

Peter König · Norbert Krüger

Symbols as self-emergent entities in an optimization process of feature extraction and predictions

Received: 10 July 2005 / Accepted: 10 January 2006 / Published online: 23 February 2006
© Springer-Verlag 2006

Abstract In the mammalian cortex the early sensory processing can be characterized as feature extraction resulting in local and analogue low-level representations. As a direct consequence, these map directly to the environment, but interpretation under natural conditions is ambiguous. In contrast, high-level representations for cognitive processing, e.g. language, require symbolic representations characterized by expression and syntax. The representations are binary, structured and disambiguated. However, do these fundamental functional distinctions translate into a fundamental distinction of the respective brain areas and their anatomical and physiological properties? Here we argue that the distinction between early sensory processing and higher cognitive functions may not be based on structural differences of cortical areas; instead similar learning principles acting on input signals with different statistics give rise to the observed variations of function. Firstly, we give an account of present research describing neuronal properties at early stages of sensory systems as a consequence of an optimization process over the set of natural stimuli. Secondly, addressing a stage following early visual processing we suggest to extend the unsupervised learning scheme by including predictive processes. These contain the widely used objective of temporal coherence as a special case and are a powerful approach to resolve ambiguities. Furthermore, in combination with a prior on the bandwidth of information exchange between units it leads to a condensation of information. Thirdly, as a crucial step, not only are predictive units optimized, but

the selectivity of the feature extractors are adapted to allow optimal predictability. Thus, over and beyond making useful predictions, we propose that the predictability of a stimulus be in itself a selection criterion for further processing. In a hierarchical system the combined optimization process leads to entities that represent condensed pieces of knowledge and that are not analogue anymore. Instead, these entities work as arguments in a framework of transformations that realize predictions. Thus, the criteria of predictability and condensation in an optimization of sensory representations relate directly to the two defining properties of symbols of expression and syntax. In this paper, we sketch an unsupervised learning process that gradually transforms analogue local representations into discrete binary representations by means of four hypotheses. We propose that in this optimization process acting in a hierarchical system, entities emerge at higher levels that fulfil the criteria defining symbols, instantiating qualitatively different representations at similarly structured low and high levels.

1 Introduction

In recent years we saw a rapid increase of our knowledge about sensory processing in the mammalian brain under natural conditions (c.f. Kayser et al. 2004). Starting from neurophysiological investigations of early visual processing a large part of this work focuses on a quantitative description of signal processing and characterizing the properties of representations of stimuli (Maunsell and Newsome 1987; Ringach 2004). For example, the activation of simple cells in primary visual cortex is well described by a convolution of the stimulus with a kernel defining its linear receptive field (Jones and Palmer 1987). Besides orientation, subsets of neurons in V1 are sensitive to different visual features such as optic flow, colour and disparity (Hubel and Wiesel 1959, 1962). In the resulting representation no structured interaction between constituents exists and the activation of neurons occurs in a graded fashion. The information represented in the different visual features is incomplete and

P. König (✉)
Department of Neurobiopsychology,
Institute of Cognitive Science, University Osnabrück,
Albrechtstr. 28, 49076 Osnabrück, Germany,
E-mail: pkoenig@uos.de
Tel.: +49-541-9692399
Fax: +49-541-9692596

N. Krüger
Cognitive Vision Group,
Institut for Medieteknologi og Ingeniørvidenskab,
Aalborg University Copenhagen,
Ballerup, Denmark

ambiguous since it is extracted by means of *local* filter operations (Aloimonos and Shulman 1989; Krüger and Wörgötter 2004). A prominent example is the aperture problem in optic flow computation. Another example is the extraction of Three-dimensional (3D), structures by stereoscopic images. Here the loss of information in the mapping from 3D world to the 2D retina necessitates an interpretation of the representations in the light of additional constraints (e.g. Klette et al. 1998). Summarizing, the processing in low-level sensory systems can be characterized as feature extraction and the resulting representations are local, analogue, ambiguous and map directly to the environment.

More recently, an additional path has been approached to understand early visual processing. It is based on unsupervised learning of neuronal representations of natural stimuli. A milestone was the discovery that orientation selective responses of simple cells as found in primary visual cortex can be understood as optimally sparse representations of natural images (Olshausen and Field 1996, 2004; Hyvärinen and Hoyer 2000). Further progress has been made addressing selectivity over features and in other visual modalities (Hurri and Hyvärinen 2003; Einhäuser et al. 2003; Hafner et al. 2004; Berkes and Wiskott 2005, S. Onat et al., submitted). Although there are still numerous unsolved problems (see Sect. 7 below), current progress fosters optimism that a substantial part of feature selectivity in the primary sensory areas are the result of unsupervised learning applied to natural images making use of a small number of objectives.

Currently, an investigation of the physiological basis of higher cognitive processes becomes feasible and moves into the centre of interest. In these processes, the use of categories and symbols plays a prominent role. Recent investigations provide evidence for category specific pop-out effects (Hershler and Hochstein 2005), effects of categories in perceptual learning (Ashby and Maddox 2005) and category-specific visual responses of single neurons in the human cortex (Kreiman et al. 2000; Quiroga et al. 2005). The most prominent example, however, is the human mastery of language. The use of language allows the representation and manipulation of symbols. The condensation of individual stimuli in categories implies a huge loss of information. However, in contrast to low-level neuronal representations, discarded information relates in a large part to irrelevant details, and ambiguities that have to be interpreted are much less prominent. A standard notion of symbols in a certain representational framework (e.g. Honavar and Uhr 1994) is that

Expression: Symbols are condensed and discrete semantic representatives are for certain pieces of knowledge.

Syntax: On which operations can be performed and which correspond to relevant functional relations in this framework.

Symbol manipulation typically identifies symbolic expressions, decomposes given expressions and generates new ones by syntax. The syntax specifies which combinations

of symbols are valid expressions, and structurally different assemblies of symbols may have different meanings. Hence, representations for cognitive processing such as high-level vision and language seem to require symbolic representations that are binary, disambiguated and show a certain degree of structure.

A fundamental problem connected to symbols is the origin of their meaning. In formal systems its relations to other symbols and operators define the meaning of a symbol (Hilbert 1928). In a purely perceptual system the meaning of symbols may come from the structural properties of the environment as well as the body and purposes of the system itself. This issue has become known as the so-called symbol-grounding problem (Harnard 1990). It has been argued that symbols are interpreted correctly only by a perception-action cycle (Steels 2003). This deep philosophical issue has triggered many debates and no satisfactory explanation of the symbol-grounding problem has been reached yet.

The two antipodes, processing of local, analogue and potentially ambiguous signals versus manipulation of discrete and structured symbols, have lead to two major research directions, found to a large extend in Neural Networks research and classical Artificial Intelligence respectively. Within the respective domains both approaches have reasonable success. However, does the fundamental distinction drawn between these disciplines translate into a fundamental distinction of the respective brain areas and physiological properties? In the following we discuss three viewpoints. They are deliberately chosen to, represent extreme answers to chart out clearly the space of possible solutions to this problem.

Firstly, cortical areas involved in local analogue signal processing and cortical areas involved in the manipulation of symbols could exhibit genetically determined different anatomical structures and circuits serving the qualitatively different functions. Indeed, it is well known for many years that the laminar structure of cortex shows regional variations and can be used to define cortical areas (Brodmann 1906), and only recently has it been speculated that few genes may serve as the foundation of high level skills like language (Vargha-Khadem et al. 2005). Nevertheless, several arguments discourage jumping to conclusions. With the exception of primary sensory and motor areas, these variations in laminar structure are small and show different patterns in different individuals (Braitenberg and Schüz 1991). Furthermore, anatomical studies reveal that the functional micro-circuits may be similar across different areas (Douglas and Martin 2004). Finally, in higher areas these variations do not match functionally defined areas. Hence, even after many decades of research, the functional significance of structural variations in cortical laminae is little understood and is far from an explanation of the variety of functions of cortical areas.

Secondly, mechanisms of symbol processing could emerge on a higher level of description of neural dynamics. This approach has some similarities with a theory proposed many years back by Lashley (c.f. Orbach 1998). It implies that

the structure of cortical circuits is generic and potentially serves any function (equipotentiality), and substantial parts of the cortex are involved in any task (mass action). Although this theory has attractive features (for a discussion see, e.g., Phillips and Singer 1997), the original interpretation of Lashley is highly controversial and not obvious to reconcile with the growing evidence of functional specialization in the human cortex (Grill-Spector and Malach 2004).

Thirdly, the distinction between early sensory processing and higher cognitive functions may not be based on structural differences of cortical areas; instead, similar learning principles acting on input signals with different statistics give rise to the observed variations of function. Quantitative differences in the form of time constants, convergence and divergence of projections as well as span of tangential connections can further shape response properties, but would not be essential as such. This would also imply that the approach, which is successful in the investigation of early sensory areas, could be applied to higher levels (R. Wyss et al., submitted).

In this work, we further investigate the third point of view. Addressing a stage following early visual processing, the unsupervised learning scheme is extended to include predictive processes. Our argumentation leads to four increasingly speculative hypotheses that become outlined in the following sections. As a central result it is claimed that in this process, entities emerge that fulfil the two criteria defining symbols.

2 Learning of feature maps from natural scenes

Following the flow of information, we first study sensory processing of stimuli in early areas in the visual system. In the primary visual cortex, most neurons can be classified into one of two generic cell types. The simple cells respond selectively to bars and gratings presented at a specific position, orientation, spatial frequency, and contrast polarity (Hubel and Wiesel 1962; Schiller et al. 1976). The neurons of the other type, complex cells, also respond to bars or gratings of adequate orientation and spatial frequency. They, however, respond equally well regardless of the contrast polarity of the stimulus and its precise location within the region of the receptive field (Hubel and Wiesel 1962; Kjaer et al. 1997).

This work follows an early proposal, that the properties of neurons in sensory systems should be specifically adapted to the behaviour of the animal (Barlow 1961). In the frog retina, for example, individual ganglion cells are perfectly suited to detect prey in the form of flies (Lettingvin et al. 1959) which the frog most certainly likes to catch. The association of features and behaviour in more developed species is less direct. Recent work links the properties of simple cells in primary visual cortex to the statistics of the natural environment. Optimally sparse representations of natural stimuli lead to orientation selective receptive fields with spatial properties matching those of simple cells (Olshausen and Field 1996). Please note that according to this concept the receptive fields of neurons may adapt without explicit supervision or a direct reinforcement signal (T. Kulvicius et al., submit-

ted). The objective functions code aspects that only allow for indirect arguments for the relevance for behaviour, e.g. reducing energy consumption. A similar argument can be made for optimally stable representations matching characteristic properties of complex neurons (Körding et al. 2004; Berkes and Wiskott 2005). Here also the relevance for behaviour is indirect, as important aspects of visual stimuli are supposed to change slower than irrelevant detail. When applied to a whole group of neurons, both types of objective functions require a decorrelation of response properties of individual neurons. On first sight this might seem a trivial step avoiding redundant representations. But indeed, it is at the heart of current state-of-the-art deconvolution techniques as described in Chichocki and Amari (2002) and serves an important purpose here as well (Hyvarinen et al. 2003; Hurri and Hyvärinen 2003). Further theoretical results also selectivity address with respect to other optimization criteria and visual features, such as motion (Berkes and Wiskott 2005), disparity (S. Onat et al., submitted) and colour (Einhäuser et al. 2003). In this way, local filter operations similar to those known from neurophysiological investigations are based on the structural properties of natural scenes and governed by a number of principles such as sparseness, stability and independence. This allows a substantial part of feature selectivity in early visual processing to be, understood on the basis of unsupervised learning according to a small number of objectives. Realistically, no current theoretical model can describe selectivity for all these features of the classical receptive field of simple and complex cells simultaneously in a unified model, let alone complex contextual effects. Even worse, a detailed analysis of the receptive field substructure reveals systematic deviations from the present theoretical predictions: e.g. simple cells are found to have few subfields and often respond well to low spatial frequency stimuli (Ringach et al. 2002). Hence, much more remains to be explored even in primary visual cortex (Olshausen and Field 2005).

If unsupervised learning is successful, how far does it take us? Whereas many results pertaining to primary sensory processing are available, few studies address unsupervised learning in hierarchical systems. Those, however, indicate that at subsequent levels representations of more complex features emerge (Wiskott and Sejnowski 2002; R. Wyss et al., submitted; M. Franzius et al., submitted). Hence, one may speculate that the approach described above and adopted by a number of research labs generalizes to higher levels of sensory processing.

Hypothesis 1: Properties of sensory representations at different levels of a processing hierarchy can be understood on the basis of optimization of objective functions, such as sparseness and temporal coherence.

However, are the differences in the statistics of natural visual stimuli that different species experience sufficient to explain a different organization of the visual hierarchy? Obviously, different behavioural needs have to be incorporated into the architecture of the sensory system (Gibson

1979). The argument has been put forward that optimally stable representations favour relevant stimuli (Wiskott and Sejnowski 2002; Körding et al. 2004). It is based on the view that irrelevant aspects, i.e. noise, are uncorrelated in space and time and therefore changing on a fast timescale. However, the reverse conclusion that all relevant stimuli do not change fast does not hold. Thus, although the approach in general may hold in a complete hierarchical system, we have to reconsider which objective functions are most useful. In Sect. 5, we describe possible extensions of currently used objective functions.

3 Predictive mechanisms for disambiguation

The problem of extracting relevant features also shows up in the form of resolving ambiguities. In the first stages of the visual system, stimuli are analyzed locally. An important example is the computation of depth information in stereo images. The relative shift of corresponding regions in the 2D image (disparity) is an indicator of depth in the 3D environment (e.g. Klette et al. 1998). However, multiple candidate patches might occur that are similar to a patch in the other image. In homogeneous areas this problem crops up in an extreme form. The local signal is essentially constant, and all matches in the homogeneous area are possible. The converse, that no match exists, may arise due to occlusion of a region in only one of the two images. Therefore no unique solution exists of the correspondence problem, and the process to reconstruct 3D information from stereo images is ambiguous. A related problem arises in matching image regions in time to determine motion vectors. Here as well local information is usually not sufficient to resolve the correspondence problem (e.g. Ullman 1979). The human visual system can make use of contextual information to derive reliable scene descriptions (e.g. Aloimonos and Shulman 1989). An elegant approach is the use of motion information to resolve ambiguities of stereo images. A stereo match directly generates a hypothesis on the 3D structure. When the motion information is taken into account, the match in the subsequent frame set can be predicted. An invalid match results in an erroneous prediction that can be quickly invalidated (see Box 1).

Hypothesis 2: Predictions across visual events are a powerful approach to resolve ambiguities.

In this line of thought, we follow the intuition that predictions relate to different points in time. But this is not a necessary restriction. Instead, a similar argument can be made for spatial relations. Here the statistical properties of typical, in our case natural, visual stimuli determine the conditional probability of local properties in other regions. This can be seen as a spatial prediction. This concept is in close analogy to the well-known Gestalt laws that are reflected in the connective structure of V1 (Gilbert and Wiesel 1989; Watt and Phillips 2000), as well as the statistics of natural images (Krüger 1998; Geisler et al. 2001; Elder and Goldberg

2002; Betsch et al. 2004). Hence, the concept of predictions to resolve ambiguities can be generalized and applied in widely varying contexts (e.g. Krüger and Wörgötter 2004).

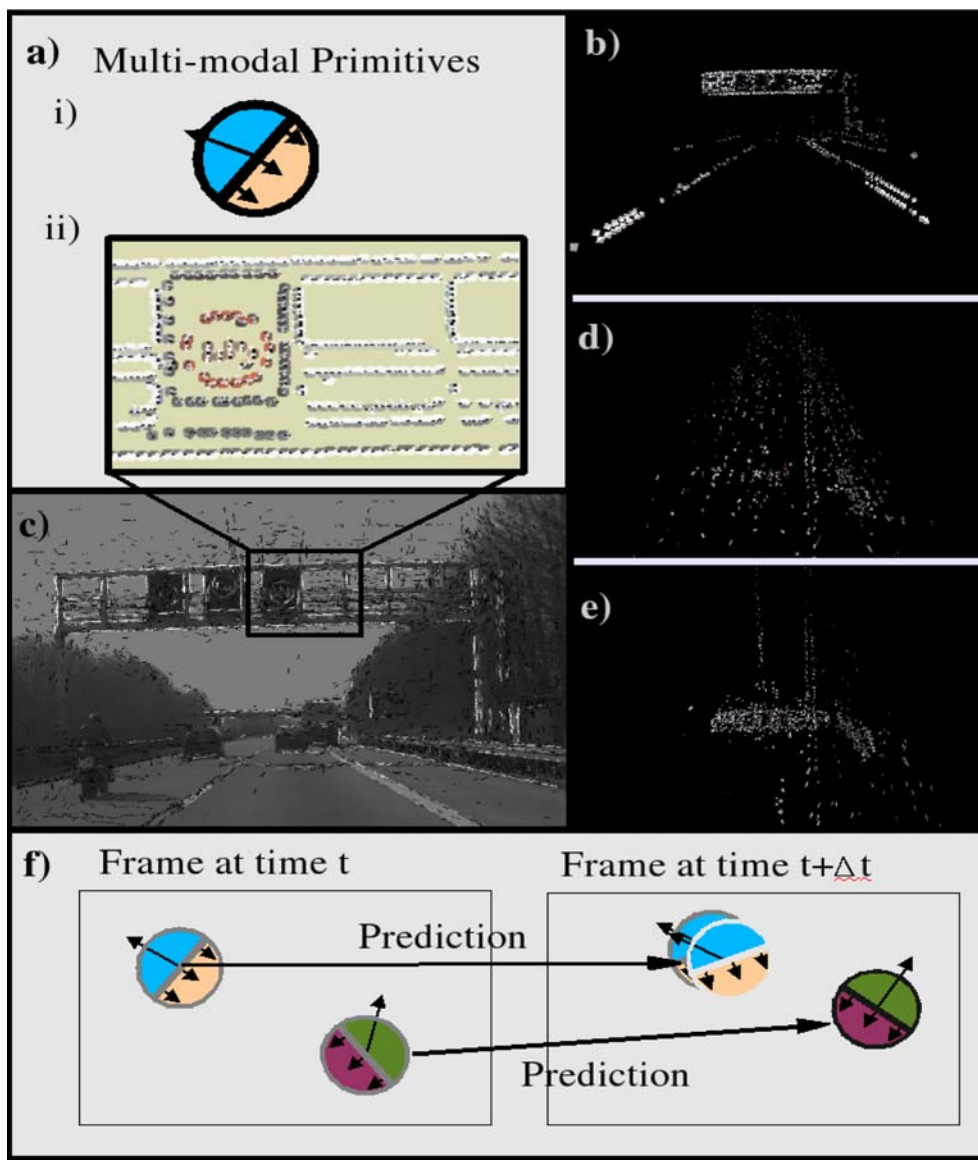
4 Predictions and objective functions

What is the relation of predictions to the objective function approach described above? In our discussion of the early visual system, sparseness and stability, are prominent examples. We discuss the relation of predictive mechanisms to both of these.

An intuitive approach is that behaviourally relevant stimuli allow predictions about future stimuli. Taking again the example of the frog's retina, spotting a small flying insect supposedly allows the frog to predict its position a short time later (Lettvin et al. 1959). Otherwise the frog would not know where to throw its tongue in order to catch the insect. Of course, these predictions will not be perfect and the insect may escape by taking a sudden turn. However, if spotting a flying insect does not increase the probability that it will be in any spatial region a short time later when the frog reacts, the visual input might as well be ignored. Indeed, it has been recently put forward, that “nonpredictive information is useless to the organism” (Tishby et al. 1999; Bialek et al. 2001). Bialek, Nemenman and Tishby present a deep and careful analysis of the relation of predictive information to the complexity and learning theory. Applied to the problem at hand, from this point of view, optimally stable representations seem to be a crude 0-th order approximation for optimally predictable representations; if something does not change, it is trivially predictable. In terms of objective functions replacing stability, predictability itself can become an additional criterion for the usefulness of a feature:

Predictability: A good feature gives rise to the prediction of other temporally and/or spatially distinct features and needs to be predictable from these.

With the inclusion of predictive processes in the unsupervised learning we are faced with the problem that such predictions code relational events. As a consequence these predictions work in a higher dimensional space than the original filter responses. A full sampling of this relational space becomes computationally intractable. Therefore we suggest that to make efficient use of predictions, we also need a change of data format. The information coded in a set of filter responses in a local part of the visual field needs to be condensed such that the space of possible predictions becomes manageable. In the technical system introduced in (Krüger et al. 2004; Krüger and Wörgötter 2005) the condensation of local visual information into a semi-symbolic format (see Fig. 1a) is a prerequisite for the utilization of predictive mechanisms for disambiguation of the locally extracted information provided by the early visual processing. Extensions of already established principles in unsupervised learning schemes such as sparseness (Field 1987) may lead to such a property. This leads to an additional criterion:



Box 1 Disambiguating stereo images by predictions based on motion information. **a** (i) Schema of a multi-modal primitive. It represents a local image patch in terms of a low-dimensional descriptor containing information about the local orientation (indicated by the *diagonal black line*), phase (indicated by the *long arrow*), colour (indicated by the contrast of *blue and yellowish region*), local motion (indicated by the *three parallel arrows*), disparity and a descriptor of the ‘homogeneity’, ‘edge-ness’, or ‘junction-ness’ of the local patch. (ii) Representation of a part of a real world stereo image (shown in *c*) in terms of multi-modal primitives. Comparing the number of bits used to represent an image by primitives versus the original image yields a condensation rate of more than 95%. For details see (Krüger and Wörgötter 2004). **b** Front view of the disambiguated scene. A good 3D representation even in difficult image regions can be observed. **c** Left frame of a stereo-image pair showing a typical highway scene. Representation of part of the image by local multi-modal primitives is shown in the *inset*. Superimposed small *oriented bars* denote predictions for this frame based on preceding frames. Many predictions are obviously erroneous (shown in *black*), some are correct and match with the present sensory evidence (shown in *white*). **d** Top view of the scene shown in (*c*) with 3D candidate set for correspondences in left and right frames extracted by standard stereo matching, i.e. not using information obtain from preceding frames. Increasing constraints of 3D candidates reduces the number of invalid 3D correspondences, but simultaneously many valid correspondences are discarded as well (data not shown). **e** Utilizing the information on ego motion we convert each candidate 3D match to a prediction for the subsequent frame. After a few iterations a more complete representation with only few outliers is generated, visible in the top view of the disambiguated 3D representation. **f** Accumulation of evidence based on spatial-temporal predictions. Based on the sensory information available at time t and knowledge of the ego-motion a prediction of the local primitives at time $t + \Delta t$ is computed. When a prediction matches the subsequently sampled data, the confidence of a candidate match as well as the associated 3D interpretation is increased, otherwise decreased. Here and in the other panels the confidences are represented by the brightness of the surrounding *circle* and the *orientation bar*. Candidate matches below a certain confidence level are discarded

Condensation: Since predictive mechanisms work in a higher dimensional relational space for an efficient coding, the local information has to be condensed.

Taking both criteria of prediction and condensation into account we can formulate our third hypothesis:

Hypothesis 3: Cortical Processing of sensorial information can be explained by a mutual optimization of condensation and predictions.

Hence, the predictive mechanisms fit seamlessly into the concept of objective functions. They supersede the temporal coherence objective and necessarily interact with a prior sparseness.

5 A concrete approach

The essence of learning feature selectivity as described in previous work rests on the definition of an objective function $\mathbf{E}^I(\mathbf{F})$. It is dependent on a set of local filter operations $\mathbf{F} = \{f_i | i : 1, \dots, n\}$ that are applied to a set of natural stimuli \mathbf{I} , mapping each stimulus \mathbf{I} onto a real value: $f_i(I) = r_i$. The result of the filter operation can be viewed as the activity of neurons.¹ As a next step the objective function is optimized by a method of choice, e.g. gradient descent. In case different aspects have to be optimized simultaneously, $\mathbf{E}^I(\mathbf{F})$ is composed as a sum of a number of terms, addressing the individual aspects. To achieve condensation of information a sparseness/decorrelation principle possibly connected with constraints on the connectivity structure of the neural net is a good start.

$$\mathbf{E}^I(\mathbf{F}) = \Psi_{\text{decorr}} + \Psi_{\text{sparse}}.$$

The relative weight of the two terms depends on the precise formulae. For a combination of a stability and decorrelation (Hipp et al. 2005) report that this is not a crucial issue and the results of the optimization process vary little within one order of magnitude for the relative weight.

$$\Psi_{\text{Decorr}} = -\frac{2}{n(n-1)} \sum_{i_1} \sum_{i_2 > i_1} \frac{\text{cov}_I(r_{i_1}, r_{i_2})^2}{\text{var}_I(r_{i_1}) \times \text{var}_I(r_{i_2})},$$

$$\Psi_{\text{Sparse}} \equiv -\frac{1}{n} \sum_i \left\langle \log \left(1 + \frac{r_i^2}{\langle r_i^2 \rangle} \right) \right\rangle.$$

Here $\langle r_i \rangle$ denotes the temporal average of activity of the i -th neuron.

¹ Please note that the set of stimuli is assumed to be fixed. In general, when the sensory system is part of a behaving agent, the statistical properties of stimuli may be dependent on the generated behaviour (Verschure and Pfeifer 1992). Although these are interesting aspects, they are beyond the scope of the present work.

To apply a gradient method we need a parameterization of non-linear features. Previous work used two-subunit energy detectors or general second order polynomials (Wiskott and Sejnowski 2002; Körding et al. 2004).

A second constraint proposed here is to use the predictability of features as an objective function:

$$\Psi_{\text{pred}} = \frac{1}{NN_p} \sum_{f(j)=i} \sum_i \frac{\text{cov}_I(p_{f(j)}(\tilde{r}(t)), r_i(t + \Delta t))^2}{\text{var}_I(p_{f(j)}(\tilde{r}(t))) \times \text{var}_I(r_i(t))}.$$

Here $p_{f(j)}(\tilde{r}_i(t)) \in \mathbf{P} = \{p_i | i : 1, \dots, n\}$ is a prediction made for neuron j on a set of responses $\tilde{r}_i(t)$, which in general may be based not only on the neuron under consideration, and NN_p is the average number of predictions acting on a neuron.²

This objective function uses the notion of predictive units and minimizing the difference between prediction of the activities resulting from feature extraction and actual future observations. Although the objective functions for decorrelation and predictions superficially appear similar, there are profound differences:

The predictive units may not map input patterns onto real numbers, but on an activity pattern of the same dimensionality. In that case, the space of possible functions would even be larger, and a low dimensional parameterization is not only a convenience, but a necessity. We suggest using local affine transformations for this purpose; second order autoregressive models might be an interesting alternative.

Ψ_{decorr} and Ψ_{pred} have opposite sign. That means that we, want low covariance of the local filter operations but high covariance of the predictions. In other words, we require local decorrelation and explicit use of global correlations. This is in accordance to the re-interpretation of Barlow's original idea that redundancy reduction is a principle underlying sensorial processing (Barlow 1961, 2001). As pointed out in (Barlow 2001), redundancies are essential for sensorial processing and need to be preserved and utilized. This line of argumentation matches the general concept, that it is in the interest of any agent to predict relevant stimuli (Roelfsema 2002; Wörgötter and Porr 2005). In this way, future rewards may be maximized, or dangerous actions avoided (Schultz and Dickinson 2000). Furthermore, in some aspects it might be compared to Kalman filters. These provide an optimal mixture of a noisy measurement and a prediction based on previous measurements. Here in contrast, optimal predictors are coevolved with non-linear filters that can be optimally predicted.

A most important aspect of the suggested unsupervised learning scheme in comparison to previous work is that not only the predictive units are optimized, but also that the selectivity of the feature extractors are adapted to allow optimal predictability. This leads to an objective function

$$\mathbf{E}^I(\mathbf{F}, \mathbf{P}) = \Psi_{\text{decorr}} + \Psi_{\text{sparse}} + \Psi_{\text{pred}}$$

² Note that here we use predictions in time. An analogue formula can also be used for spatial predictions as occurring for example in Gestalt laws such as good continuation or symmetry.

depending on the set of filters \mathbf{F} as well as a set of predictions \mathbf{P} . But it also emphasizes an important difference: The hypothesis states not only that the cortex tries to predict future stimuli, but additionally it proposes that the predictability of a stimulus is in itself a selection criterion for further processing.

To exemplify our approach, in Box 2 we compute Ψ_{Decorr} , Ψ_{Sparse} and Ψ_{Pred} for representations on the pixel-level and for the level of visual primitives (as described in Box 1). At the pixel level we represent the image sequence by neuronal responses coding the three (r,g,b)-colour values. In this way we can compute Ψ_{Decorr} and Ψ_{Sparse} according to the formulas given above. Since the change of pixel positions between frames is known, a prediction for colour values in consecutive frames can be formulated and Ψ_{Pred} can be computed. In a way analogous to that, as for the pixel level we can compute the three terms at the level of primitives making, however, use of a more powerful prediction taking the local orientation as an additional modality into account. It can be seen that in this example all three terms show higher values for the higher level primitive representation indicating that the differences in the representations are reflected in the three objective functions.

For a higher level representation we identify sparse positions with our visual primitives described in Box 1a. Here we represent the image sequence by neuronal responses coding the local orientation of the edges in a population code. In this way we can again compute Ψ_{Decorr} and Ψ_{Sparse} . The change of positions in the image can be computed as for the pixel level. In addition, in this representation we can compute the change of orientation under the motion (see, e.g., Krüger et al. 2002). Thus, we can realize a more sophisticated prediction by extending the prediction of a position to another modality (orientation). Then, like for the pixel level we can compute Ψ_{Pred} .

In the figure above, we see that the three terms have significantly different values for the two different levels. Besides the increase of the decorrelation and sparseness term, it also becomes clear that more sophisticated and powerful predictions can be defined at the higher level representation. As a consequence, we can deduce that when we impose the additional constraint predictability in the objective function we shall find features of structure other than pixels in our search space that fulfil the objectives in a better way.

The analysis above gives an example of the contribution of the three terms Ψ_{Decorr} , Ψ_{Sparse} and Ψ_{Pred} that we have suggested for an extended objective function. However, we are well aware that in future work a systematic investigation is required, including a realization of unsupervised learning upon a large database of natural stimuli.

6 Emergence of symbols

The process outlined above optimizes predictions at the same time as feature selectivity. When this is implemented in a processing hierarchy, optimizing sensory representations and matching predictors in parallel, it will produce entities which

not only represent the input stimuli in a sensible way but code context structure in terms of relations of visual events. We postulate that the data format for the learned feature selectivity will lead to representations which differ fundamentally in the dynamic properties compared to early vision representations.

The criteria of predictability and condensation relate directly to the two defining properties of symbols of expression and syntax. The optimized process of feature selectivity and predictions requires entities that represent condensed pieces of knowledge that are not analogue anymore. Furthermore, these entities work as arguments in a complex structural framework of transformations that realize predictions. These transformations generate new entities and relate them to other entities. Thus, the high level representations differ from the feature selectivity that results from learning without predictive mechanisms in the way that symbol-like structures can emerge:

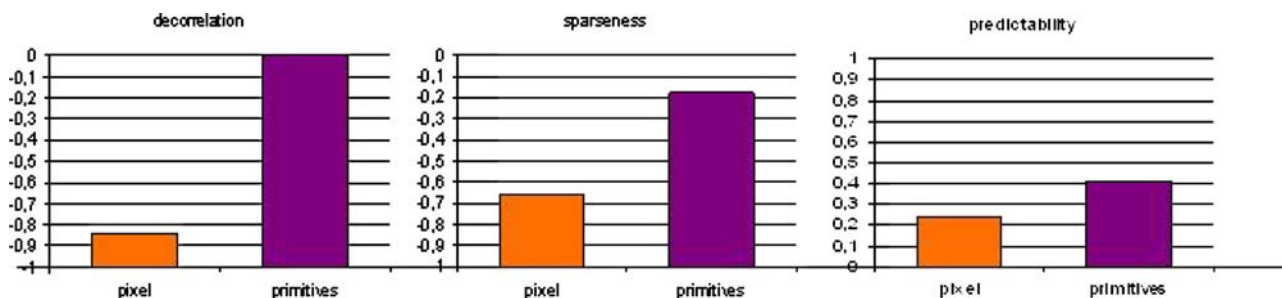
Hypothesis 4: In the process of mutual optimization of features and predictions, symbols emerge as condensed entities on which predictions are performed.

The symbols become representatives of structures and their relations in the real world. By that world knowledge becomes intrinsic structure of the sensorial machinery. As a consequence, these high level representations are justified only by the matching of high-level sensory representations; like symbols in a formal system they are not directly justified in the real world but they become indirectly justified by their predictive relations which express structural properties of natural scenes.

7 Discussion

We sketched a process that gradually transforms analogue local representations into discrete binary representations by means of four hypotheses. They are based on the notion that a proper understanding of early cognitive vision requires the integration of predictive processes. Our central assumption is that in this process entities emerge that fulfil the criteria defining symbols and that by a mutual learning of feature selectivity and predictions of these features, symbol-like structures emerge. By this, we outline a process in which symbol-like structures can become justified by an extension of unsupervised learning schemes already successfully applied in early vision.

This proposal builds on many previous studies. Tishby, Bialek and coworkers have pointed out that information on the environment, by necessity, is outdated by the time it reaches cortex. In order to be useful for behaviour, it must be turned into a prediction of the environmental context at the time of action (Tishby et al. 1999; Bialek et al. 2001). Furthermore, the concept of coherence infomax as proposed by Phillips and Singer (1997) introduces the constraint that cortical processes should maximize the mutual information



Box 2 For stereo sequences with known motion and depth structure we compute the three terms Ψ_{decorr} , Ψ_{sparse} and Ψ_{pred} at two levels of representations: the pixel level and the level of visual primitives (see box 1). At the pixel level, we represent an image sequence by a neuronal activity $r(k, l, m, t)$ where (k, l) is the pixel position, m represents the (r, b, g) colour channel and t represent time. This allows us to compute Ψ_{decorr} and Ψ_{sparse} according to the given formulas. Note that once the 3D position of a point in the first frame is known as well as the camera motion, we can compute the 3D position of that point in the next frame. Since the motion is known, we can compute for each pixel position where it will appear in the next frame. In other words, for each pixel position $(k, l)_t$ at time t we can realize a temporal prediction $p((k, l)_t) = (\tilde{k}, \tilde{l})_{t+1}$ where $(\tilde{k}, \tilde{l})_{t+1}$ is the pixel position where the occurrence of the corresponding pixel at time $t + 1$ is predicted. Using a common first order approximation we assume that the pixel intensity value does not change under the motion. Hence we arrive at a prediction $p(r(k, l, m, t)) = r(\tilde{k}, \tilde{l}, m, t)$ and can compute Ψ_{pred}

between parallel channels, a close analogy of a spatial prediction. Hence, the line of thought presented in this article is firmly grounded in previous work on signal processing and hypotheses that the principles of prediction and condensation might be more general and apply naturally to high-level cognitive processes involving symbolic representations also.

The four hypotheses inherit an increasing degree of speculation and also their experimental verification is increasingly difficult.

The first claim (H1) offers itself as a straightforward generalization of a concept, which was rather successful in primary sensory areas. As a consequence, predictions are straightforward and experimentally accessible. Receptive fields of neurons found in secondary visual cortex for example, should maximize a few well defined objective functions over the set of natural stimuli. Manipulations of the environment during development should lead to adaptation of sensory representations that are optimal with respect to the stated objective functions. This is very much in line with experiments of Merzenich and his colleagues (Nakahara et al. 2004; Zhang et al. 2001). However, experiments from the very same laboratory demonstrate that not only the statistics, but also associated rewards have a significant impact on the plasticity of sensory representations. Whether such effects can be described in one coherent framework has to be investigated.

The second hypothesis is suggested using the example of disambiguating stereo images using motion information. A generalization to other features and modalities might not be that obvious. Firstly, motion itself is often considered a primary feature. It can give rise to shape information, and neurons throughout the brain are sensitive to motion cues. But this need not to be a conflict. Indeed, recent experiments on motion sensitivity of retinal ganglion cells uncover a close relation to predictions of future stimuli (Berry et al. 1999). Thus, the marked sensitivity of neurons in the visual system to motion cues might be directly related to predicting the future.

The third hypothesis (H3) is supported by indirect evidence on the plasticity of auditory representations (c.f. Buonomano and Merzenich 1998). Recent work on optimal representations in a hierarchical system applying a stability objective supports the hypothesis as well. Yet, this is obviously just a beginning.

The final suggestion (H4) is the most difficult to test. A simulation with a technical agent in a controlled real world environment offers the best perspective of an investigation of the complete system and a test of high-level representations. An experimental test of human high level representations at that level of temporal and spatial resolution does not seem currently feasible.

We started with a discussion of cortical microanatomy, and now have to consider the implications of the present work for our view on the tessellation of cortex into functionally different areas. The intensively investigated primary sensory areas display a distinctively different laminar pattern. Furthermore, this pattern relates specifically to different classes of afferent fibres (Callaway 1998). It might be argued that such special structural properties severely limit any claim on general learning properties. Such a point is well taken, and in spite of the name unsupervised “learning”, it is essentially an optimization procedure. It can act on different time-scales, within an individual as well as within a population. This implies that the peculiarities of distal optical signals, which are constant on an evolutionary time scale, require specialized circuits for optimal processing, which may be taken care of by specific genetic adaptations on a long time scale. Proximal signals, like higher order representations, are more variable on an evolutionary time scale and hence do induce limited structural adaptation. In this view, the specific laminar structure of primary sensory and motor areas is an argument not against, but in favour of the hypotheses put forward here.

Acknowledgements We would like to thank Michael Felsberg, Bill Phillips, Laurenz Wiskott, Florentin Wörgötter and Saskia Nagel for

comments on an earlier version of the manuscript. Special thanks goes to Nicolas Pugeault who performed the simulations shown in Box 2.

References

- Aloimonos J, Shulman D (1989) Integration of visual modules: an extension of the Marr paradigm. Academic Press, Boston
- Ashby FG, Maddox WT (2005) Human category learning. *Annu Rev Psychol* 56:149–178
- Barlow H, Blakemore C, Pettigrew JD (1967) The neural mechanisms of binocular depth discrimination. *J Physiol (Lond)* 193:327–342
- Barlow HB (1961) Possible principles underlying the transformation of sensory messages. In: Rosenblith WA (ed). *Sensory communication*, vol 1961. MIT, Cambridge, pp 217–234
- Barlow HB (2001) Redundancy reduction revisited. *Network Comput Neural Syst* 12(3):241–254
- Berkes P, Wiskott L (2005) Slow feature analysis yields a rich repertoire of complex cell properties. *J Vision* 5(6):579–602
- Berry MJ II, Brivanlou IH, Jordan TA, Meister M (1999) Anticipation of moving stimuli by the retina. *Nature* 398:334–338
- Betsch BY, Einhäuser W, Körding KP, König P (2004) The world from a cat's perspective—statistics of natural videos. *Biol Cybern* 90(1): 41–50
- Bialek W, Nemenman I, Tishby N (2001) Predictability, complexity, and learning. *Neural Comput* 13(11):2409–2463
- Braitenberg V, Schüz A (1991) *Anatomy of the cortex*. Springer, Berlin Heidelberg New York
- Brodmann K (1906) Beiträge zur histologischen Lokalisation der Grosshirnrinde. Fünfte Mitteilung: über den allgemeinen Bauplan des Cortex pallii bei den Mammalieren und zwei homologe Rindenfelder im besonderen. Zugleich ein Beitrag zur Furchenlehre. *J Psychol Neurol* 6:275–400
- Buonomano DV, Merzenich MM (1998) Cortical plasticity: from synapses to maps. *Annu Rev Neurosci* 21:149–186
- Callaway EM (1998) Local circuits in primary visual cortex of the macaque monkey. *Annu Rev Neurosci* 21:47–74
- Chichocki A, Amari S-I (2002) Adaptive blind signal and image processing. In: *Learning algorithms and applications*. Wiley, New York
- Douglas RJ, Martin KA (2004) Neuronal circuits of the neocortex. *Annu Rev Neurosci* 27:419–451
- Einhausen W, Kayser C, Körding KP, König P (2003) Learning distinct and complementary feature selectivities from natural colour videos. *Rev Neurosci* 14(1–2):43–52
- Elder JH, Goldberg RM (2002) Ecological statistics of Gestalt laws for the perceptual organization of contours. *J Vision* 2(4):324–353
- Field D (1987) Relations between the statistics of natural images and the response properties of cortical cells. *J Opt Soc of Am* 4(12): 2379–2394
- Geisler WS, Perry JS, Super BJ, Gallogly DP (2001) Edge co-occurrence in natural images predicts contour grouping performance. *Vis Res* 41:711–724
- Gibson JJ (1979) *The ecological approach to visual perception*. Houghton Mifflin, Boston, MA
- Gilbert CD, Wiesel TN (1989) Columnar specificity and intrinsic horizontal and cortico-cortical connections in cat visual cortex. *J Neurosci* 9:2432–2442
- Grill-Spector K, Malach R (2004) The human visual cortex. *Annu Rev Neurosci* 27:649–677
- Hafner VV, Fend M, König P, Körding KP (2004) Predicting properties of the rat somatosensory system by sparse coding. *Neural Inf Process* 4:11–18
- Harnard S (1990) The symbol grounding problem. *Physica D* 42: 335–346
- Hershler O, Hochstein S (2005) At first sight: a high-level pop out effect for faces. *Vis Res* 45:1707–1724
- Hilbert D (1928) Die Grundlagen der Mathematik. *Abhandlungen aus dem mathematischen Seminar der Universität Hamburg* 6:65–85
- Hipp J, Einhäuser W, Conrath J, König P (2005) Unsupervised learning of somatosensory representations for texture discrimination using a temporal coherence principle. *Network Comput Neural Syst* (in press)
- Honavar V, Uhr L (1994) *Artificial intelligence and neural networks: steps toward principled integration*. Academic, New York, NY
- Hubel DH, Wiesel TN (1959) Receptive fields of single neurones in the cat's striate cortex. *J Physiol* 148:574–591
- Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* 160: 106–154
- Hurri J, Hyvärinen A (2003) Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Comput* 15: 663–691
- Hyvärinen A, Hoyer P (2000) Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Comput* 12:1705–1720
- Hyvarinen A, Hurri J, Vayrynen J (2003) Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *J Opt Soc Am A Opt Image Sci Vis* 20(7):1237–1252
- Jones JP, Palmer LA (1987) An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J Neurophysiol* 58(6):1233–1258
- Kayser C, Körding KP, König P (2004) Processing of complex stimuli and natural scenes in the visual cortex. *Curr Opin Neurobiol* 14: 468–473
- Kjaer TW, Gawne TJ, Hertz JA, Richmond BJ (1997) Insensitivity of V1 complex cell responses to small shifts in the retinal image of complex patterns. *J Neurophysiol* 78(6):3187–3197
- Klette R, Schlüns K, Koschan A (1998) *Computer vision—three-dimensional data from images*. Springer, Berlin Heidelberg New York
- Körding KP, Kayser C, Einhäuser W, König P (2004) How are complex cell properties adapted to the statistics of natural stimuli? *J Neurophysiol* 91(1):206–212
- Kreiman G, Koch C, Fried I (2000) Imagery neurons in the human brain. *Nature* 408:357–361
- Krüger N (1998) Collinearity and parallelism are statistically significant second order relations of complex cell responses. *Neural Process Lett* 8(2):117–129
- Krüger N, Ackermann M, Sommer G (2002) Accumulation of object representations utilizing interaction of robot action and perception. *Knowl Based Syst* 15:111–118
- Krüger N, Lappe M, Wörgötter F (2004) Biologically motivated multimodal processing of visual primitives. *AISB J* 1(5):417–428
- Krüger N, Wörgötter F (2004) Statistical and deterministic regularities: utilisation of motion and grouping in biological and artificial visual systems. *Adv Imaging Electron Phys* 131:82–147
- Krüger N, Wörgötter F (2005) Multi-modal primitives as functional models of hyper-columns and their use for contextual Integration. In: *Proceedings of the 1st international symposium on brain, vision and artificial intelligence 2005*, LNCS 3704. Springer, Berlin Heidelberg New York, p 157–166
- Lettvin JY, Maturana HR, McCulloch WS, Pitts WH (1959) What the frog's eye tells the frog's brain. *Proc IRE* 47:1940–1951
- Linsker R (1988) Self-organization in a perceptual network. *Computer* 21:105–117
- Maunsell JHR, Newsome WT (1987) Visual processing in monkey extrastriate cortex. *Annu Rev Neurosci* 10:363–401
- Nakahara H, Zhang LI, Merzenich MM (2004) Specialization of primary auditory cortex processing by sound exposure in the “critical period”. *Proc Natl Acad Sci USA* 101:7170–7174
- Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381(6583):607–609
- Olshausen BA, Field DJ (2004) Sparse coding of sensory inputs. *Curr Opin Neurobiol* 14(4):481–487
- Olshausen BA, Field DJ (2005) How close are we to understanding v1? *Neural Comput* 17(8):1665–1699

- Phillips WA, Singer W (1997) In search of common foundations for cortical computation. *Behav Brain Sci* 20:657–683
- Orbach J (1998) *The neuropsychological theories of Lashley and Hebb*. University Press of America
- Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I (2005) Invariant visual representation by single neurons in the human brain. *Nature* 435(7045):1102–1107
- Ringach RL (2002) Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J Neurophysiol* 88:455–463
- Ringach DL, Hawken MJ, Shapely R (2002) Receptive field structure of neurons in monkey visual cortex revealed by stimulation with natural image sequences. *J Vision* 2:12–24
- Ringach DL (2004) Mapping receptive fields in primary visual cortex. *J Physiol* 558:717–728
- Roelfsema PR (2002) Do neurons predict the future? *Science* 295(5553):227
- Schiller PH, Finlay BL, Volman SF (1976) Quantitative studies of single-cell properties in monkey striate cortex. III. Spatial frequency. *J Neurophysiol* 39:1334–1351
- Schultz W, Dickinson A (2000) Neuronal coding of prediction errors. *Annu Rev Neurosci* 23:473–500
- Steels L (2003) Evolving grounded communication for robots. *Trends Cog Sci* 7(7):308–312
- Tishby NZ, Pereira F, Bialek W (1999) The information bottleneck method. In: Hajek B, Sreenivas RS (eds) *Proceedings of the 37th Allerton Conference on communication, control and computing*, Urbana, Illinois, 1999. University of Illinois, Illinois
- Ullman S (1979) *The interpretation of Visual Motion*. MIT, Cambridge, MA
- Vargha-Khadem F, Gadian DG, Copp A, Mishkin M (2005) FOXP2 and the neuroanatomy of speech and language. *Nat Rev Neurosci* 6:131–138
- Verschure PFMJ, Pfeifer R (1992) Categorization, representations, and the dynamics of system-environment interaction: a case study in autonomous systems. In: Meyer JA, Roitblat H, Wilson S (eds) *From animals to animats: proceedings of the 2nd international conference on simulation of adaptive behavior*, Honolulu, Hawaii. MIT, Cambridge, MA pp 210–217
- Watt RJ, Phillips WA (2000) The function of dynamic grouping in vision. *Trends Cog Sci* 4(12):447–454
- Wiskott L, Sejnowski TJ (2002) Slow feature analysis: unsupervised learning of invariances. *Neural Comput* 14(4):715–770
- Wörgötter F, Porr B (2005) Temporal sequence learning, prediction, and control: a review of different models and their relation to biological mechanisms. *Neural Comput* 17(2):245–319
- Zhang LI, Bao S, Merzenich MM (2001) Persistent and specific influences of early acoustic environments on primary auditory cortex. *Nat Neurosci* 4:1123–1130