Appeared in: Neural Computation 2001 (13), p:389–410.

Learning Object Representations using a priori Constraints within ORASSYLL *

Norbert Krüger
Institut für Informatik
Christian-Albrechts-Universität zu Kiel
Preußerstraße 1-9, 24105 Kiel, Germany
nkr@ks.informatik.uni-kiel.de

Abstract

In this paper a biologically plausible and efficient object recognition system (called ORASSYLL) is introduced which is based on a set of *a priori* constraints motivated by findings of developmental psychology and neurophysiology. These constraints are concerned with the organisation of the input in local and corresponding entities, the interpretation of the input by its transformation in a highly structured feature space and the evaluation of features extracted from an image sequence by statistical evaluation criteria. In the context of the bias/variance dilemma the functional role of *a priori* knowledge within ORASSYLL is discussed. In contrast to systems in which object representations are defined manually the introduced constraints allow an autonomous learning from complex scenes.

1 Introduction

The necessity of the existence of a certain amount of *a priori* knowledge within a system which has to deal with a high dimensional learning problem such as object recognition, is well-founded in, e.g. (Geman et al., 1995; Abu-Mostafa, 1995). However, the definite selection and formalisation for a specific task domain is still an open question. Many artificial object recognition systems *implicitly* apply a certain amount of *a priori* knowledge. The aim of the work presented here is to put the concrete choice of predetermined structural constraints in the focus of attention.

*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).

Plausible requirements for predetermined structural constraints are discussed and definitions of suitable constraints for object recognition are given. Furthermore, their formalisation within an artificial system called ORASSYLL (**O**bject **R**ecognition with **A**utonomous learned and **S**parse **SY**mbolic representations based on **L**ocal **L**ine detectors) is discussed. I will claim that findings of developmental psychology and neurophysiology indicate that the human visual system already possesses certain structural constraints at birth and supports their definition. The study of the primate's visual system enables us to look at the result of an evolutionary learning algorithm which has established predetermined structural constraints which may be extracted up to a certain degree from experimental data and which can be realized within technical systems. As a result of this controlled application of *a priori* knowledge and in contrast to approaches applying a manually designed object representation, within ORASSYLL model–based representations can be extracted autonomously or with only little manual intervention within a perception—action—cycle.

This work is organised as follows: In section 2 arguments for the necessity of *a priori* constraints are summarised and the relation of evolutionary learning and individual learning are discussed. Motivated by genetically determined structure in the human brain and observations of the behaviour of newborns and infants (described in section 3), I will define constraints for visual learning on which the object recognition system ORASSYLL is based (section 4). These constraints are concerned with the organisation of the input in local and corresponding entities utilising interaction of action and perception (section 4.1), the interpretation of the input by its transformation in a highly structured feature space resulting in a sparse object representation (section 4.2) and the evaluation of features extracted from an image sequence by statistical evaluation criteria (section 4.2). A technical description of ORASSYLL focussing on the learning of object representations and on the role of the introduced constraints is given in section 5.

2 The Bias/Variance Dilemma

Learning is inherently faced with the bias/variance dilemma (Geman et al., 1995): If the starting configuration of the system has many degrees of freedom, it can learn from and specialise to a wide variety of domains, but it will in general have to pay for this advantage by having many internal degrees of freedom —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 at all —the "bias" problem. As a conclusion (Geman et al., 1995) argue "bias needs to be *designed* to each particular problem". This conclusion is the starting point of this paper in which I will suggest concrete choices of bias in terms of structural constraints which are realized within ORASSYLL.

If artificial neural networks are considered as being caught in the variance part of the bias/variance dilemma, another type of object recognition system suffering from too much bias exists. These systems can be called "model based systems", see, e.g., (Yuille, 1991; Lanitis et al., 1997). As an example, (Lanitis et al., 1997) can successfully locate faces by matching a manually defined face model with a certain number of free parameter enabling the adaptation to a specific face in a specific pose. Because the model of the face is defined manually, each time the algorithm is applied to a new object class, a new representation has to be designed manually again. In this way, for example in (Cootes et al., 1995), resistors are localised within the framework of the object representation in (Lanitis et al., 1997).

Each concrete choice of *a priori* knowledge is a crucial point. A wrong choice may lead to the exclusion of good solutions in the search space. An amount of predetermined knowledge that is too restricted may result in an increase of the search space, leading to unrealistic learning time and bad generalisation. This trade–off requires firstly that the *a priori* constraints should, on the one hand, be powerful in the sense that they cover essential structure of the problem they deal with. Secondly, they should be general, such that they can be applied in many situations and do not restrict the system to deal only with very specific sub–problems. Of course, these two necessary properties are not sufficient to give a unique definition for structural constraints. Indeed there is a certain amount of arbitrariness, and a proof for THE *a priori* constraints (e.g., based on the statistics of the input data of the visual system and a precise definition of the task it has been designed for), is far beyond the stage of visual science today and of course far beyond the scope of this work.

How can we escape the bias/variance dilemma? The existence of a pattern recognition system — the human visual system — able to deal with its surroundings efficiently *and* with sufficient adaptivity raises hope about this possibility. The predetermined structural constraints have evolved during evolution and appear to be well suited to organise visual experience. Therefore they seem to cover essential structure of the physical world and it is a valuable opportunity to look at results of biology to become inspired for suitable definitions of constraints. In this sense, nowadays the Kantian idea (Kant, 1781) to establish a *table* of *a priori* constraints which organises perception can be supported, guided and justified by a good amount of neuropysiological and psychophysical data which is discussed in the next section.

3 The Development of the Visual System

To motivate the constraints applied within ORASSYLL, a short description of the development of human visual skills is given. I will restrict myself to the aspects which are relevant for predetermined structural knowledge applied within ORASSYLL. In subsection 3.1 the observable (regarding the behaviour of infants) or "outer" development of the human visual system is summarised. In subsection 3.2 some aspects of the neurophysiological or "inner" correlate of this development are described. Both ways of description, "inner" and "outer", give evidence that the human brain neither is a blank table nor a system which is completely determined at birth but confirm that structural knowledge is already existent and guides post-natal learning.

3.1 Developmental Psychology of Visual and Gripping Abilities

Newborns already possess a remarkable set of visual abilities. They are able to distinguish lines (Rauh, 1995) and colours (Jones-Molfese, 1992). Movement, high contrast, and faces are "interesting features" to which the newborn infant directs its attention (Goren, 1975). Furthermore, newborns are able to follow (clumsily) a moving object by rotating their head and eyes (Rauh, 1995). Spelke (Spelke, 1993) demonstrated that the Gestalt principle "common fate" is already utilised by infants of an age of four months but also demonstrated the appearance of the Gestalt principles "collinearity" and "parallelism" at an age of seven months.

The process of gaining control of arms and hands and their interaction with the visual system develops within the first six months. At approximately 4–6 months an infant can perform a visually controlled movement of its arms and hands. Infants under the age of 4 months mainly characterise an object by its movement or position. They perceive an object as "something at a certain position" or "something moving with a certain velocity". Objects have no "above, below, left, right, in front or behind" (Bower, 1971). In coincidence with gaining visual control about their movement the object representation of infants starts to be based on higher features, such as, form and size (Bower, 1971). An interesting analogy of ORASSYLL and human aquisition of object representation is the use form features by infants at an age when they are able to carry out a perception—action—cycle which is also necessary within the artificial system presented here (see section 4.1).

3.2 Development of the Visual System: Neurobiology

The relation of intrinsic properties of brain areas to extrinsic influences on the structure of the cortex is a controversially discussed question. As an extreme viewpoint (Creutzfeld, 1977) proposed that all cortical neurons are initially equipotent and that laminar and areal differences in the organisation of the cortex are induced exclusively by extrinsic influence. There exist indeed impressive phenomena of extrinsic effects on structuring of brain areas (see e.g., (Sur et al., 1988; Blakemore and Cooper, 1970)). However, also a considerable number of counterexamples show the restricted adaptivity of the visual system, e.g., in the case of strabism and astigmatism (Atkinson and Braddick, 1989). Furthermore, the evolution of complex structures without post–natal visual experiences, e.g., orientation maps (Wiesel and Hubel, 1974; Gödecke and Bonhoeffer, 1996), confirms the importance of genetic predetermination.

In (Krüger, 1998b) I give a short overview about the results of research about the cortogenesis of the visual system which are relevant for the choices of constraints within ORASSYLL. Summarising the most important items, it is argued that the connections of brain areas and the receptive field size of neurons in different areas are largely predetermined and established at birth. Even the features extracted in some areas (orientation, movement and colour) (Wiesel and Hubel, 1974; Gödecke and Bonhoeffer, 1996), i.e., the coarse sensitivity of neurons, are basically initiated before the first post–natal visual experience. Other features, such as extraction of disparity information, depend on extrinsic influence and do not develop without it. Input dependent fine tuning and local normalisation processes (parallel to the development of lateral synaptical connections) develop during the first months of visual experience. The arrangement of features in computational maps (Knudsen et al., 1987) is a major principle applied throughout the brain, both in the early stages of visual processing (Hubel and Wiesel, 1979) as well as in higher stages (Tanaka, 1993) and probably, at least for area V1, is genetically determined (see (Wiesel and Hubel, 1974; Gödecke and Bonhoeffer, 1996)).

The utilisation of Gestalt principles (such as, collinearity and parallelism) also evolves with visual experience (see section 3.1), possibly making use of statistical properties of natural images (see also (Phillips and Singer, 1997)). It is an interesting "by–product" of ORASSYLL that the Gestalt principles collinearity and parallelism can be detected as significant relations of the class of natural images after interpreting the Gabor wavelets according to the constraints applied within ORASSYLL (see section 4.2 and (Krüger, 1998a)).

Developmental psychology and neurophysiological research give indications for the impressive adaptivity of the visual system and its capability to extract significant information from experience. Both ways of description also indicate a large amount of genetic prestructuring. Maybe the important question is not so much the *relative weight* of genetic predetermination and adaptation but a *precise definition* of the predetermined structural constraints. The experiments in striate cortex (Wiesel and Hubel, 1974; Gödecke and Bonhoeffer, 1996) already give good hints about predetermined structures, such as feature choices and their organisation in computational maps.

4 The *a priori* Constraints

Inspired by the constraints imposed by the human visual system (as described in section 3), in this section predetermined structural constraints for visual learning are introduced which will be realized within ORASSYLL. Each pattern recognition system has to apply a certain amount of *a priori* knowledge. What may be different in this work is that these structural constraints are put into the focus of attention and are the starting point of designing the object recognition system. For the constraints I will discuss analogies to constraints imposed by the human visual system (see also table 1) and differences and relations to design choices in other systems.

I do not assume that the constraints introduced here are complete in the sense that they cover all necessary constraints for an object recognition system which is as efficient as the human visual system. Firstly, here the focus is on

Constraint	Analogy in Visual System	Functional Role within ORASSYLL
PL1 Independence	localised receptive fields (Hubel and Wiesel, 1979)	reducing complexity by splitting the input space in subspaces
PL2 Correspondence	interaction of arm movement and perception (perception–action–cycle) (Rauh, 1995; Koenderink, 1992)	learning with comparable entities
PF1 Feature Selection	genetically determined orientation sensitive Gabor–shaped neurons (Wiesel and Hubel, 1974; Jones and Palmer, 1987)	reduction of search space by forcing description by specific features with symbolic meaning
PF2 Feature Arrangement	genetically determined order of orientation maps in V1 (Wiesel and Hubel, 1974) or (Gödecke and Bonhoeffer, 1996)	combination of similar and separation of dissimilar entities by a metric
PF3 Hierarchical Processing	applied within whole cortex (see, e.g. (Oram and Perrett, 1994))	sharing of resources, speed up of feature processing
PF4 Sparse Coding	equal response probability of neurons across images and low response probability for a single image (Palm, 1980; Field, 1994), date spread in V1 compared to ganglia cells	low memory requirements and high storage capacity, reduction of space of relations within object representations, speed up of matching
PE1 Maximal Discrimination		speed up of learning by internal evaluation of features instead of e.g., error back-propagation
PE2 Minimal Redundancy	transformation of redundancy of input data into cognitive maps (Barlow, 1961)	reduction of redundancy by representing similar entities by one entity utilizing metric

Table 1: Constraints (first column), analogy in human visual system (second column) and functional role within ORASSYLL (third column).

shape processing and other clues such as colour or disparity information are ignored. Secondly, in its current form ORASSYLL is a 2D–approach. Thirdly, for object representation on higher stages of visual processing (higher than V1) additional constraints (e.g., Gestalt principles) probably have to be taken into account. However, the constraints utilised within ORASSYLL are already sufficient to extract autonomously 2D-representations of objects from images and these representations can be applied to difficult vision tasks.

The *a priori* constraints can be divided into constraints concerning the organisation of the input (PL1–2, subsection 4.1), constraints concerning feature selection, feature organisation, and feature processing (PF1–4, subsection 4.2), and constraints concerning statistical feature evaluation (PE1–2, subsection 4.3). Their relationship to biological and psychological findings and their functional role within ORASSYLL is summarised in table 1.

4.1 Locality and the Correspondence Problem

The system's input is spatially organised: the input is divided into non–overlapping subparts (PL1: independence) and reflects a certain consistency of moving ob-

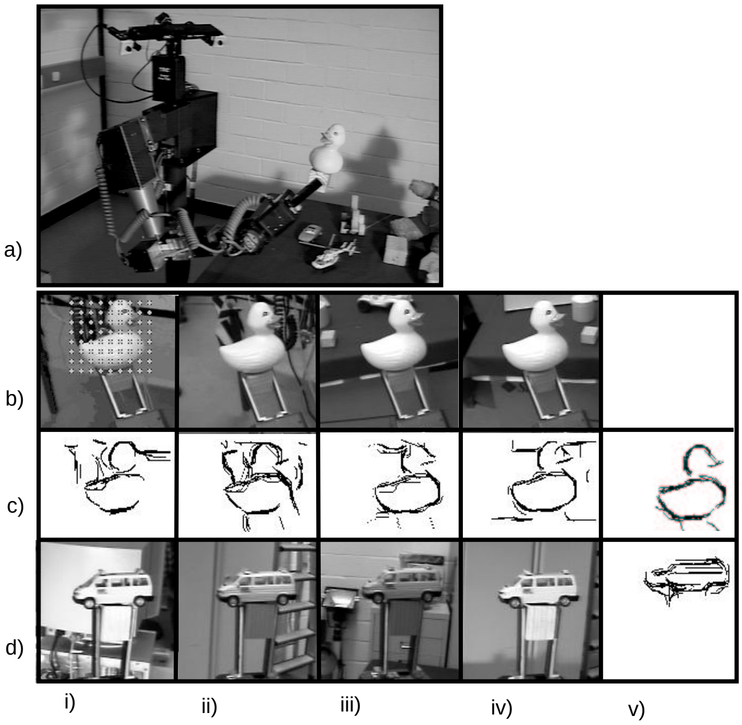

Figure 1: **a)** The robot arm with the camera. **b)** The "retinal" images produced by following the robot arm holding a toy–duck. **c,i–iv)** Significant Features per Instance extracted in an rectangular region (shown in b,i). **c,v)** Learned representation. **d)** Training data and learned representation for a toy car.

jects in time (PL2: correspondence). This organisation of the input decreases the relevant feature space and facilitates learning.

PL (Locality)

(1) (Independence) Features at distant locations are assumed to be independent.

(2) (Correspondence) Only features corresponding to the same landmark for different examples of the same object class are used for learning.

PL1: The first part of the constraint PL already implies the locality of features: a local feature, i.e., a feature which describes a quality of a local part of an image, does not interact with features corresponding to other landmarks. Corresponding to the limited receptive field size of neurons in V1 the features will be local (see PF1). It has been demonstrated that local filters similar to the filters applied within ORASSYLL are the result of feature extraction by Independent Component Analysis (Bell and Sejnowski, 1996).

The splitting of the feature space in smaller independent subspaces is a design principle of many artificial systems (e.g., (Alpaydin and Jordan, 1996; Wiskott et al., 1997)), nevertheless others (e.g., (Turk and Pentland, 1991)) apply *global* functions to the input image.

PL2: The second part of PL comprises a fundamental problem of vision, the "correspondence problem". For any learning algorithm, it is indispensable to ensure that comparable entities (i.e., comparable landmarks) are used as training data. Looking at a single image of an object makes it difficult to distinguish between features corresponding to the background and features corresponding to the object; if the object to be learned is moving, a number of factors will produce high variation in the data: *i) Translation* of the object leads to the appearance of corresponding landmarks at different positions in the image and varying background and scale; *ii) Rotation* may lead to the occurrence of very dissimilar views of the same object; *iii) Variation of illumination* causes shadows on the object and amplifies or diminishes the occurrence of textures or edges.

A sensible learning of a certain view of an object seems to be impossible if not at least some of these sources of variation are eliminated. The easiest way is to solve the correspondence problem manually, but of course this is a serious restriction. At least for one instance of an object class, interaction of motor control and vision can help to create controlled training data: 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, training data is produced 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 figure 1a,b). The method is comparable to an arm movement and grasping controlled by vision such as 4–6 months old infants are able to (see section 3.1) and reflects the strong relation between action and perception (Koenderink, 1992; Sommer, 1997). Interestingly the infants' concept of objects changes dramatically at this stage of development (see section 3.1). The ability to create a situation in which an object appears under controlled conditions may help, as in the object recognition system, to learn a suitable representation of objects.

4.2 Feature Selection and Feature Organisation

The spatially organised input is transformed to a structured feature space, or, in other words, the input is seen through the "glasses" of this feature space. The *a priori* constraints introduced here are concerned with the choice of features itself and their transformation to a more "symbolic" meaning (PF1: feature selection), their inter–relation (PF2: feature arrangement), their computation (PF3: hierarchical processing) and their role within the object representation (PF4: sparse coding). An important difference to other object recognition systems is the exploitation of a rich structure of the feature space for learning: The locality of features and the mechanism described in figure 1 ensures that only comparable local entities are input for learning. These features can be compared by a metric of the feature space. Interestingly, the specific interpretation of Gabor wavelet responses within ORASSYLL allows for the detection of Gestalt principles in the input data.

PF (Feature Assumptions)

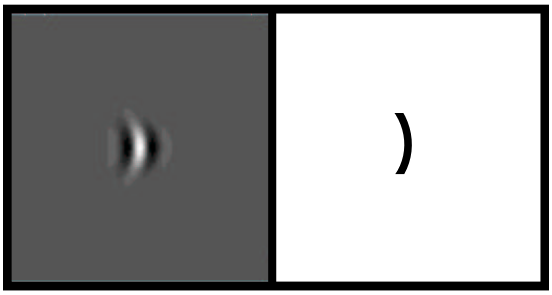

Figure 2: Path corresponding to a banana wavelet. a: Arbitrary wavelet. b: Symbolic representation by the corresponding curve. c: Visualisation of a representation of an object class.

---

(1) (Feature Selection) Significant features of a localised area of the two–dimensional projection of the visual world are localised (curved) line segments.

(2) (Feature Metric) A metric defines a distance between these features indicating the differences in their properties orientation, curvature and position.

(3) (Hierarchical Processing) These features are computed from simpler features in a hierarchical fashion.

(4) (Sparse Coding) An object is coded as a sparse and spatially ordered arrangement of these features.

PF1: In ORASSYLL Gabor or banana wavelets and their symbolic analogue, "(curved) local line segments", are used as basic features which are given *a priori* (see figure 2). The restriction to Gabor or banana wavelets gives a significant reduction of the search space. Instead of allowing, e.g., all linear filters as possible features (as realized, e.g., in the scalar–product of a back propagation neuron), a restriction to a small subset is imposed. Neurophysiological experiments (Wiesel and Hubel, 1974) (see also section 3.2) show that also the human visual system makes certain kind of feature choices before their first acts of postnatal visual experience.

Within the framework of ORASSYLL it has been shown in (Krüger, 1998a) that the Gestalt principles collinearity and parallelism can be detected and described as second order statistics of *normalised* Gabor wavelet responses. This non-linear normalisation was initially developed to solve a certain subproblem within ORASSYLL: The comparison of a symbolic feature with a certain position within the grey–level image. The normalisation transforms Gabor wavelet responses such that they express the system's confidence for the presence or absence of a local line segment, i.e., it represents their interpretation in terms of constraint PF1. Surprisingly, this transformation effects the second order statistics of Gabor wavelet responses significantly (see figure 3a–d). Looking only at the non–normalised Gabor wavelet responses the Gestalt principles collinearity and parallelism are barely detectable (see figure 3e–f).

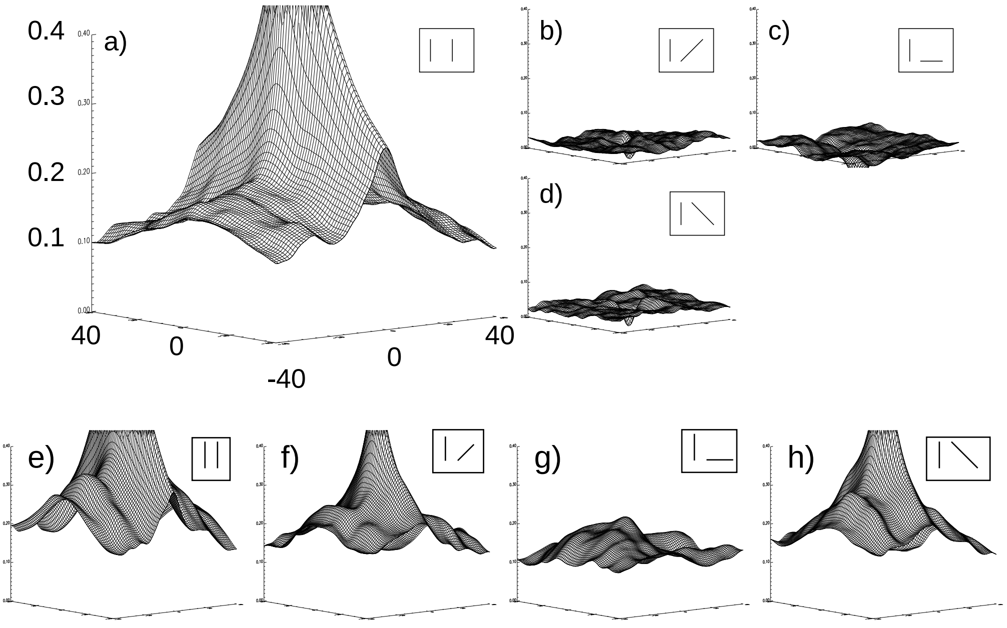

Figure 3: The cross–correlation of pairs of normalised Gabor wavelet responses (a–d) and unmodified Gabor wavelet responses (e–h) of four orientations on a large set of natural images: a,e) horizontal - horizontal, b,f) horizontal - diagonal, c,g) horizontal - vertical, d,h) horizontal - diagonal. The x- and y-axes represent the separation of the kernels (labeling of all axes for a–h is the same than in a) and the z-axis represents the correlation. In a) parallelism and collinearity are clearly visible: Collinearity is detectable as ridge in the first diagram and parallelism appears as global property expressed in the flat part of the surface in the first diagram clearly above the surfaces corresponding to non–parallel orientations.

PF2: In ORASSYLL, the ordered arrangement of features is achieved by a metric defining a distance between features indicating their differences in the properties orientation, curvature and position. This metric organisation is essential for learning in the object recognition system, because it allows to cluster similar features and thus to determine representatives for such clusters.
PF3: In the visual cortex of primates hierarchical processing of features of increasing complexity and increasing receptive field size occurs. In the object recognition system the main advantage of hierarchical processing (see figure 4) is speed–up and reduction of memory requirements. Hierarchical processing is a widely used principle, realized in most of the neural network systems.
PF4: Sparse coding is discussed as a coding scheme realized in the primate's visual system (Field, 1994). A sparse representation can be defined as a coding of an object by a *small number* of *binary* features taken from a *large feature space.* In ORASSYLL an expansion of the feature space is forced by extracting a number of features for each pixel. For the representation of an object only about 10^{-6} of all available features are required. In this sense, the objects (see, e.g., figure 2c) are represented sparsely. ORASSYLL differs in this aspect to many other object recognition systems which apply compact representations (e.g., (Turk and

Pentland, 1991)). The sparseness of representation allows a fast matching because only a few features have to be checked within an image. Furthermore the space of possible feature relation within the object representation is reduced in a sparse representation. Taking into account that important aspects are coded by these relations (e.g., Gestalt relations), this potentially facilitates learning within the space of multiple order relations.

4.3 Evaluation of Features

In contrast to the constraints (PF1–PF4) which define the feature space itself, now two criteria are introduced for the selection of "good" features to represent an object class. These constraints enable the system to use multiple visual experiences of instances of the same object class to extract significant information. In contrast to, e.g., back-propagation ((Rumelhart et al., 1986)), in which a "global" criteria is optimised, the two constraints represent criteria which guide and speed up the learning algorithm by evaluating intermediate stages of processing (see also the discussion of the "credit assignement problem" in, e.g. (Arbib, 1994)).

PE (Evaluation)

(1) (Maximal Discrimination) Features are preferred whose values on images vary little within classes and vary much between classes.

(2) (Minimal Redundancy): Redundant information shall be eliminated from the system.

PE1, PE2: Analogies to the constraints PE1 and PE2 cannot be detected by biological experiments, as it is possible for the constraints PF1–PF4. However, (Barlow, 1961) discusses redundancy reduction as an important principle underlying the transformation of sensory messages in the brain. Furthermore, both constraints are applied implicitly in many pattern recognition algorithms (see, e.g., (Fisher, 1923; Krüger, 1997)).

5 Description of the Object Recognition System

In this section I will give a technical description of ORASSYLL which is based on the *a priori* constraints introduced in section 4. Here it is not my aim to give a detailed technical description of the whole system but to describe how the applied constraints enable learning of object models for realistic and difficult tasks (for details concerning the complete system, see (Krüger, 1998b; Krüger and Peters, 2000)).

5.1 The Feature Space

In this subsection the realization of the constraints PF1–PF4 (introduced in section 4) is described. By utilising these constraints the input is transformed into

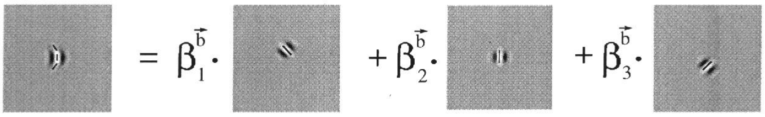

Figure 4: Hierarchical processing: The more complex banana wavelet on the left is approximated by the weighted sum of Gabor wavelets on the right.

a metrically organised feature space. Images are interpreted as an assemble of spatially organised local line segments which are processed hierarchically.

Gabor and 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).[1] 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 $F^{\vec{b}}$ (see figure 2a).

Banana wavelets are generalised Gabor wavelets (for Gabor wavelets see, e.g., (Daugman, 1985)), they possess additional to frequency and orientation the parameters curvature and elongation. The approach introduced here does not necessitate the usage of banana wavelets and is also applicable with Gabor Wavelets (see figure 5d,i,ii,iv for object representations with only straight line segments).

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 wavelet $B^{\vec{b}}$ with its center at pixel position $\vec{x}$ in an image. PF2 is realized by 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. For the exact definition of the metric first a distance measure is defined for the orientation–curvature subspace (α, c) expressing the Moebius topology thereof. Setting

$$d((\alpha_1, c_1), (\alpha_2, c_2)) =$$
$$\min\left\{\sqrt{\tfrac{(\alpha_1-\alpha_2)^2}{e_\alpha^2} + \tfrac{(c_1-c_2)^2}{e_c^2}},\ \sqrt{\tfrac{((\alpha_1-\pi)-\alpha_2)^2}{e_\alpha^2} + \tfrac{(c_1+c_2)^2}{e_c^2}},\ \sqrt{\tfrac{((\alpha_1+\pi)-\alpha_2)^2}{e_\alpha^2} + \tfrac{(c_1+c_2)^2}{e_c^2}}\right\}$$

on the subspace (α, c). Then a distance measure on the complete coordinate space is defined by

$$d(\vec{c}_1, \vec{c}_2) =$$
$$\sqrt{\frac{(x_1-x_2)^2}{e_x^2} + \frac{(y_1-y_2)^2}{e_y^2} + \frac{(f_1-f_2)^2}{e_f^2} + d((\alpha_1, c_1), (\alpha_2, c_2))^2 + \frac{(s_1-s_2)^2}{e_s^2}}. \quad (1)$$

The values $\vec{e} = (e_x, e_y, e_f, e_\alpha, e_c, e_s)$ define a cube of volume 1 in the features space, i.e. they define the weights for the different properties such as orientation and curvature. In the standard settings $\vec{e} = (4, 4, 0.01, 0.3, 0.4, 3.0)$ is used.

Non–Linear Transformations of the Filter Responses The feature processing of ORASSYLL consists of a two–step non–linear transformation of the

[1] For Gabor wavelets $\vec{b}$ is reduced to (f, α)

complex filter responses. In the first step, the magnitude of the filter response is extracted after the convolution of $B^{\vec{b}}$ with the image I. In contrast to the complex filter response oscillating with phase, the magnitude of the response is more stable under slight variation of position, see (Pötzsch et al., 1996). A filter $B^{\vec{b}}$ causes a strong response at a certain pixel position when the local structure of the image at that pixel position is similar to $B^{\vec{b}}$.

The magnitude of the filter responses depends significantly on the strength of edges in the image. However, here I am only interested in the presence and not in the strength of edges. Thus, in a second step a function $N(\ )$ normalises the real valued filter responses $r(\vec{c})$ into the interval $[0, 1]$. The value $N\left(r(\vec{c})\right)$ represents the likelihood of the presence or absence of a local line segment corresponding to $\vec{c} = (\vec{x_0}, \vec{b})$. This normalisation is based on the "Above Average Criterion":

AAC a line segment corresponding to the banana wavelet $\vec{c}$ is present if the corresponding banana wavelet response is distinctly above the average response.

The normalisation is realized by mapping $r(\vec{c})$ by a sigmoid function N. $N\left(r(\vec{c})\right)$ returns a small value, when $r(\vec{c})$ is below an average response E and a high value if it is close to the maximum response Max. Both parameters, average E and maximum Max, are computed using local *and* global information. Because of this locality, they vary with pixel position. The influence of the normalisation to the second order statistics of Gabor wavelet responses is demonstrated in figure 3.

Approximation of Banana Wavelets by Gabor Wavelets: To reduce computational requirements for the extraction of the large feature space an algorithm to approximate banana wavelets from Gabor wavelets and banana wavelet responses from Gabor wavelet responses is defined (see figure 4). By this hierarchical processing (PF3) a speed up to a factor 5 can be achieved.

5.2 Learning

In this subsection a sparse object representation (PF4) is extracted from single images or image sequences. Features caused from background structure or illumination can be eliminated by a learning scheme which makes use of the structure of the feature space (PF1, PF2). The learning algorithm applies the evaluation criteria PE1 and PE2 as internal criteria to determine significant features for an object class. It presupposes the correspondence of local landmarks in different images (PL1, PL2) which can be achieved by interaction of perception and action.

Extracting Significant Features Per Instance: Here the aim is to extract the local structure in an image in terms of (curved) local line segments. A *significant feature per instance* of an object is defined 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 1c,i–iv, 5b,d) and 7b,i–iv)) show the significant features per instance for some objects. Each wavelet is described by a curve with same orientation, curvature and size. Lower frequencies are represented as thicker line segments.

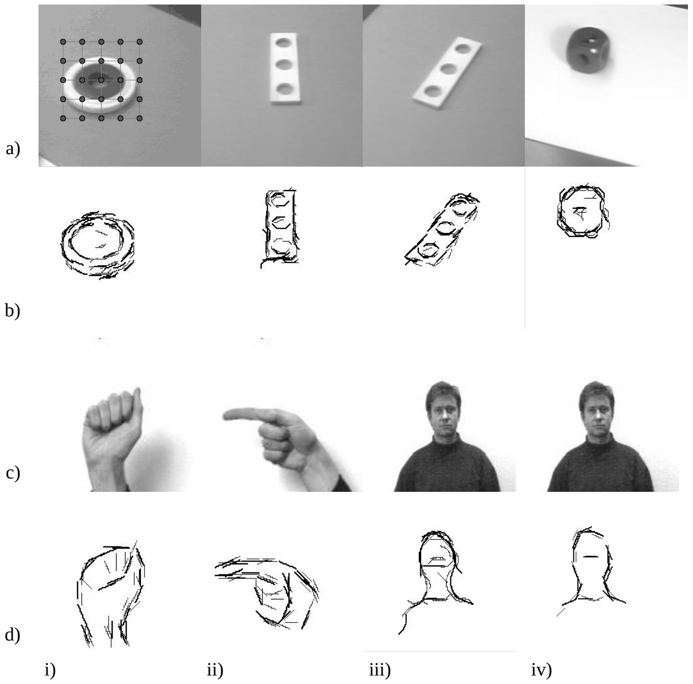

Figure 5: One–shot learning: Row a) and c) show the objects to be learned in front of homogeneous background. Row b) and d) show the extracted representations. For all objects a rectangular grid was roughly positioned on the object as in the first image a,i).

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 inter-columnar competition giving a more specific coding of this feature (Oram and Perrett, 1994). Therefore, the feature space is divided in locally distinct columns (PL1) in which related features (or features with close distance (PF2)) are represented.

One–shot learning: By positioning a rectangular grid on an object (see figure 5a,i) in front of homogeneous background and extracting significant features per instance as described above, suitable representations of objects can be extracted. These representations were already successfully applied to difficult discrimination tasks. For instance, for a difficult 10–class problem in hand posture classification a recognition rate of 80% could be achieved with representations extracted from single images (see row 2 in table 2).

Clustering: In case of non–homogeneous background and uncontrolled illumination one shot–learning would create representations with line segments corresponding to background (see figure 1c,i–iv). In this case a more sophisticated learning scheme has to be applied: After extracting the significant features per instance in different pictures an algorithm to extract invariant local features for a *class of objects* is applied. Here the task is the selection of the *relevant features* for the object class from the noisy features extracted from the training examples (see figure 7b,i–iv and 1c,i–iv). A significant feature should be independent of background, illumination or accidental qualities of a certain example of the ob-

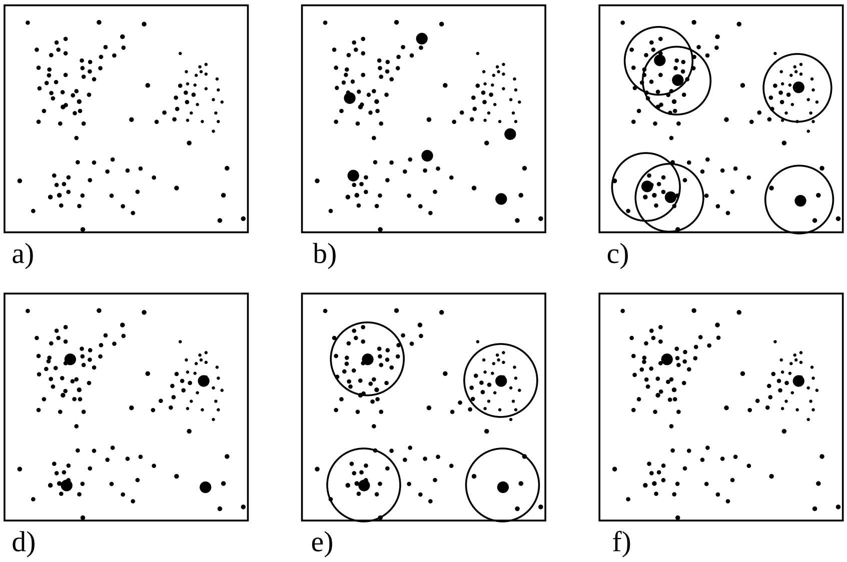

Figure 6: Clustering: a) Distribution of the significant features per instance extracted at a certain landmark. b) Codebook initialisation. c) Codebook vectors after learning. d) Substituting sets of codebook vectors with small distance by their center of gravity. e) Counting the number of elements within a certain radius. f) Deleting codebook vectors representing insignificant features.

ject class, i.e., it should be invariant under these transformations (PE1). This is realized within the algorithm by measuring the probability of occurrence of features in a local area of the feature space for different examples. The metric (PF2) allows the grouping of similar features into one bin, but it also allows the reduction of redundant information (PE2) by avoiding multiple similar features in the learned representation. In this way it becomes possible to learn sparse object representations (PF4) in very difficult situations (see figure 7). The correspondence problem (PL2) is assumed to be solved, i.e., it is assumed that the position of certain landmarks of an object to be known on pictures of different examples of these objects. In figure 7 corresponding landmarks are determined manually, in figure 1 this manual intervention is substituted by motor controlled feedback

The learning algorithm works as follows (illustrated for two dimensions in figure 6): **a)** Let $\mathcal{I}$ be a set of pictures of different examples of a class of objects of certain orientation and approximately equal size. $I^{(j,k)}$ represents a local area in the j-th image in $\mathcal{I}$ with the k-th landmark as its center. Let $\vec{s}_{ij}^{k}$ be the i-th significant feature per instance extracted in the area $I^{(j,k)}$. All $\vec{s}_{ij}^{k}$ for a specific k are collected in one set S^k. **b)** Then the LBG–vector quantisation algorithm (Linde et al., 1980) is applied to S^k (see figure 6). After vector quantization a codebook C^1 expresses the vectors $\vec{s}_{ij}^{k}$ with a certain number n_{C^1} of code book vectors $\vec{c}_i^{1} \in C^1 \subset \mathcal{C}, \vec{c}_i^{1} : 1, \ldots, n_{C^1}$ (figure 6b).

The LBG–algorithm reduces the distortion error, i.e., the average error occurring when all elements of S^k are replaced by the nearest codebook vector in C^1. In case of high densities of elements $\vec{s}_{ij}^{k}$ in S^k it may be advantageous in terms of the distortion error to have code book vectors $\vec{c}$ and $\vec{c}'$ with small distance $d(\vec{c}, \vec{c}')$ (PF2). But the significant features for a certain class of objects are expected to express independent qualities (L1), i.e., they are expected to have large distances in the feature space. **c,d)** Therefore a smaller codebook C^2 is constructed in which the $\vec{c}, \vec{c}' \in C^1$ with close distances are replaced by their center of gravity: Let $r_1 \in I\!R^+$ be fixed. For all $\vec{c} \in C^1$ the number of $\vec{c}' \in C^1$ with distance $d(\vec{c}, \vec{c}') < r_1$ (figure 6c) is computed. If there exists at least one such $\vec{c}' \neq \vec{c}$ all the codebook vectors in C^1 with $d(\vec{c}, \vec{c}') < r_1$ are substituted by their center of gravity (figure 6d). C^2 now represents a code book with less or equal number of elements than C^1, with redundant codebook vectors being eliminated. **e,f)** Now the important features for the k-th landmark of a certain object can be defined as those codebook vectors $\vec{c} \in C^2$ for which a certain percentage p of $\vec{s}^{k}_{ij}$ exists with $d(\vec{c}, \vec{s}^{k}_{ij}) < r_2$ (figure 6e,f).

Autonomous Learning: To achieve correspondence (PL2) and to avoid manual intervention the mechanism described above can be applied. Then the flexible grid can be substituted by a rectangular grid and the interaction of the camera and the motor controlled feedback ensures that landmarks are positioned at corresponding pixel position on the object (see figure 1) and the very same learning algorithm as described in section 5.2 for manually defined landmarks can be applied (see figure 1v for autonomously learned representations).

5.3 Matching

To use the learned representation for location and classification of objects Elastic Graph Matching (EGM) (Lades et al., 1993; Wiskott et al., 1997) is used. To apply EGM a similarity function between a graph labelled with the learned local line segments and a certain position in the image is defined. It simply averages *local similarities.* These local similarities express the system's confidence whether a pixel in the image represents a certain landmark. The graph is adapted in position and scale by optimising 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 each learned feature and pixel position in the image it is simply checked whether the corresponding normalised filter response (see section 5.1) is high or low, i.e., the corresponding feature is present or absent. Because of the sparseness (PF4) of the representation only a few of these checks have to be made, therefore the matching is fast. Because it is only made use of the important features, the matching is efficient.

Simulations: The test sets of hand postures contain images of 10 different hand postures in front of homogeneous background with controlled illumination (table 1, row 1–3, 240 images) and with a second set containing images with inhomogeneous background and varying illumination (row 4–6, 200 images). Matching with ten representations (one for each hand posture) takes 9.5 seconds and recognition rate was 93% (first row). The simulations corresponding to second row were performed with representations extracted by one-shot learning. The

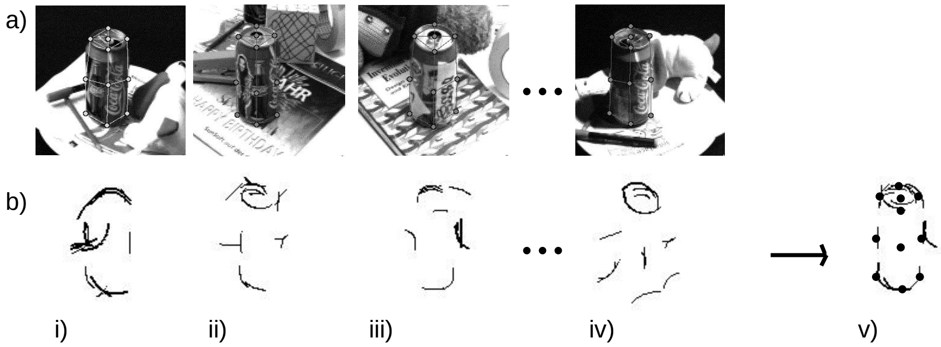

Figure 7: **a)** Pictures for training. **b,i–iv)**: Significant features per instance describing beside relevant information also accidental features such as background, shadow or surface textures. **b,v)** The learned representation.

performance is still remarkably high (80%). The performance with the bunch graph approach as described in (Triesch and von der Malsburg, 1996) is given in the third row. Results for the test set with uncontrolled background and illumination is shown in row 4–6. For the first test set performance within the bunch graph approach (Wiskott et al., 1997; Triesch and von der Malsburg, 1996) is comparable to ORASSYLL. For the second and more difficult set, performance of ORASSYLL is significantly better.

In (Loos et al., 1998) face detection with binarised banana wavelets was performed on a very large data set (more than 700 pictures) with size variation of faces between 40 and 60 pixel, inhomogeneous background and uncontrolled illumination. For this set performance was 95%. For the problem of face–finding it has been demonstrated in (Krüger, 1998b) that performance could be increased from 54% to 77% compared to the bunch graph approach on an extremely difficult test set with a significant speed up. It also could be performed successfully matching with cans, toys and other objects. Especially in case of uncontrolled illumination and inhomogeneous background significant improvement compared to (Lades et al., 1993; Wiskott et al., 1997) could be achieved. Furthermore, in (Krüger, 1998b; Krüger and Peters, 2000) ORASSYLL was compared extensively to the older system (Lades et al., 1993; Wiskott et al., 1997) as well as to a bunch of other object recognition systems.

6 Conclusion and Outlook

ORASSYLL is founded on reflections about the necessity, structure and the amount of *a priori* knowledge an artificial vision system might require. Genetically determined structures of the human visual system and findings of developmental psychology supported the definition of predetermined structural constraints within ORASSYLL. In contrast to model–free methods within ORASSYLL the input is organised within a perception–action–cycle and transformed into a highly structured feature space. Learning is not only based on trial–and–error but guided by internal statistical criteria. As a result of this controlled ap-

Matching Results for Hand Posture Classification						
	Representation		Transformation		Performance	
	nb. reps	rep	approx	sec.	sec. match	Recog.
1)	10	standard	no approx	17.0	9.5	93 %
2)	10	one instance	approx	4.9	12.4	80 %
3)	10	bunch graph		0.9	18.0	93 %
4)	10	standard	no approx	17.0	9.5	90 %
5)	10	standard	approx	4.9	9.5	80 %
6)	10	bunch graph		0.9	18.0	65 %

Table 2: Matching results for face finding and hand posture recognition (for interpretation see text).

plication of *a priori* knowledge and in contrast to approaches applying a manually designed object representation (e.g., (Yuille, 1991)), model–based representations can be extracted autonomously or with only little manual intervention. These representations were applied for difficult discrimination tasks.

An important problem remains the integration of higher stages of object representation from which much less is known compared to our knwoledge of striate cortex (for an overview see, e.g., (Hoffman, 1980)). It is possible that a key to formalise these higher levels of visual processing is the finding and formalising of appropriate *a priori* constraints. This will be a challenging task for ongoing and future research.

Acknowledgements: I would like to thank Christoph von der Malsburg, Laurenz Wiskott, Michael Pötzsch, Niklas Lüdtke, Wolfgang Banzhaf and Ladan Shams for fruitful discussion.

References

Abu-Mostafa, Y. (1995). Hints. *Neural Computation*, 7:639–671.

Alpaydin, E. and Jordan, M. (1996). Local linear perceptrons for classification. *IEEE Transactions on Neural Networks*, 7(3):788–792.

Arbib, M. (1994). *The handbook of brain theory and neural networks.* MIT–Press.

Atkinson, J. and Braddick, O. (1989). Development of basic visual functions. In Slater, A. and Bremner, G., editors, *Infant Development*, pages 7–41.

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

Bell, A. and Sejnowski, T. (1996). Edges are the 'independent components'of natural scenes. *Advances in Neural Information Processing Systems*, 9.

Blakemore, C. and Cooper, G. (1970). Development of the brain depends on the visual environment. *Nature*, 228:477–478.

Bower, T. (1971). The object in the world of the infant. *Scientific American*, 225:30–38.

Cootes, T., Taylor, C., Cooper, D., and Graham, J. (1995). Active shape models — their training and application. *Computer Vision and Image Understanding*, January:33–59.

Creutzfeld, O. (1977). Generality of the functional structure of the neocortex. *Naturwissenschaften*, 64:507–517.

Daugman, J. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by 2d visual cortical filters. *Journal of the Optical Society of America*, 2(7):1160–1169.

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

Fisher, A., editor (1923). *The Mathematical Theory of Probabilities.* Macmillan, New York.

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

Gödecke, I. and Bonhoeffer, T. (1996). Development of identical orientation maps for two eyes without common visual experience. *Nature*, 379:251–255.

Goren, C. (1975). Form perception, innate form preferences and visually mediated head–turning in the human newborn. *Paper represented at the Biennial Meeting of the Society for Research in Child Development, Denver.*

Hoffman, D., editor (1980). *Visual Intelligence: How we create what we see.* W.W. Norton and Company.

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

Jones, J. and Palmer, L. (1987). An evaluation of the two dimensional gabor filter model of simple receptive fields in striate cortex. *Journal of Neurophysiology*, 58(6):1223–1258.

Jones-Molfese, V. (1992). Responses of neonates to colored stimuli. *Child development*, 48:1092–1095.

Kant, I. (1781). *Kritik der reinen Vernunft.*

Knudsen, E., Lac, S., and Esterly, S. (1987). Computational maps in the brain. *Ann. Rev. Neuroscience.*, 10:41–65.

KOENDERINK, J. (1992). Wechsler's vision: An essay review of computational vision by Harry Wechsler. *Ecological Psychology*, 4:121—128.

KRÜGER, N. (1997). 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. (1998a). Collinearity and parallelism are statistically significant second order relations of complex cell responses. *Neural Processing Letters*, 8(2).

KRÜGER, N. (1998b). *Visual Learning with a priori Constraints.* (PhD Thesis) Shaker Verlag, Germany.

KRÜGER, N. AND PETERS, G. (2000). Orassyll: Object recognition with autonomously learned and sparse symbolic representations based on metrically organized local line detectors. *Computer Vision and Image Understanding*, 77.

LADES, M., VORBRÜGGEN, J., BUHMANN, J., LANGE, J., VON DER MALSBURG, C., WÜRTZ, R., AND KONEN, W. (1993). Distortion invariant object recognition in the dynamik link architecture. *IEEE Transactions on Computers*, 42(3):300–311.

LANITIS, A., TAYLOR, C., AND COOTES, T. (1997). Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 7:743–756.

LINDE, Y., BUZO, A., AND GRAY, R. (1980). An algorithm for vector quantizer design. *IEEE Transactions on communication*, vol. COM-28:84–95.

LOOS, H., FRITZKE, B., AND VON DER MALSBURG, C. (1998). Positionsvorhersage von bewegten objekten in groformatigen bildsequenzen. *Proceedings in Artificial Intelligence: Dynamische Perzeption*, pages 31–38.

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.

PHILLIPS, W. AND SINGER, W. (1997). In search of common foundations for cortical processing. *Behavioral and Brain Sciences*, 20(4):657–682.

PÖTZSCH, M., KRÜGER, N., AND VON DER MALSBURG, C. (1996). Improving object recognition by transforming gabor filter responses. *Network: Computation in Neural Systems*, 7:341–347.

RAUH, H. (1995). Frühe kindheit. In OERTER, R. AND MONTADA, L., editors, *Entwicklungspsychologie*, pages 167–248. Psychologie VerlagsUnion, 3 edition.

RUMELHART, D., HINTON, G., AND WILLIAMS, R. (1986). Learning representation by back–propagating errors. *Nature*, 323(9):533–536.

SOMMER, G. (1997). Algebraic aspects of designing behaviour based systems. In SOMMER, G. AND KOENDERINK, J., editors, *Algebraic Frames for the Perception and Action Cycle*, pages 1–28. Springer Verlag.

SPELKE, E. (1993). Principles of object perception. *Cognitive Science*, 14:29–56.

SUR, M., GARRAGHTY, P., AND ROE, A. (1988). Experimentally induced visual projections into auditory thalamus and cortex. *Science*, 242:1437–1441.

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

TRIESCH, J. AND VON DER MALSBURG, C. (1996). Robust classification of hand postures against complex background. *Proceedings of the Second International Workshop on Automatic Face- and Gesture recognition, Vermont*, pages 170–175.

TURK, M. AND PENTLAND, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86.

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.

YUILLE, A. L. (1991). Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, 3(1):59–70.