Extracting Slow Subspaces from Natural Videos Leads to Complex Cells

Christoph Kayser, Wolfgang Einhäuser, Olaf Dümmer,
Peter König, and Konrad Körding

Institute of Neuroinformatics, ETH / University Zürich
Winterthurerstr. 190, 8057 Zürich, Switzerland
{kayser,weinhaeu,olaf,peterk,koerding}@ini.phys.ethz.ch

Abstract. Natural videos obtained from a camera mounted on a cat's head are used as stimuli for a network of subspace energy detectors. The network is trained by gradient ascent on an objective function defined by the squared temporal derivatives of the cells' outputs. The resulting receptive fields are invariant to both contrast polarity and translation and thus resemble complex type receptive fields.

Keywords: Computational Neuroscience, Learning, Temporal Smoothness

1 Introduction

A large body of research addresses the problem of obtaining selective responses to a class of stimuli (e.g. Hebb 1949, Grossberg 1976, Oja 1982) but surprisingly few results exist on learning representations invariant to given transformations. But real world problems like recognition tasks do not only require the network to be specific to the relevant stimulus dimensions but also to be insensitive to the irrelevant dimensions (e.g. Fukushima 1988). In this paper we address the problem of learning translation invariance from natural video sequences, pursuing an objective function approach. We implement the temporal smoothness criterion as proposed by Hinton (1989) and used by Földiak (1991). A generative model containing slowly changing hidden variables is assumed. The effect of these hidden variables on linear subspaces can be described by a mixing matrix. This mixing matrix is inverted by the search for slowly varying subspace energy detectors. Instead of mathematically deriving the objective function for these subspaces from an explicit generative model we here explore the effect of a given function on learning of nonlinear detectors. We analyze the obtained slow components and compare them with properties of complex type receptive fields of cortical cells.

2 Methods

The stimuli used to train our network consist of randomly chosen 10 by 10 patches sampled from a natural video recorded by a camera mounted on a cat's

G. Dorffner, H. Bischof, and K. Hornik (Eds.): ICANN 2001, LNCS 2130, pp. 1075–1080, 2001.

head (Betsch et al. submitted). Patches from the same spatial location within the image are taken from 2 subsequent images yielding a pair of intensity vectors I_{t-1} and I_t (images are sampled at 25 Hz). Each vector is normalized to mean zero. The complete stimulus set, consisting of 11000 such pairs, is reduced in dimensionality by PCA and whitened using the procedure described in Hyvärinen and Hoyer (2000). If not stated otherwise, the number of used principal components is 30 (in the following termed PCA dimension).

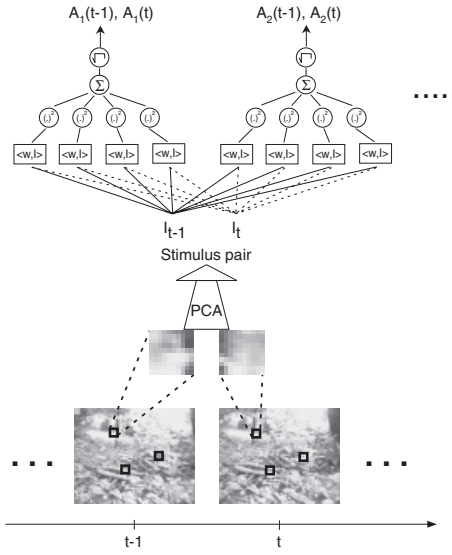

Fig. 1. Network layout. Two cells of the network together with the four sub-units are shown (top) Two images of the natural movie are shown together with patches used as stimuli (bottom).

For the reported results the network consists of 5 neurons each of which summes the input of 4 sub-units (Fig. 1). Each sub-unit has a weight-vector associated and the activity of sub-unit j of neuron i is calculated as the product $A_{ij} = W_{ij} \cdot I$. The neurons are modelled as subspace energy detectors (Kohonen 1996) and their activity is calculated as $A_i = \sqrt{\sum_j A_{ij}^2}$. The analyzed objective function is

$$O_{time} := - \sum_{\text{cells } i} \frac{\left\langle \left(\frac{\mathrm{d}}{\mathrm{dt}} A_i \right)^2 \right\rangle_t}{var_t(A_i)} \qquad (1)$$

where the mean ($<>_t$) and the variance are taken over time. In order to implement this in discrete time, the derivative is approximated by the difference of the activities for two consecutive patches, $A_i(t) - A_i(t-1)$. The variance is furthermore replaced by the product of the standard deviation taken over all the activities for the patches I_{t-1} times the standard deviation for the patches I_t. The network learns by changing the sub-unit weights W_{ij} following the (analytically calculated) gradient of (1) to a local maximum. The gradient ascent is controlled using adaptive stepsizes as described in Hyvärinen and Hoyer (2000) till a stationary state is reached. All sub-units are forced to be orthonormal in whitened space. The weights are randomly initialized with values between 0 and 1. The network layout together with two typical stimuli is shown Figure 1.

In order to quantify the properties of the learned cells their orientation and position specificity is calculated and displayed in θ-r diagrams: The cells are probed with Gaussian bars of defined orientation θ and position r as stimuli and the resulting activities displayed. From these diagrams two parameters are extracted: The *orientation specificity index* (σ_θ) is computed as the mean width of the orientation tuning over all positions. The *position specificity index* (σ_r) is

computed by first taking the standard deviation of the activity over all orientations at a fixed position and then averaging over all positions.

3 Results

In order to explore learning of invariant detectors a nonlinear network is implemented (see methods). We use neurons that compute the 2 norm of the corresponding sub-unit activities (Fig. 1). On the activities of these neurons an objective function characterizing their temporal smoothness, O_{time}, is defined and the network is trained till a stationary state is reached (Fig. 2A).

The resulting receptive fields of the sub-units largely resemble those of simple cells (Fig. 2B). After training every neuron receives input from a set of sub-units which all share the same orientation preference but differ in spatial localization. This is shown by the θ - r diagrams for the sub-units (Fig. 2C). Thus the resulting neurons are insensitive to the position of the stimuli and are therefore translation invariant (Fig. 2D). The system is also invariant with respect to the contrast polarity of the stimuli: The response for a bright bar on dark background is the same as for a dark bar on bright background. Note that this contrast polarity invariance is not learned by the network but instead is a built in feature of the transfer function of the neurons (since an even norm is used).

As an important control it is necessary to check that translation invariance is indeed a consequence of the temporal smoothness of the stimuli and not also an inherent network property. The stimulus vectors are randomly shuffled to destroy the temporal coherence of the pairs $\{I_{t-1}, I_t\}$. Figure 3A shows the resulting receptive fields of the sub-units, which no longer exhibit the systematic properties of those obtained with the stimuli in natural order. This shows that the correlations in the time domain of the video sequences are necessary for the learning of the complex like receptive fields.

Since the temporal correlation between patches in natural videos decays gradually over time (Betsch et al. Submitted) we pair frames of larger temporal distances ($\{I_{t-\Delta_n}, I_t\}$ instead of $\{I_{t-1}, I_t\}$). As expected, with growing time shift Δ_n orientation specificity decreases and the cells become more specific to position (Fig. 3B). In the limit of no correlation (large temporal distances or randomly paired frames) position and orientation specificity index become identical within error range.

In the current implementation the stimuli are whitened and all principal components up to the given PCA dimension are amplified to amplitude one whereas the other amplitudes are set to zero. One reason for this preprocessing is the large decrease in computation time when using fewer dimensions. To assess the effect of the choice of the PCA dimension the position and orientation specificity is computed for different dimensions (Fig. 3C). None of these quantities changes significantly. Inspection of the resulting sub-unit receptive fields and θ-r diagrams reveals that still complex like receptive fields are obtained (data not shown). But since the dimension of the stimulus space is now much larger than the number of feature detectors, the coverage of the stimulus space is coarse and most complex cells have similar preferences.

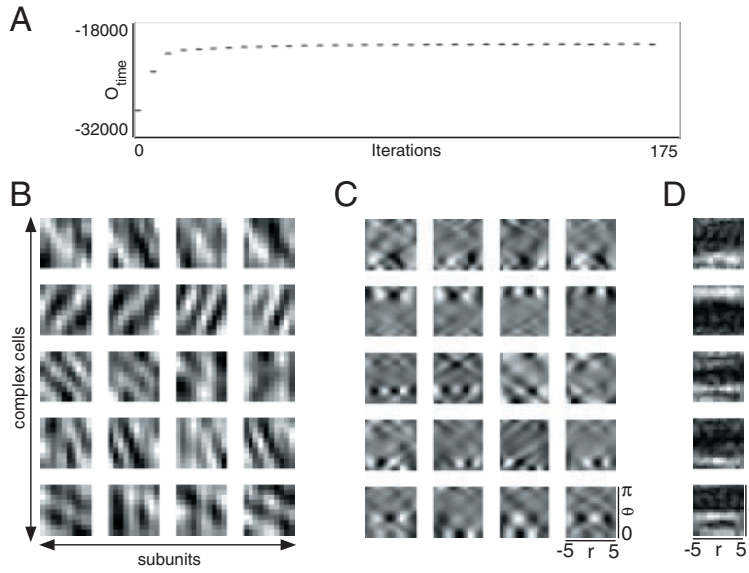

Fig. 2. Results. A) The objective function is optimized till a stationary state is reached. B) Receptive fields of the sub-units after 175 iterations. C) θ -r diagrams for these sub-units. The diagram shows the response strength of the unit for bars of different position (x-axis) and different orientation (y-axis). D) θ -r diagrams for the complex cells.

4 Discussion

The presented results show that complex like receptive fields can be learned by extracting the slowly varying subspaces of natural stimuli. The obtained receptive fields are comparable to those of Hyvärinen and Hoyer (2000) who use a different approach, independent subspace analysis (ISA). ISA uses the same network layout but implements a different objective, independence of the cells' responses, which is comparable to sparse coding. Whereas they use natural photographs taken from PhotoCDs we exploit the temporal domain of natural image sequences.

Another network for learning transformation invariant filters is the adaptive-subspace self-organizing map (ASSOM) proposed by Kohonen (1996). There also the neurons are modelled as sub-space energy detectors but the network learns a two dimensional map such that the activity maximum moves slowly over the network. The cells are implicitely forced to extract slowly varying features resulting in an approach comparable to the work of Foldiak and to the one presented here. Opposed to the ASSOM, the objective function approach incorporates the temporal smoothness in an explicit way and the results shown here were obtained from more natural stimuli.

The fact that quite different objectives lead to similar receptive fields poses the question to which degree the objectives of temporal smoothness and independence are equivalent.

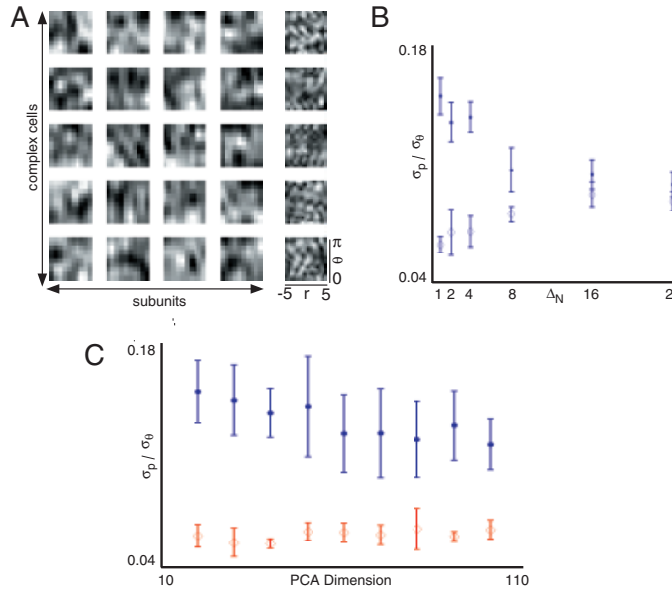

Fig. 3. Controls. A) (Left) Receptive fields of the sub-units for a network trained with randomly paired stimuli (no temporal coherence). (Rightmost column) θ -r diagrams for the (no longer complex) cells. B) Increasing the time lag Δ_N between two subsequent stimuli decreases the orientation specificity σ_θ (circles) and increases the position specificity σ_r (diamonds). Errorbars denote the standard deviation over all cells in the network. C) σ_θ (circles) and σ_r (diamonds) are shown as a function of the PCA Dimension.

It is interesting to note that the temporal smoothness function is very well compatible with a number of physiological mechanisms found in the mammalian cortex (Körding and König 2000). In this respect it is of importance that optimizing the objective function only needs information locally available to the cell.

A number of issues remain for further research: Different PCA dimensions require different subspace sizes and different numbers of neurons for optimal stimulus space coverage. Incorporating a dynamic subspace size in the objective function approach might recruit the optimal number of sub-units needed.

The presented results are obtained by using the 2-Norm of the subspace as transfer function for the cells. In this way the network becomes very similar to the classical energy models for complex cells which are supported by electrophysiological evidence. Some research on the other hand advocates stronger nonlinearities. Riesenhuber and Poggio (2000) for example propose the max function, which corresponds to the infinity norm. It seems likely that this network property can also be learned using the same objective function. Learning the norm of the sub-spaces might be worthwhile since it incorporates learning the nonlinearity of the network. Furthermore this could also lead to an explicitly learned contrast polarity invariance which so far is built in.

Concluding, temporal coherence is a method for learning complex type receptive fields from natural videos, and seems very well suited for learning different network properties of biological systems in which temporal information is ubiquitous.

Acknowledgments

This work was supported by the SNF (CK, PK) and the Boehringer Ingelheim Fonds (KPK). Furthermore we thank A. Hyvärinen and P. Hoyer for making their code available to the public. We are grateful to B. Betsch and C. Arielle for help with the acquisition of the stimulation videos.

References

Betsch, B.Y., Körding, K.P., Einhäuser, W., König, P. What cats see - statistics of natural images. Submited

Földiak, P. (1991). Learning Invariance from Transformation Sequences. Neural Computation, 3(2):194-200.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. Neural Networks, 1, 119-130.

Grossberg, S. (1976) A neuronal model of attention, reinforcement and discrimination learning. International Review of Neurobiology, 18:263-327.

Hebb, D.O. (1949) The Organization of Behaviour: A Neurophysiological Theory. Wiley

Hinton, G.E. (1989) Connectionist Learning Prcedures. Artificial Intelligence, 40:185-234.

Hyvarinen, A., and Hoyer, P.O. (2000). Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. Neural Comput., 12, 1705-20.

Kohonen, T. (1996). Emergence of Invariant-Feature Detectors in the Adaptive-Subspace SOM. Biological Cybernetics, 75(4):281-291.

Körding, K.P., König, P. Learning with two sites of synaptic integration. Network: Computation in neural systems, 11:1-15.

Oja, E. (1982) A simplified neuron model as a principal component analyzer. J. of Mathematical Biology, 15:267-273.

Riesenhuber, M., Poggio, T. (1999) Hierachical model of object recognition in cortex. Nature Neuroscience, 2(11):1019-1025