such as human heads or cans. The representation of an object class by sparse banana responses is validated by the success with which the system solves this task, as measured by the recognition rate. The principles of our approach differ from those in Olshausen's and Field's approach (Olshausen and Field 1996), who demand *information preservation* in addition to sparseness to create Gabor-like filters[1]. We doubt that information preservation is a suitable criterion for higher visual processing. The aim of human visual processing is to extract the information which is needed to survive and react in the environment by solving tasks such as interpreting a scene and recognizing an object or a facial expression, that means the aim is not *reconstruction* but *interpretation*. We believe that *task driven principles* should substitute the principle of information preservation for learning features of higher complexity. Our system creates abstract representations (see figure 2v) of a class of objects by reducing the information of a local area to line-like features from which graylevel images cannot be reconstructed. Since these representations can be recognized by humans and since linedrawings in general can be recognized as fast as graylevel images (Biederman and Ju 1988) this kind of abstract representation seems to contain the information needed to solve the recognition task.
In addition to the advantages of sparse coding already mentioned, we now discuss the following advantages: reduction of memory requirements, speed-up of the matching process, and simplification of determining relations among the features. A sparse code leads to representations with low memory requirements. In the former system (Lades et al. 1992; Wiskott et al. 1997) (from which our system is derived) an object is represented by a graph whose nodes are labeled by jets, where a jet contains the responses of a set of Gabor filters (all centered at the same pixel position). This kind of representation stores all the filter responses independent of whether they are needed to describe the considered object or not. Our representation only contains those responses which have high values before sparsification and thus represent the salient information of a 2D–view of an object in a specific scene. For the representation of one view of a specific object the required memory is reduced by a factor 40 compared to the former system and for the representation of classes of objects the reduction is even in the order of factor 1500.[2]
A functional advantage of a sparse representation is a fast matching process, since the time needed to compare a representation with the features at a certain position in an image goes nearly linearly with the number of features in the representation. This functional advantage is achieved on a sequential computer. Requiring only a small amount of processing capacities may be advantageous even for parallel systems, such as the brain, if many potential objects are tested in parallel.
Among others, (Biederman 1987) suggests that it is not a single feature which is important in the representation of an object but the *relations* among features. At the present stage of our approach only topographic relations expressed in the graph structure are represented. Banana wavelets represent features with certain complexity which describe suitable abstract properties (e.g., orientation and curvature). In future work we will aim to utilize these abstract properties to define Gestalt relations between banana wavelets, such as parallelism, symmetry, and connectivity. These abstract properties of our features enable the formalization of these relations. Furthermore, sparse coding leads to a decrease in the number of possible relations for an object description (only the relations between the few "active" features have to be taken into account). Therefore, the reduction of the space of relations and the describable abstract properties of these features make the space of those relations manageable. In the *reduction of the space of relations* we see an additional advantage of sparse coding which, to our knowledge, has not been mentioned in the literature.

In summary, sparse coding allows for representations with low memory requirements, which lead to a speed-up in the matching process. Furthermore, sparse coding potentially simplifies the determination of relations among features.

[1] Since the filters are not learned but only valued in our system, only the principles of both approaches and not the algorithms themselves can be compared.

[2] In the former system (Lades et al. 1992; Wiskott et al. 1997) a typical graph with 30 nodes and 40 complex-valued entries in every jet contains 2400 real values. If a representation of a class of objects is considered, even about 10^5 real values are needed because a bunch of about 50 object graphs is taken to represent a class of, e.g., human heads in frontal pose. For single objects the 2400 real values have to be compared with the about 60 integer values needed for our sparse representation (every of the about 60 binary features has to be stored by one index specifying the six labels frequency, orientation, curvature, elongation, x- and y-position). For classes of objects the 10^5 real values have also to be compared with about 60 integer values because our sparse representations of classes of objects require an amount of binary features which is similar to the amout for single objects.

3.4 Ordered Arrangement of Features

The order in the arrangement of features is a major principle aplied throughout the brain, both in the early stages of signal processing (Hubel and Wiesel 1979) and in the higher stages (Tanaka 1993). It is realized by computational maps (Knudsen, du Lac and Esterly 1987). These maps are organized in a columnar fashion. According to (Oram and Perrett 1994; Tanaka 1993) the columnar organization enables the assignment of a feature to a more general feature class (generalization) and also to discriminate between fine differences within a feature class (specialization).
In our system, the ordered arrangement of the banana features is achieved by the metric described in section 2.1. This metric defines a similarity between two features in the six-dimensional banana space. The metric organization of the banana responses is essential for learning in our object recognition system, because it allows to cluster similar features and thus to determine representatives for such clusters (section 2.2). By this kind of generalization we are able to reduce redundancies in our representation. A columnar organization is not yet defined in our system and thus general and special feature responses as described above are not distinguished. However, if columns may be defined as small local areas in the banana space, the criterion C2 utilized for the extraction of 'significant features for one example' may represent an intercolumnar competition giving a more specific coding of the unspecific response of the whole small region.
In summary, in our system an ordered arrangement of features is achieved by a metric in banana space. This metric enables a competition of neighboring features resulting in sparsified responses. Furthermore, the metric is essential to learn representations of object classes.

4 Conclusion and Outlook

We have discussed the biological plausibility of the artificial object recognition system described in (Krüger, Peters and v.d. Malsburg 1996). We were not interested in the detailed modelling of certain brain areas, we did not even utilize "neurons" or "hebbian plasticity" in our algorithm. Instead, we tried to apply principles of biological processing, i.e., we tried a modelling of the brain on a more *functional level*. As our system is already able to solve difficult vision tasks with high performance (not comparable to the human visual system but comparable to other artificial object recognition systems) we can evaluate the quality of our system by the *performance for a certain task*. This enables us to justify modifications of our system not by biological plausibility but by *efficiency*. As the human visual system is the best visual pattern recognition system we believe that biological plausibility and efficiency are not contradictory qualities but can be increased simultaneously. However, for this kind of interaction it is necessary to look at the brain as a system solving a certain task, i.e., as an algorithm. We think that the insight in abstract principles of cortical processing as utilized in our artificial object recognition system helps to create efficient artificial systems, and the understanding of the functional meaning of these principles in the artificial system can support the understanding of their role in cortical processing.
Following that line of thinking we have applied such principles within an object recognition system and have decribed their functional meaning: the bias/variance dilemma and the ability of a representation of objects on a higher level leads to a certain feature choice; local feedback is used for the processing of curvature and for sparsification; with hierarchical processing and a sparse representation we could reduce time for matching and memory requirements; the value of a sparse coding for the detection and utilization of Gestalt principles was discussed; the trade off between memory requirements and speed of processing for coding schemes such as "local coding" and "assembly coding" could be exemplified; and we have utilized an ordered arrangement of features for learning and redundancy reduction.
We claim that in the current state the object recognition system gives a reasonable model of the form path of lower levels of the visual system. We have demonstrated an efficient application of a biologically plausible feature selection utilizing sparse coding, hierarchical processing, and ordered arrangement of features. Concerning higher level processing, i.e., on a higher level than V1, we do not claim that our object recognition has the same plausibility. Giving just one example of an aspect of higher visual processing not covered by our system, we point to the Gestalt principles utilized by humans for the interpretation of natural scenes. However, we assume our system is also a good basis to model higher stages of visual processing, because it is a plausible approximation of lower stages of visual processing. In section 3 we already discussed a possible

representation of geon–like primitives based on banana wavelets as well as a possible formalization of Gestalt principles utilizing sparse coding and the abstract properties of our features. Furthermore, the ability to give a characterization of the Gestalt principles collinearity and parallelism within the framework of our object recognition system (Krüger 1997b) encourages us to merge biological plausibility and effectivity in an artificial systems with even better performance than in its current state.

Acknowledgement:

We would like to thank Christoph von der Malsburg, Jan Vorbrüggen, Peter Hancock, Laurenz Wiskott, Rolf Würtz, Thomas Maurer, Carsten Prodoehl, and two anonymous reviewer for fruitful discussions.

References

Atick, J. 1992. Could information theory provide an ecological theory of sensory processing? *Network*, 3:133–251.

Barlow, H. 1961. Possible principles underlying the transformation of sensory messages. *Sensory Communication*, pages 217–234.

Barlow, H. 1972. Single units and cognition: A neurone doctrine for perceptual psychology. *Perception*, 1:371–394.

Baum, E., Moody, J., and Wilczek, F. 1988. Internal representation for associative memory. *Biological Cybernetics*, pages 217–228.

Biederman, I. 1987. Recognition by components: A theory of human image understanding. *Psychological Review*, 94(2).

Biederman, I. and Ju, G. 1988. Surface vs. edge-based determinants of visual recognition. *Cognitive Psychology*, 20:38–64.

Biederman, I. and Kalocsai, P. Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society: Biological Sciences, in press.*

Dobbins, A. and Zucker, S. 1987. Endstopped neurons in the visual cortex as a substrate for calculating curvature. *Nature*, 329:438–441.

Field, D. 1994. What is the goal of sensory coding? *Neural Computation*, 6(4):561–601.

Földiák, P. 1990. Forming sparse representation by local anti–hebbian learning. *Biological Cybernetics*, 64.

Geman, S., Bienenstock, and Doursat, R. 1995. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58.

Georgopoulus, A. 1990. Neural coding of the direction of reaching and a comparison with saccadic eye movement. *Cold. Spring. Harb. Symp. Quant. Biol.*, 55:849–859.

Hubel, D. and Wiesel, T. 1979. Brain mechanisms of vision. *Scientific American*, 241:130–144.

Hummel, J. and Biederman, I. 1992. Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99:480–517.

Kapadia, M., Ito, M., Gilbert, C., and Westheimer, G. 1995. Improvement in visual sensitivity by changes in local context: Parallel studies in human observers and in v1 of alert monkeys. *Neuron*, 15:843–856.

E.I. Knudsen, S. du Lac, S.D. Esterly 1987. Computational maps in the brain. *Ann. Rev. Neuroscience.*, 10:41–65.

Krüger, N. 1997a. An algorithm for the learning of weights in discrimination functions using a priori constraints. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, July:764–768.

Krüger, N. 1997b. Collinearity and parallism are statistically significant second order relations of complex cell responses. Technical report, Institut für Neuroinformatik, Bochum, IRINI 97–15. http://www.neuroinformatik.ruhr-uni-bochum.de/ini/ALL/PUBLICATIONS/IRINI/irinis97.html.

Krüger, N., Peters, G., and von der Malsburg, C. 1996. Object recognition with a sparse and autonomously learned representation based on banana wavelets. Technical report, Institut für Neuroinformatik, Bochum. http://www.neuroinformatik.ruhr-uni-bochum.de/ini/ALL/PUBLICATIONS/IRINI/irinis96.html.

Krüger, N., Pötzsch, M., and von der Malsburg, C. 1997. Determination of face position and pose with a learned representation based on labeled graphs. *Image and Vision Computing*, August:665–673.

Lades, M., Vorbrüggen, J., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R., and Konen, W. 1992. Distortion invariant object recognition in the dynamik link architecture. *IEEE Transactions on Computers*, 42(3):300–311.

Olshausen, B. and Field, D. Sparse coding with an overcomplete basis set: A strategy employed by V1. *accepted for Vision Research.*

Olshausen, B. and Field, D. 1996. Natural image statistics and efficient coding. *Network*, 7:333–339.

Oram, M. and Perrett, D. 1994. Modeling visual recognition from neurobiological constraints. *Neural Networks*, 7:945–972.

Palm, G. 1980. On associative memory. *Biological Cybernetics*, 36:19–31.

Shevelev, I., Lazareva, N., Tikhomirov, A., and Sharev, G. 1995. Sensitivity to cross–like figures in the cat striate neurons. *Neuroscience*, 61:965–973.

Singer, W. 1995. Time as coding space in neocortical processing: A hypothesis. In Gazzaniga, M., editor, *The Cognitive Neuroscience*, pages 91–104. MIT Press.

Sparks, D., Lee, C., and W.H. Rohrer 1990. Population coding of the direction, amplitude and velocity of saccadic eye movements by neurons in superior colliculus. *Cold. Spring. Harb. Symp. Quant. Biol.*, 55:805–811.

Tanaka, K. 1993. Neuronal mechanisms of object recognition. *Science*, 262:685–688.

Treisman, A. 1986. Features and objects in visual processing. *Scientific American*, 255(5):114–125.

Wiesel, T. and Hubel, D. 1974. Ordered arrangement of orientation columns in monkeys lacking visual experience. *J. Comp. Neurol.*, 158:307–318.

Wiskott, L., Fellous, J., Krüger, N., and von der Malsburg, C. 1997. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 775–780.

Zipser, K., Lamme, V., and P.H. Schiller 1976. Context modulation in primary visual cortex. *Journal of Neuroscience*, 16(22):7376–7389.