Learning Object Representations Using A Priori Constraints Within ORASSYLL

Norbert Krüger
Institut für Neuroinformatik, Ruhr-Universität Bochum, 44780 Bochum, Germany

In this article, a biologically plausible and efficient object recognition system (called ORASSYLL) is introduced, based on a set of a priori constraints motivated by findings of developmental psychology and neurophysiology. These constraints are concerned with the organization 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 that has to deal with a high-dimensional learning problem such as object recognition is well founded (Geman, Bienenstock, & Doursat, 1995; Abu-Mostafa, 1995). However, the definite selection and formalization 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 into focus.

Plausible requirements for predetermined structural constraints are discussed, and definitions of suitable constraints for object recognition are given. Furthermore, their formalization within an artificial system called ORASSYLL (object recognition with autonomous learned and sparse symbolic representations based on local line detectors) is discussed. I will claim that findings of developmental psychology and neurophysiology indicate that the human visual system 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
that has established predetermined structural constraints that may be extracted up to a certain degree from experimental data and 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, model-based representations within ORASSYLL can be extracted autonomously or with only little manual intervention within a perception-action cycle.

In section 2 arguments for the necessity of a priori constraints are summarized, and the relation of evolutionary learning and individual learning is discussed. Motivated by genetically determined structure in the human brain and observations of the behavior 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 organization of the input in local and corresponding entities utilizing 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 focusing on the learning of object representations and 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 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—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 that “bias needs to be designed to each particular problem.” This conclusion is the starting point of this article, in which concrete choices of bias are suggested in terms of structural constraints 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 suffers from too much bias. These systems can be called model-based systems (see, e.g., Yuille, 1991; Lanitis, Taylor, & Cootes, 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 parameters, 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, Taylor, Cooper, and Graham (1995)—resistors are local-
ized 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 generalization. This trade-off requires first that the a priori constraints should be powerful in the sense that they cover the essential structure of the problem they deal with. Second, they should be general, such that they can be applied in many situations and do not restrict the system to deal with only very specific subproblems. 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 organize visual experience. Therefore, they seem to cover essential structures of the physical world, giving us a valuable opportunity to look at results of biology to become inspired for suitable definitions of constraints. In this sense, the Kantian idea (Kant, 1781) of establishing a table of a priori constraints that organizes perception can be supported, guided, and justified by a good amount of neuropysiological and psychophysical data, discussed in the next section.

3 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 restrict myself to the aspects relevant for predetermined structural knowledge applied within ORASSYLL. In section 3.1 the observable (regarding the behavior of infants), or outer, development of the human visual system is summarized. In section 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 is neither a blank table nor a system completely determined at birth; rather, structural knowledge is already existent and guides postnatal 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 colors (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
Norbert Krüger

(Rauh, 1995). Spelke (1993) demonstrated that the Gestalt principle of “common fate” is already used by 4-month-old infants and also demonstrated the appearance of the Gestalt principles collinearity and parallelism at an age of 7 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 to 6 months, an infant can perform a visually controlled movement of arms and hands. Infants under the age of 4 months mainly characterize 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 acquisition of object representation is the use of 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 controversial question. As an extreme viewpoint Creutzfeld (1977) proposed that all cortical neurons are initially equipotent and that laminar and areal differences in the organization 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, Garraghty, & Roe, 1988; Blakemore & Cooper, 1970). However, a considerable number of counterexamples also show the restricted adaptivity of the visual system, as in the case of strabism and astigmatism, for example (Atkinson & Braddick, 1989). Furthermore, the evolution of complex structures without postnatal visual experiences, such as orientation maps (Wiesel & Hubel, 1974; Gödecke & Bonhoeffer, 1996), confirms the importance of genetic predetermination.

I have already given a short overview of the results of research about the cortogenesis of the visual system that is relevant for the choices of constraints within ORASSYLL(Krüger, 1998b). I argue there 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 color) (Wiesel & Hubel, 1974; Gödecke & Bonhoeffer, 1996), that is, the coarse sensitivity of neurons, are basically initiated before the first postnatal 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 normalization processes (parallel to the development of lateral synaptic connections) develop during the first months of visual experience. The arrangement of features in computational maps (Knudsen, Lac, & Esterly,
1987) is a major principle applied throughout the brain in the early stages of visual processing (Hubel & Wiesel, 1979) and higher stages (Tanaka, 1993) and probably, at least for area V1, is genetically determined (see Wiesel & Hubel, 1974; Gödecke & Bonhoeffer, 1996).

The utilization 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 & Singer, 1997). It is an interesting by-product of ORASSYLL that the Gestalt principles of 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 & Hubel, 1974; Gödecke & Bonhoeffer, 1996) already give good hints about predetermined structures, such as feature choices and their organization 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, which will be realized within ORASSYLL, are introduced. 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 the focus of attention and 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 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 as efficient as the human visual system. First, here the focus is on shape processing; other clues, such as color or disparity information, are ignored. Second, in its current form, ORASSYLL is a two-dimensional (2D) approach. Third, 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. Nevertheless, the constraints used within ORASSYLL are 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 organization of the input (PL1–2, section 4.1); constraints concerning feature selection, feature organization, and feature processing (PF1–4, section 4.2);
Table 1: Constraints, Analogy in the Human Visual System, and Functional Role Within ORASSYLL.

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Analogy in Visual System</th>
<th>Functional Role Within ORASSYLL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL1: Independence</td>
<td>Localized receptive fields (Hubel &amp; Wiesel, 1979)</td>
<td>Reducing complexity by splitting the input space in subspaces</td>
</tr>
<tr>
<td>PL2: Correspondence</td>
<td>Interaction of arm movement and perception (perception-action-cycle) (Rauh, 1995; Koenderink, 1992)</td>
<td>Learning with comparable entities</td>
</tr>
<tr>
<td>PF1: Feature selection</td>
<td>Genetically determined orientation-sensitive Gabor-shaped neurons (Wiesel &amp; Hubel, 1974; Jones &amp; Palmer, 1987)</td>
<td>Reduction of search space by forcing description by specific features with symbolic meaning</td>
</tr>
<tr>
<td>PF2: Feature Arrangement</td>
<td>Genetically determined order of orientation maps in V1 (Wiesel &amp; Hubel, 1974; Gödecke &amp; Bonhoeffer, 1996)</td>
<td>Combination of similar and separation of dissimilar entities by a metric</td>
</tr>
<tr>
<td>PF3: Hierarchical processing</td>
<td>Applied within whole cortex (see, e.g., Oram &amp; Perrett, 1994)</td>
<td>Sharing of resources; speed-up of feature processing</td>
</tr>
<tr>
<td>PF4: Sparse coding</td>
<td>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</td>
<td>Low memory requirements and high storage capacity; reduction of space of relations within object representations; speed-up of matching</td>
</tr>
<tr>
<td>PE1: Maximal discrimination</td>
<td>Transformation of redundancy of input data into cognitive maps (Barlow, 1961)</td>
<td>Speed-up of learning by internal evaluation of features instead of, e.g., error backpropagation</td>
</tr>
<tr>
<td>PE2: Minimal redundancy</td>
<td>Transformation of redundancy of input data into cognitive maps (Barlow, 1961)</td>
<td>Reduction of redundancy by representing similar entities by one entity utilizing metric</td>
</tr>
</tbody>
</table>
and constraints concerning statistical feature evaluation (PE1–2, section 4.3). Their relationship to biological and psychological findings and their functional role within ORASSYLL are summarized in Table 1.

4.1 Locality and the Correspondence Problem. The system’s input is spatially organized. The input is divided into nonoverlapping subparts (PL1: independence) and reflects a certain consistency of moving objects in time (PL2: correspondence). This organization of the input decreases the relevant feature space and facilitates learning.

PL (Locality).

PL1 (independence): Features at distant locations are assumed to be independent.

PL2 (correspondence): Only features corresponding to the same landmark for different examples of the same object class are used for learning.

4.1.1 PL1. The first part of the constraint PL already implies the locality of features: a local feature (i.e., a feature that 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 & Sejnowski, 1996). The splitting of the feature space in smaller independent subspaces is a design principle of many artificial systems (e.g., Alpaydin & Jordan, 1996; Wiskott, Fellous, Krüger, & von der Malsburg, 1997). Nevertheless, others (e.g., Turk & Pentland, 1991) apply global functions to the input image.

4.1.2 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:

- **Translation** of the object, which leads to the appearance of corresponding landmarks at different positions in the image and varying background and scale
- **Rotation**, which may lead to the occurrence of very dissimilar views of the same object
- **Variation of illumination**, which 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 unless at least some of these sources of variation are eliminated. The easiest way is to solve the correspondence problem manually, but 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 three-dimensional (3D) position, training data are 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 Figures 1a and 1b). The method is comparable to an arm movement and grasping controlled by vision such as infants 4 to 6 months old are able to (see section 3.1) and reflects the strong relation between action and perception (Koenderink, 1992; Sommer, 1997). Interestingly 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, in learning a suitable representation of objects.

4.2 Feature Selection and Feature Organization. The spatially organized input is transformed to a structured feature space; 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 and their transformation to a more “symbolic” meaning (PF1: feature selection), their interrelation (PF2: feature arrangement), their computation (PF3: hierarchical processing), and their role within the object representation (PF4: sparse coding). An important difference from 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 ensure 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).

PF1 (feature selection): Significant features of a localized area of the two-dimensional projection of the visual world are localized (curved) line segments.

PF2 (feature metric): A metric defines a distance between these features indicating the differences in their properties orientation, curvature, and position.

PF3 (hierarchical processing): These features are computed from simpler features in a hierarchical fashion.

PF4 (sparse coding): An object is coded as a sparse and spatially ordered arrangement of these features.

4.2.1 PF1. In ORASSYLL Gabor or banana wavelets and their symbolic analog, (curved) local line segments, are used as basic features that 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, say, all linear filters as possible features (as realized, e.g., in the scalar product of a backpropagation neuron), a restriction to a small subset is imposed. Neurophysiological experiments (Wiesel & Hubel, 1974) (see also section 3.2) show that the human visual system also makes certain kinds of feature choices before their first acts of postnatal visual experience.

Within the framework of ORASSYLL it has been shown (Krüger, 1998a) that the Gestalt principles of collinearity and parallelism can be detected and described as second-order statistics of normalized Gabor wavelet responses. This nonlinear normalization was initially developed to solve a certain sub-
problem within ORASSYLL: the comparison of a symbolic feature with a certain position within the gray-level image. The normalization transforms Gabor wavelet responses such that they express the system’s confidence for the presence or absence of a local line segment—that is, it represents their interpretation in terms of constraint PF1. Surprisingly, this transformation affects the second-order statistics of Gabor wavelet responses significantly (see Figures 3a–d). Looking only at the nonnormalized Gabor wavelet responses, the Gestalt principles of collinearity and parallelism are barely detectable (see Figures 3e–f).

4.2.2 PF2. In ORASSYLL, the ordered arrangement of features is achieved by a metric defining a distance between features indicating their differences in the property orientation, curvature, and position. This metric organization is essential for learning in the object recognition system, because it allows the clustering of similar features and thus the determination of representatives for such clusters.

4.2.3 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.

4.2.4 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. ORAS-
SYLL differs in this aspect from many other object recognition systems that apply compact representations (e.g., Turk & 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 relations 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) that define the feature space itself, two criteria are now 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, say, backpropagation (Rumelhart, Hinton, & Williams, 1986), in which global criteria are optimized, the two constraints represent criteria that guide and speed up the learning algorithm by evaluating intermediate stages of processing (see also the discussion of the credit assignment problem in, e.g., Arbib, 1994).

PE (Evaluation).

PE1 (maximal discrimination): Features are preferred whose values on images vary little within classes and vary much between classes.

PE2 (minimal redundancy): Redundant information shall be eliminated from the system.

4.3.1 PE1 and 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 give a technical description of ORASSYLL, based on the a priori constraints introduced in section 4. My aim is not to give a detailed 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 & Peters, 2000).

5.1 The Feature Space. In this section the realization of the constraints PF1–PF4 is described. By utilizing these constraints, the input is transformed into a metrically organized feature space. Images are interpreted as an assembly of spatially organized local line segments that are processed hierar-
5.1.1 Gabor and Banana Wavelets. A banana wavelet $\mathcal{B}^b$ is a complex-valued function, parameterized by a vector $\mathbf{b}$ of four variables $\mathbf{b} = (f, \alpha, c, s)$ expressing the attributes frequency ($f$), orientation ($\alpha$), curvature ($c$), and elongation ($s$).\footnote{For Gabor wavelets $\mathbf{b}$ is reduced to $(f, \alpha)$} It can be understood as a product of a curved and rotated complex wave function $F^b$ and a stretched two-dimensional Gaussian $G^b$, bent and rotated according to $F^b$ (see Figure 2a).

Banana wavelets are generalized Gabor wavelets (for Gabor wavelets, see, e.g., Daugman, 1985) that possess, in addition to frequency and orientation, the parameters of curvature and elongation. The approach introduced here does not necessitate the use of banana wavelets and is also applicable with Gabor wavelets (see Figure 5d.i,ii,iv, for object representations with only straight line segments).

5.1.2 The Feature Space. The six-dimensional space of vectors $\mathbf{c} = (\mathbf{x}, \mathbf{b})$ is called the feature space, with $\mathbf{c}$ representing the wavelet $\mathcal{B}^b$ with its center at pixel position $\mathbf{x}$ in an image. PF2 is realized by a metric $d(\mathbf{c}_1, \mathbf{c}_2)$. Two coordinates $\mathbf{c}_1, \mathbf{c}_2$ are expected to have a small distance $d$ when their corresponding kernels are similar—that is, they represent similar features. For the exact definition of the metric, first a distance measure is defined for the orientation-curvature subspace $(\alpha, c)$ expressing the Moebius topology thereof, setting

$$d((\alpha_1, c_1), (\alpha_2, c_2)) =$$

$$\min \left\{ \sqrt{\frac{(\alpha_1 - \alpha_2)^2}{\epsilon_{\alpha}^2} + \frac{(c_1 - c_2)^2}{\epsilon_c^2}}, \sqrt{\frac{(\alpha_1 - \pi)^2}{\epsilon_{\alpha}^2} + \frac{(c_1 + c_2)^2}{\epsilon_c^2}}, \sqrt{\frac{(\alpha_1 + \pi)^2}{\epsilon_{\alpha}^2} + \frac{(c_1 + c_2)^2}{\epsilon_c^2}} \right\}$$

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

$$d(\mathbf{c}_1, \mathbf{c}_2) =$$

$$\sqrt{\frac{(x_1 - x_2)^2}{\epsilon_x^2} + \frac{(y_1 - y_2)^2}{\epsilon_y^2} + \frac{(f_1 - f_2)^2}{\epsilon_f^2}} + d((\alpha_1, c_1), (\alpha_2, c_2))^2 + \frac{(s_1 - s_2)^2}{\epsilon_s^2}$$

(5.1)

The values $\mathbf{\epsilon} = (\epsilon_x, \epsilon_y, \epsilon_f, \epsilon_c, \epsilon_s)$ define a cube of volume 1 in the features space, that is, they define the weights for the different properties, such as orientation and curvature. In the standard settings, $\mathbf{\epsilon} = (4, 4, 0.01, 0.3, 0.4, 3.0)$ is used.
5.1.3 Nonlinear Transformations of the Filter Responses. The feature processing of ORASSYLL consists of a two-step nonlinear transformation of the complex filter responses. In the first step, the magnitude of the filter response is extracted after the convolution of $B^b$ with the image $I$. In contrast to the complex filter response oscillating with phase, the magnitude of the response is stabler under a slight variation of position (see Pötzsch, Krüger, & von der Malsburg, 1996). A filter $B^b$ causes a strong response at a certain pixel position when the local structure of the image at that position is similar to $B^b$.

The magnitude of the filter responses depends significantly on the strength of edges in the image. However, here I am interested only in the presence, and not the strength, of edges. Thus, in a second step, a function $N$ normalizes the real valued filter responses $r(\tilde{c})$ into the interval $[0, 1]$. The value $N(r(\tilde{c}))$ represents the likelihood of the presence or absence of a local line segment corresponding to $\tilde{c} = (\tilde{x}_0, \tilde{b})$. This normalization is based on the “above-average criterion”:

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

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

5.1.4 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. 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 that 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 the interaction of perception and action.
Figure 3: Cross-correlation of pairs of (a–d) normalized Gabor wavelet responses and (e–h) unmodified Gabor wavelet responses 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 as in $a$), and the $z$-axis represents the correlation. In $a$, parallelism and collinearity are clearly visible. Collinearity is detectable as a ridge in the first diagram, and parallelism appears as a global property expressed in the flat part of the surface in the first diagram clearly above the surfaces corresponding to nonparallel orientations.

$$+ = \beta^r_1 \cdot \underline{\sigma} + \beta^r_2 \cdot \underline{\rho} + \beta^r_3 \cdot \underline{\sigma}$$

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

5.2.1 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: it has to cause a strong response (C1) and it has to represent a maximum within a local area of the feature space (C2). Figures 1c.i–iv, 5b, 5d, and 6b.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. (a, c) The objects to be learned in front of a homogeneous background. (b, d) 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 that indicates the general presence of a feature, whereas C2 represents the inter-columnar competition giving a more specific coding of this feature (Oram & 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.

5.2.2 One-Shot Learning. By positioning a rectangular grid on an object (see Figure 5a.i) in front of a homogeneous background and extracting significant features per instance, suitable representations of objects can be extracted. These representations have already been 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).

5.2.3 Clustering. In case of a nonhomogeneous background and uncontrolled illumination, one-shot learning would create representations with line segments corresponding to background (see Figure 1c.i–iv). In this case,
Figure 6: Clustering. (a) Distribution of the significant features per instance extracted at a certain landmark. (b) Code book initialization. (c) Code book vectors after learning. (d) Substituting sets of code book vectors with small distance by their center of gravity. (e) Counting the number of elements within a certain radius. (f) Deleting code book vectors representing insignificant features.

A more sophisticated learning scheme has to be applied. After extracting the significant features per instance in different pictures, an algorithm extracts 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 Figures 6b.i–iv and 1c.i–iv). A significant feature should be independent of background, illumination, or accidental qualities of a certain example of the object class; in other words, it should be invariant under these transformations (PE1). This is realized

Table 2: Matching Results for Hand Posture Recognition.

<table>
<thead>
<tr>
<th>Number</th>
<th>Representation</th>
<th>Transformation</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Representations</td>
<td>Approximation</td>
<td>Seconds</td>
</tr>
<tr>
<td></td>
<td>Representations</td>
<td>Seconds</td>
<td>Matching</td>
</tr>
<tr>
<td>1)</td>
<td>10 Standard</td>
<td>No approx</td>
<td>17.0</td>
</tr>
<tr>
<td>2)</td>
<td>10 One instance</td>
<td>Approx</td>
<td>4.9</td>
</tr>
<tr>
<td>3)</td>
<td>10 Bunch graph</td>
<td>Approx</td>
<td>4.9</td>
</tr>
<tr>
<td>4)</td>
<td>10 Standard</td>
<td>No approx</td>
<td>17.0</td>
</tr>
<tr>
<td>5)</td>
<td>10 Standard</td>
<td>Approx</td>
<td>4.9</td>
</tr>
<tr>
<td>6)</td>
<td>10 Bunch graph</td>
<td>Approx</td>
<td>0.9</td>
</tr>
</tbody>
</table>
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 6). The correspondence problem (PL2) is assumed to be solved; it is assumed that the position of certain landmarks of an object is known on pictures of different examples of these objects. In Figure 6, 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 7). Let $I$ be a set of pictures of different examples of a class of objects of certain orientation and approximately equal size (see Figure 6a). $I_{j,k}$ represents a local area in the $j$th image in $I$ with the $k$th landmark as its center. Let $E_{sk_{ij}}$ be the $i$th significant feature per instance extracted in the area $I_{j,k}$. All $E_{sk_{ij}}$ for a specific $k$ are collected in one set $S_k$. Then the LBG vector quantisation algorithm (Linde, Buzo, & Gray, 1980) is applied to $S_k$ (see Figure 6b). After vector quantization, a code book $C_1$ expresses the vectors $E_{sk_{ij}}$ with a certain number $n_{C_1}$ of code book vectors $E_{c_1i} \in C_1 \subset C$, $i = 1, \ldots, n_{C_1}$ (see Figure 7b).

The LBG algorithm reduces the distortion error—the average error occurring when all elements of $S_k$ are replaced by the nearest code book vector in $C_1$. In case of high densities of elements $E_{sk_{ij}}$ 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); they are expected to have large distances in the feature space. Therefore a smaller code book $C_2$ is constructed in which the $\vec{c}, \vec{c}' \in C_1$ with close distances are replaced by their center of gravity. Let $r_{1} \in R^{+}$ be fixed. For all $\vec{c} \in C_1$, the number of
\( \bar{c} \in C^1 \) with distance \( d(\bar{c}, \bar{c}') < r_1 \) (see Figure 7c) is computed. If there exists at least one such \( \bar{c}' \neq \bar{c} \), all the code book vectors in \( C^1 \) with \( d(\bar{c}, \bar{c}') < r_1 \) are substituted by their center of gravity (see Figure 7d). \( C^2 \) now represents a code book with fewer than or an equal number of elements as \( C^1 \), with redundant code book vectors being eliminated. Now the important features for the \( k \)th landmark of a certain object can be defined as those code book vectors \( \bar{c} \in C^2 \) for which a certain percentage \( p \) of \( \bar{s}_{ij}^k \) exists with \( d(\bar{c}, \bar{s}_{ij}^k) < r_2 \) (see Figures 7e and 7f).

5.2.4 Autonomous Learning. To achieve correspondence (PL2) and 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). Then the same learning algorithm as described in section 5.2 for manually defined landmarks can be applied (see Figure 1.v 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 labeled 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 optimizing the total similarity. The graph with the highest similarity determines the size and position of the objects within the image.

In a nutshell, the local similarity is defined as follows: For each learned feature and pixel position in the image, it is simply checked whether the corresponding normalized filter response (see section 5.1) is high or low—that is, whether 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 made use of only the important features, the matching is efficient.

5.3.1 Simulations. The test sets of hand postures contain images of 10 different hand postures in front of a homogeneous background with controlled illumination (see Table 1, rows 1–3, 240 images) and with a second set containing images with an inhomogeneous background and varying illumination (rows 4–6, 200 images). Matching with 10 representations (one for each hand posture) takes 9.5 seconds; the recognition rate was 93% (first row). The simulations corresponding to the second row were performed with representations extracted by one-shot learning. The performance is still remarkably high (80%). The performance with the bunch graph ap-
proach, 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 are shown in rows 4–6. For the first test set performance within the bunch graph approach (Wiskott et al., 1997; Triesch & von der Malsburg, 1996) is comparable to ORASSYLL. For the second and more difficult set, performance of ORASSYLL is significantly better.

Loos, Fritzke, and von der Malsburg (1998) performed face detection with binarized banana wavelets was performed on a very large data set (more than 700 pictures) with size variation of faces between 40 and 60 pixels, an inhomogeneous background, and uncontrolled illumination. For this set, performance was 95%. For the problem of face finding it has been demonstrated (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 an inhomogeneous background, significant improvement compared to Lades et al. (1993) and Wiskott et al. (1997) could be achieved. Furthermore, in Krüger (1998b) and 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 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 organized within a perception-action cycle and transformed into a highly structured feature space. Learning is not only based on trial and error but is guided by internal statistical criteria. As a result of this controlled application 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 a 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 knowledge of striate cortex (for an overview see, e.g., Hoffman, 1980). It is possible that a key to formalizing these higher levels of visual processing is finding and formalizing appropriate a priori constraints. This will be a challenging task for ongoing and future research.
Acknowledgments

I thank Christoph von der Malsburg, Laurenz Wiskott, Michael Pötzsch, Niklas Lüdtke, Wolfgang Banzhaf, and Ladan Shams, Peter Hancock, and Helge Ritter for fruitful discussion. This work was supported by grants from the German Ministry for Science and Technology 01IN504E9 (NEU-ROS), 01M3021A4 (Electronic Eye), and the EC Human Capital and Mobility Program (ERBCHRX CT930097).

References


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


Received June 25, 1998; accepted March 23, 2000.