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

Peter König[1] and Norbert Krüger[2]

[1]Institute of Cognitive Science, Dept. Neurobiopsychology (University of Osnabrück, Germany)

[2]Cognitive Vision Group, MediaLab (Aalborg University Copenhagen, Denmark)

Correspondence to:

Peter König

Institute of Cognitive Science

University Osnabrück

Albrechtstr. 28

49076 Osnabrück

Germany

Phone: +49-541-9692399

Fax: +49-541-9692596

Email: pkoenig@uos.de

Abstract

In the mammalian cortex the early sensory processing, e.g. vision, 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 identical 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 is 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 on 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 characterising 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; Hubel and Wiesel 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 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 3-dimensional 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 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 results also address selectivity with respect to other features, such as motion (Berkes and Wiskott in press), disparity (Onat et al. submitted) and colour (Einhäuser et al. 2003). This allows the perspective that a substantial part of feature selectivity in early visual processing can be understood on the basis of unsupervised learning according to a small number of objectives. In this way, local filter operations similar to ones known from neuropysiological investigations become grounded in the structural properties of natural scenes and by a number of principles such as sparseness, stability and independence.

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 to 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

(SE) symbols are condensed and discrete semantic representatives for certain pieces of knowledge (Expression)

(SS) on which operations can be performed that correspond to relevant functional relations in this framework (Syntax).

Symbol manipulation typically identifies symbolic expressions, decomposes given expressions and generates new 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 (Harnad 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 chosen deliberately representing 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 it has 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 from 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 microcircuits 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 1996). It implies that the structure of cortical circuits is generic and potentially serves any function (equipotentiality) and substantial parts of 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 identical 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 (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 criteria SE and SS defining symbols.

2 Learning of Feature maps from natural scenes

Following the flow of information, we first study processing of stimuli in early sensory 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 (Lettvin et al. 1959), 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 (Kulvicius et al. submitted). 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 in press). Here as well the relevance for behaviour is indirect, as important aspects of visual stimuli are supposed to change slower than irrelevant detail. Further progress has been made addressing selectivity to other features and in other visual modalities (Hurri and Hyvärinen 2003; Hafner et al., 2004). Although there are still numerous unsolved problems, 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.

If unsupervised learning is so 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 in subsequent levels representations of more complex features emerge (Wiskott and Sejnowski 2002; Wyss et al. submitted; Franzius et al. submitted). Hence, one may speculate that the approach described above and taken 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 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). The argument is based on the view that not relevant aspects, i.e. noise, to be 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 the following section, we describe possible extensions of currently used objective functions.

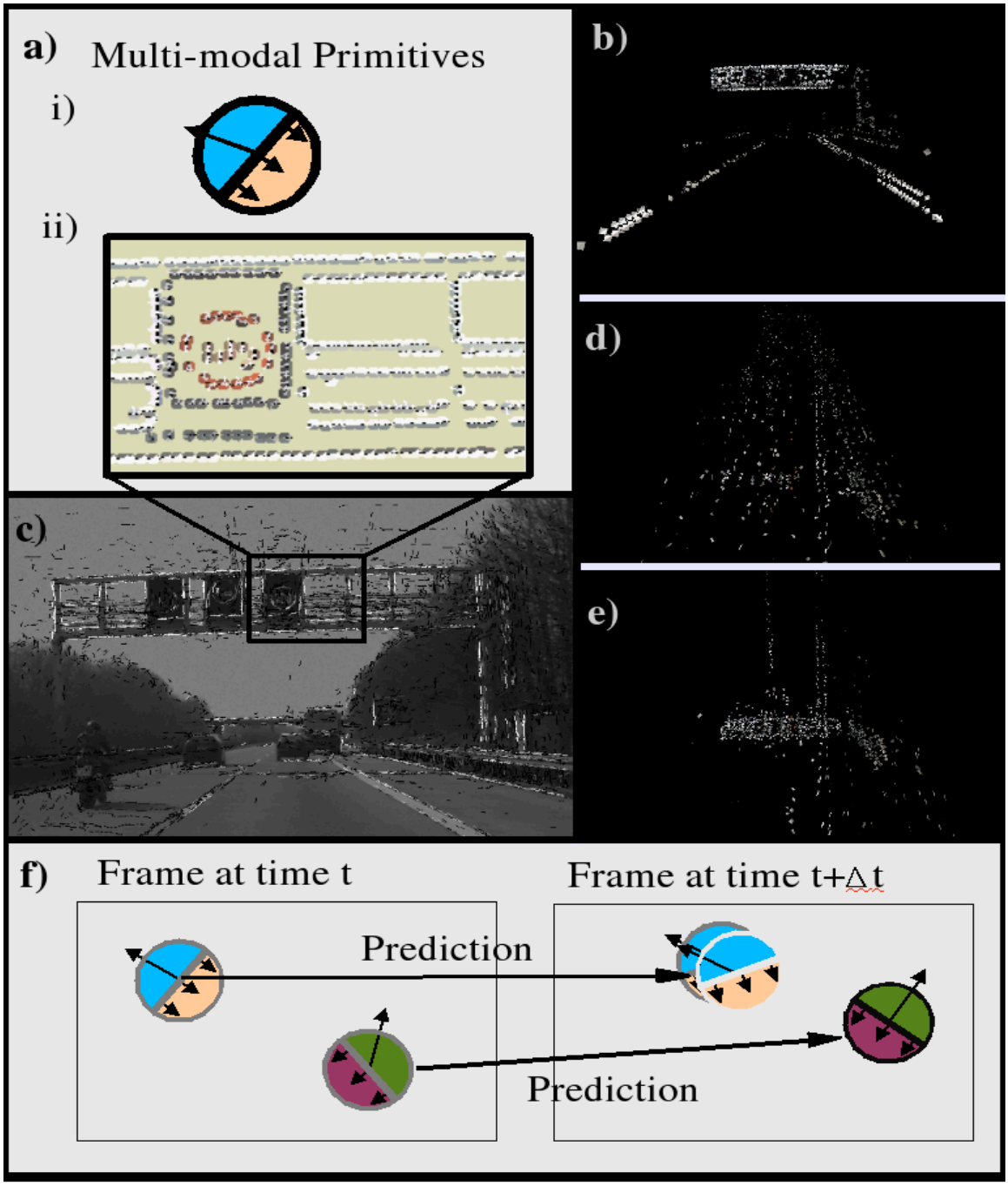
3 Predictive mechanisms for disambiguation

The problem of extracting relevant features also shows up in the form of resolving ambiguities. At 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 two dimensional image (disparity) is an indicator of depth in the three dimensional 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 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 three-dimensional 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 3-dimensional 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 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).



Box 1. **a)** Condensed representation of local image information by multi-modal primitives. One primitive represents a local image patch in terms of a low-dimensional descriptor containing information about the local orientation, phase, colour, local motion, disparity and a descriptor of the 'homogeneity', 'edge-ness', or 'junction-ness' of the local patch. Condensation rate is higher than 95% as computed by the number of bits used to represent an image by primitives compared to the original image. For details see (Krüger

et al. 2004). **b)** Front view of primitives extracted from the scene in c) and found as reliable after the disambiguation process. **c-f)** Disambiguating stereo images by predictions based on motion information. **c)** Highway scene with correct and erroneous predictions. The grey level indicates the resulting confidence level, with white representing highest confidence. Note the large amount of non-verified predictions represented by the nearly black line segments. **d)** Top view of the 3D candidate set extracted from an ordinary stereo matching. Obviously it contains many invalid matches. When the threshold is increased for accepting a match as a candidate for correspondence many valid correspondences are discarded as well (data not shown). However, with the information on ego motion we convert each candidate match to a prediction for the next frame. After a few iterations a more complete representation with only few outliers is generated. **e)** Shows a top view of the disambiguated 3D representation. **f)** Accumulation scheme based on spatial-temporal predictions. When a prediction matches the subsequently sampled data the confidence of a candidate match as well as the associated 3-dimensional interpretation is increased, otherwise decreased. Confidences are represented by the brightness of the surrounding circle and the orientation bar.

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 my escape by taking a sudden turn. However, if spotting a flying insect does not increase the probability that it is in any spatial region a short time later when the frog reacts, the visual input might as well be ignored. 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 (CS): A good feature gives rise to the prediction of other temporally and/or spatially distinct features and needs to be predictable from those.

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 efficiently 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 in press) the condensation of local visual information into a semi-symbolic format (see figure 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:

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

Taking both criteria CS and CE into account we can formulate our third hypothesis:

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

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

5 A concrete approach

The essence of learning feature selectivity as described in previous work rests on the definition of an objective function $\mathbf{E}^{I}(\boldsymbol{F})$. It is dependent on a set of local filter operations

$\boldsymbol{F}=\{f_i|i:1,..,n\}$ that are applied to a set of natural stimuli $\boldsymbol{I}$, mapping each stimulus $I \_ \boldsymbol{I}$ 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 the case different aspects have to be optimized simultaneously, $\mathbf{E}^{\boldsymbol{I}}(\boldsymbol{F})$ is composed as a sum of a number of terms, addressing the individual aspects. To achieve condensation of information (CE) a sparseness/decorrelation principle possibly connected with some constraints on the connectivity structure of the neural net is a good start.

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

The relative weight of the two terms depends on the precise formulae. For a combination of a stability and decorrelation (Hipp et al. submitted) 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.

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

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

To apply a gradient method we need a parameterization of non-linear features. Previous work used 2-subunit energy detectors or general 2nd 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:

---

[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.

$$\Psi_{pred} = \frac{2}{N\ (N\ -1)} \sum_{f(j)=i} \sum_{i} \frac{\mathrm{cov}_I(p_{f(j)}(\tilde{r}\ (t)), r_i\ (t+\Delta t))^2}{\mathrm{var}_I(p_{f(j)}(\tilde{r}\ (t))) \cdot \mathrm{var}_I(r_i\ (t))}$$

Here $p_{f(j)}(r\sim(t)) \in \boldsymbol{P}=\{p_i|i:1,..,n\}$ is a prediction made for neuron j on a set of responses r~(t), which in general may be based not only on the neuron under consideration.[2] Note that the predictions may not be based on only one response but on a set of responses.

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 do not map input patterns onto a real numbers, but on an activity pattern of the same dimensionality. Thus, the space of possible functions is even larger, and a low dimensional parameterization is not only a convenience, but a necessity. We suggest using local affine transformations for this purpose; 2nd order autoregressive models might be an interesting alternative.

Ψ_{decorr} and Ψ_{pred} have opposite sign. That means that we are after low covariance of the local filter operations but after high covariance of the predictions. In other words, we require local decorrelation and an 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). Furthermore, it shares many similarities with the concept of coherence infomax as proposed by Phillips and Singer (1997). 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.

---

[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.

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

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

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

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 which 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 (CS) and condensation and (CE) relate directly to the two defining properties of symbols of expression (SE) and syntax (SS). 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 grounded only by the matching of high-level sensory representations; like symbols in a formal system they are not

directly grounded in the real world but they become indirectly grounded 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 outlined. They are based on the notion that a proper understanding of early cognitive vision requires the integration of predictive processes (CS, CE). Our central assumption is that in this process entities emerge that fulfil the criteria defining symbols (SS, SE) and that by a mutual learning of feature selectivity and predictions on these features symbol-like structures emerge. By this, we outline a process in which symbol-like structures can become grounded by an extension of unsupervised learning schemes already successfully applied in early vision.

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 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 needs 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 2nd 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 most difficult to test. A simulation with a technical agent in a controlled real world environment offers the best perspective on an investigation of the complete system and a test of high-level representations. An experimental test of human high level representations on that level of temporal and spatial resolution does not seem currently feasible.

What is the significance of known differences in microanatomy of cortical 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.

Acknowledgement

We would like to thank Michael Felsberg, Bill Phillips, Laurenz Wiskott and Florentin Wörgötter for comments on an earlier version of the manuscript.

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—78

Barlow H, Blakemore C, Pettigrew JD (1967) The neural mechanisms of binocular depth discrimination. J Physiology (London) 193:327—342

Barlow HB (1961) Possible Principles underlying the Transformation of Sensory Messages. In Rosenblith, W.A. (editor). Sensory Communication, MIT 1961:217—234.

Barlow HB (2001) Redundancy reduction revisited. Network: Computation in Neural Systems. 12(3):241—254

Berkes P, Wiskott L (in press) Slow feature analysis yields a rich repertoire of complex cell properties. J Vision

Berry MJ 2nd, Brivanlou IH, Jordan TA, Meister M (1999) Anticipation of moving stimuli by the retina. Nature 398:334—8.

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

Braitenberg V, Schüz A (1991) Anatomy of the cortex. Berlin, Springer Verlag

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 Psychologie Neurologie 6:275-400

Buonomano DV, Merzenich MM (1998) Cortical plasticity: from synapses to maps. Annu Rev Neurosci 21:149—86

Callaway EM (1998) Local circuits in primary visual cortex of the macaque monkey. Annu Rev Neurosci 21:47-74

Douglas RJ, Martin KA (2004) Neuronal circuits of the neocortex. Annu Rev Neurosci 27:419-51

Einhäuser 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. Journal of the Optical Society of America 4(12): 2379—2394.

Franzius M, Einhäuser W, König P, Körding KP (submitted) Learning a hierarchical model of cortical function from natural stimuli

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. Boston, MA: Houghton Mifflin

Gilbert CD, Wiesel TN (1989) Columnar specificity and intrinsic horizontal and cortico-cortical connections in cat visual cortex. J Neuroscience 9:2432-42

Grill-Spector K, Malach R (2004) The human visual cortex. Annu Rev Neurosci 27:649-77

Hafner VV, Fend M, König P and Körding KP (2004) Predicting properties of the rat somatosensory system by sparse coding. Neural Information Processing 4:11-18

Harnard S (1990). The symbol grounding problem. Physica D 42:335-346, 1990

Hershler O, Hochstein S (2005) At first sight: A high-level pop out effect for faces. Vis Res 45:1707-24

Hilbert D (1928) Die Grundlagen der Mathematik. Abhandlungen aus dem mathematischen Seminar der Unversität Hamburg 6:65-85

Hipp J, Einhäuser W, Conradt J, König P (submitted) Unsupervised learning of somatosensory representations for texture discrimination using a temporal coherence principle.

Honavar V, Uhr L (1994) Artificial Intelligence and Neural Networks: Steps Toward Principled Integration. (Ed) New York, NY: Academic Press

Hubel DH, Wiesel TN (1959) Receptive fields of single neurones in the cat's striate cortex. J Physiol 148:574-91

Hubel DH, Wiesel TN (1962) Receptive Fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiology 160:106-154

Hurri J, Hyvärinen A (2003) Simple-cell-like receptive fields maximize temporal coherence in natural video. Neural Comput 15:663– 691

Hyväarinen 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

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):1233-58

Kayser C, Körding KP, König P (2004) Processing of complex stimuli and natural scenes in the visual cortex. Curr Opin Neurobiol 14:468-73

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):3187-97

Klette R, Schlüns K, Koschan A (1998) Computer Vision - Three-Dimensional Data from Images. Springer.

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):206-12

Kreiman G, Koch C, Fried I (2000) Imagery neurons in the human brain. Nature 408:357-61

Krüger N (1998) Collinearity and parallelism are statistically significant second order relations of complex cell responses. Neural Processing Letters 8(2):117—129

Krüger N, Lappe M, Wörgötter F (2004) Biologically motivated multi-modal processing of visual primitives. AISB Journal 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 (in press). Multi-modal Primitives as functional Models of Hyper-columns and their use for contextual Integration. Proceedings of the 1st International Symposium on Brain, Vision and Artificial Intelligence 2005, Lecture Notes in Computer Science, Springer.

Kulvicius T, Porr B, Wörgötter F (submitted) Behaviorally Guided Development of Primary and Secondary Receptive Fields.

Lettvin JY, Maturana HR, McCulloch WS, Pitts WH (1959) What the Frog's Eye Tells the Frog's Brain. Proc IRE 47:1940-51

Linsker R (1988) Self-organization in a perceptual network. Computer 21:105-17

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 101:7170-4

Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381(6583):607-9

Olshausen BA, Field DJ (2004) Sparse coding of sensory inputs. Curr Opin Neurobiol 14(4):481-7

Onat S, Kayser C, König P (submitted) On the time course of disparity in natural visual stimuli.

Phillips WA, Singer W (1997) In search of common foundations for cortical computation. Behav Brain Sci 20:657-83

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):1102-7

Ringach DL (2004) Mapping receptive fields in primary visual cortex. J Physiol 558:717-28

Roelfsema PR (2002) Do neurons predict the future? Science 295(5553):227

Schiller PH, Finlay BL, and 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

Ullman S (1979) The interpretation of Visual Motion. MIT Press, Cambridge, MA

Vargha-Khadem F, Gadian DG, Copp A, Mishkin M. (2005) FOXP2 and the neuroanatomy of speech and language. Nat Rev Neurosci 6:131-8

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, editors, From Animals to Animats: Proceedings of the Second International Conference on Simulation of Adaptive behavior. Honolulu: Hawaii, pages 210-217. Cambridge, Ma., MIT press.

Watt RJ, Phillips WA (2000) The Function of Dynamic Grouping in Vision. Trends Cog Sci 4(12):447—154

Wiskott L, Sejnowski TJ (2002) Slow feature analysis: unsupervised learning of invariances. Neural Comput 14(4):715-70

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

Wyss R, König P, Verschure PFMJ (submitted) The computational principles of temporal stability and local memory can account for the structural and functional organization of the ventral visual system.

Zhang LI, Bao S, Merzenich MM (2001) Persistent and specific influences of early acoustic environments on primary auditory cortex. Nat Neurosci 4:1123-30