MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan

13-7

Using Non-negative Sparse Profiles in a Hierarchical Feature Extraction Network

Ingo Bax, Gunther Heidemann and Helge Ritter

Neuroinformatics Group
University of Bielefeld
P.O. Box 10 01 31, D-33501 Bielefeld, Germany

Abstract

In this contribution we utilize recent advances in feature coding strategies for a hierarchical Neocognitron-like neural architecture, which can be used for invariant recognition of natural visual stimuli like objects or faces. Several researchers have identified that sparseness is an important coding principle for learning receptive field profiles that resemble response properties of simple cells in visual cortex. However, an ongoing discussion is concerned with the question whether sparseness should be imposed on the latent variables as implicitly done in ICA or Sparse Coding or if it should rather be imposed directly on the feature matrix. Since answers to this question have so far not been unique and were rather qualitative in nature, this paper investigates the two possibilities by applying a recently introduced algorithm for Non-negative Matrix Factorization with Sparseness Constraints (NMFSC) to feature learning in a hierarchical recognition network. For this network, we compare recognition performance on several difficult image datasets under varying sparseness settings.

1 Introduction

While computational vision has made significant progress in the recognition of isolated objects under fixed imaging conditions, unrestricted environments are still a major challenge. Since biological vision is highly successful in solving problems like illumination variation, cluttered scenes, object deformations and occlusion, computer vision research increasingly draws upon physiological and psychophysical findings. Modern approaches that follow this paradigm often rely on the early findings by Hubel and Wiesel [6], who determined receptive fields of *simple cells* and *complex cells* in the primary visual cortex of mammals, and by Barlow [1], who analyzed the behavior of these cells and firstly suggested that their response properties might emerge from an efficient coding strategy in the sense of information theory.

A computational model to account for the idea of efficient coding was introduced by Olshausen and Field [9], who proposed the notion of *sparse coding* as a strategy of learning receptive fields from natural image data. The method produces results qualitatively similar to those obtained by Independent Component Analysis (ICA) [2].

A recognition architecture that is based on a hierarchical organization of layers of simple and complex cell arrays was introduced by Fukushima [4], called the *Neocognitron*. The network performs invariant recognition of simple visual stimuli like paper clip objects. More recently, Wersing and Körner [11] introduced a variation of the Neocognitron architecture, which learns receptive field profiles using a special type of sparse coding algorithm with invariance constraints to perform robust recognition of natural stimuli, e.g. objects and faces. The authors could show astonishing invariance performance on a variety of disrupted test datasets.

The present work extends the approach of [11] by incorporating recently proposed advances in feature coding strategies:

- Chennubhotla and Jepson [3] showed that sparse coding in some cases fails to extract a "good" representation of the data and suggested to impose the constraint on the feature matrix instead.
- Lee and Seung [7] proposed that a non-negativity constraint should be imposed on the matrices to obtain parts based representations and introduced two algorithms for Non-negative Matrix Factorization (NMF), that can be used to model receptive field learning.
- Hoyer [5] argued, that NMF not always succeeds to extract parts based representations and introduced an enhancement, that allows to explicitly control the amount of sparseness in both the feature matrix and the latent variables, which he calls Non-negative Matrix Factorization with Spareness Constraints (NMFSC).

The major advantage of NMFSC is that former approaches are subsumed (at least qualitatively), therefore, NMFSC allows to investigate the effects of non-negativity, sparseness on the feature matrix and sparseness of the latent variables (or even all of them together) in a common framework. The goal of this contribution is to integrate NMFSC into the Neocognitron-like architecture of [11] and to investigate the effects of combining the afore said mechanisms.

Section 2 describes the hierarchical model, section 3 briefly summarizes the feature coding procedure. Finally, quantitative experimental results are presented in section 4.

2 The Hierarchical Model

Figure 1 shows a diagram of the hierarchical model that is used for the experiments in this paper. It is related to the architectures proposed in [4], [11] and [10]. It has the following properties:

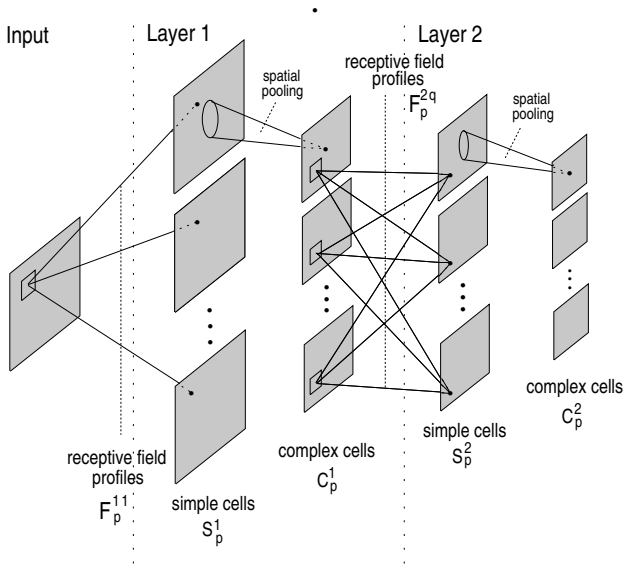

Figure 1: *The Hierarchical Model.* The network consists of two alternating layers of simple and complex cell planes. See text for explanation.

- *Topology:* The model consists of $L = 2$ layers, indexed $l = 1 \dots L$ and each holding P_l planes of two types: *simple cell planes* S_p^l and *complex cell planes* C_p^l with $p = 1 \dots P_l$. The network input is a gray value pixel image. In addition to notational convenience, we set $P_0 = 1$ and refer to the input image as C_1^0. An edge between a complex cell plane C_p^{l-1} and a simple cell plane S_p^l denotes a receptive field profile $F_q^{l\,p}$.

- *Computing simple cell plane activation:* The activation of simple cells in plane S_p^l is computed in two steps: First, we sum up the results of convolution of the activations of the complex cell planes of the previous layer C_q^{l-1} with corresponding receptive field profiles $F_q^{l\,p}$, $q = 1 \dots P_{l-1}$:

$$\hat{S}_p^l = \sum_{q=1}^{P_{l-1}} C_q^{l-1} \otimes F_q^{l\,p}, \tag{1}$$

where $\otimes$ denotes convolution. Note for the simple cells of the first layer the previous layer is simply the input image.

Second, to compute the final (binary) activation of each cell in S_p^l, the "winner takes most" plane-wise competitive mechanism, introduced in [11], is performed among all cells that are located at a position (x, y) in the planes $\hat{S}_p^l, p = 1 \dots P_l$:

$$S_p^l(x,y) = \begin{cases} 0 & \text{if} \quad M = 0 \quad \text{or} \\ & \frac{\hat{S}_p^l(x,y)}{M} < \gamma_l \quad \text{or} \\ & \frac{\hat{S}_p^l(x,y) - \gamma_l M}{1-\gamma_l} < \theta_l, \\ 1 & \text{else,} \end{cases} \tag{2}$$

where $M = \max_p \hat{S}_p^l(x,y)$, $0 < \gamma_l < 1$ is the "competition strength" and θ_l is the "activation threshold" common to all planes in layer l. See [11] for a detailed discussion on this nonlinear step.

- *Computing complex cell plane activation:* The activation of a complex cell plane C_p^l (which is smaller in size than the simple cell planes in the same layer) is directly derived from its corresponding S_p^l plane. The activation of a cell C_p^l at position (x, y) is computed by weighted spatial pooling over a neighborhood of corresponding simple cells:

$$C_p^l(x,y) = \sum_{(x',y') \in H_l(x,y)} G_l(x',y';x,y) * C_p^l(x',y'), \tag{3}$$

where $H_l(x, y)$ is a neighborhood function for layer l, that returns a set of corresponding cell positions in S^l within a square of $\sigma_l \times \sigma_l$. $G_l(x', y'; x, y)$ is a Gaussian with variance σ_l, centered at the C^1 cell position corresponding to (x, y).

Following [11], we choose for the experiments $P_1 = 4$ and use as fixed values for the first layer receptive field profiles, i.e. for $F_1^{1\,p}$, $p = 1 \dots 4$, first-order even Gabor kernels at 0, 45, 90 and 135 degrees.

Using fixed profiles for the first layer is motivated by the fact that efficient coding on natural image patches yields Gabor like receptive fields [2, 9, 5]. The use of Gabor kernels can thus be understood as a feature extraction which is not specialized to a particular domain. Together with the "winner takes most" nonlinearity, the first layer yields a "general" segmentation of the input stimulus based on four dominant edge orientations.

In contrast, the receptive field profiles $F_q^{2\,p}$ used on the second layer are domain specific. They are obtained by an unsupervised feature coding strategy that analyses the typical activation patterns of the C^1 cell planes, while the network is exposed to training images (section 3).

This way, a rejection behavior of 'unknown' parts of a stimulus can be achieved and utilized for recognition in the presence of clutter (see section 4).

3 Feature Coding using NMFSC

To obtain a training set for the feature coding procedure in layer 2, we first apply layer 1 of the network to a set of training images. Sample patches of size $d_{F^2} \times d_{F^2}$ are extracted at random positions from the activation of C^1 cell planes. Concatenating these sample patches yields vectors of dimension $d_{F^2} * d_{F^2} * P_1$. The vectors are used as the columns of a data matrix V which is subsequently decomposed using the NMFSC algorithm proposed in [5].

The algorithm solves the problem $V \approx WH$, where W denotes the feature matrix and H the latent matrix. The inner dimension of WH is set to P_2. The solution is obtained by minimizing the MSE between WH and V under explicit sparseness constraints $0 < W_s < 1$ (the sparseness of columns of W) and $0 < H_s < 1$ (the sparseness of rows of H), and the additional constraints of non-negativity for

matrices W and H. The algorithm also allows to omit W_s or H_s causing the standard learning rules of [7] to be used (refer to [5] for details).

After decomposition, each column p of W is normalized and the values are used to obtain the receptive field profiles $F_q^{2\,p}$, for $p = 1 \ldots P_2$ and $q = 1 \ldots P_1$.

4 Experimental Results

In this section we describe the experimental setup that utilizes the introduced architecture for invariant object recognition by applying a standard Nearest Neighbor Classification scheme on C^2 activations. This allows us to compare the "goodness" of different receptive field profiles in quantitative terms of classification performance. We experiment with two different datasets:

- *Dataset 1*: This dataset contains 3600 images of size 64×64, i.e. 72 views of each of 50 different objects, taken from the first 50 objects of the COIL-100 image library [8]. The dataset is divided into disjoint sets D_{train} (all even numbered views) and D_{test} (all odd numbered views). Additionally, we distort the images in D_{test} by random translation of +/- 5 pixels in x- and y-direction and random scaling of +/- 10%.

- *Dataset 2*: This dataset is similar to Dataset 1, but more difficult, in that the test images are additionally distorted by random background-clutter. Clutter is generated by randomly combining views taken from the remaining 50 objects from the COIL-100 image library [8].

4.1 Optimized vs. Random Profiles

In this experiment we consider two settings: In the first, we generate random P^2 profiles, in the second, we apply the feature coding scheme described in section 3 using D_{test} (which is identical for Dataset 1 and Dataset 2), to obtain P^2 profiles that are optimized to "fit" the image domain. (Note, that the sparseness parameters are omitted here, so standard NMF [7] is applied. The influence of the sparseness parameters will be analyzed in the second experiment in the next section).

Using these two settings, we vary the number of views that are used for training from 1 to 36. For each number, the training images are processed by the network, and for each example the activation of the complete C^2 layer is stored in a database together with the class label information. We then pass the test images from D_{test} (which are different for Dataset 1 and Dataset 2) through the architecture and perform a Nearest Neighbor comparison of the C^2 activations with the database to obtain the classification answers.

The result of this experiment is shown in Fig. 2. Interestingly, for Dataset 1 (Fig. 2, left), the classification rates for optimized profiles exhibit no significant improvement over the random profiles. However, for Dataset 2, a significant improvement can be observed. From this we conclude that the advantage of "tuning" the P^2 profiles to the image domain (here by using standard NMF) is that the overall robustness to distortion by clutter can be improved to a certain extend.

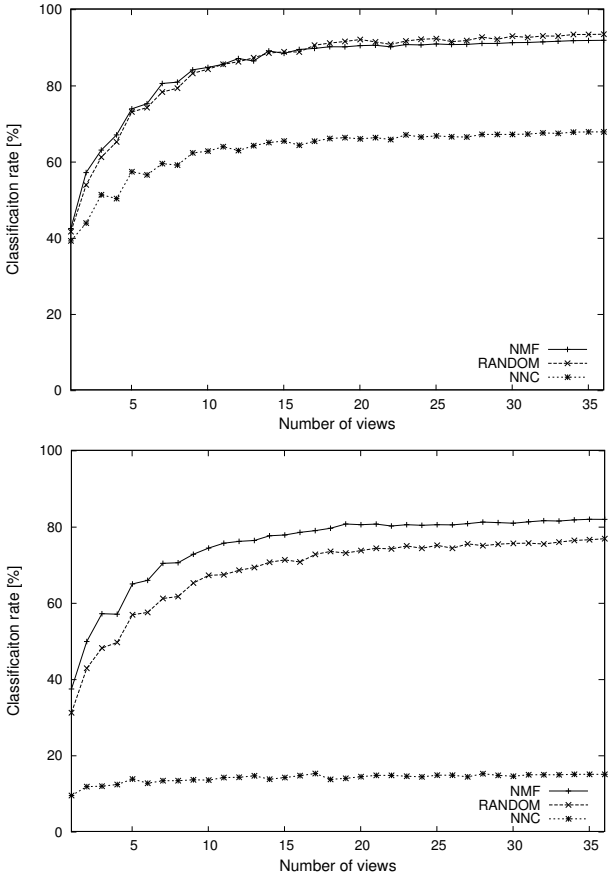

Figure 2: *Top:* Results of experiment 1 on Dataset 1. No significant performance improvement of the optimized profiles over random profiles can be observed. *Bottom:* Results of experiment 1 on Dataset 2. For the cluttered dataset, the optimized profiles exhibit better performance than random profiles. In all cases the values represent the average performance over 20 repeated runs. To provide an impression of the "difficulty" of the dataset, in both plots the NNC-curve denotes the performance of a Nearest Neighbor-Classifier applied to the unprocessed input images.

4.2 Sparseness Constraints

Based on the results from the first experiment, we now analyze whether the classification performance can be further improved by imposing explicit sparseness constraints on either the feature matrix W or on the latent matrix H. This can be done by choosing different values for W_s and H_s, resp. from the interval $[0, 1]$ (refer to section 3 and [5] for details).

Since the results of the first experiment show that saturation of the classification performance starts for approx. 15 training views, we will use this fixed number for the following experiment.

In order to analyze the effect of imposing a sparseness

constraint on the latent variables (like implicitly done on sparse coding and ICA), in a first setting we assign an increasing number between 0.025 and 0.975 to H_s with an increment of 0.025 (the W_s parameter is omitted). The results for 20 repeated runs are shown in Fig. 3, top. The dashed curve shows the average classification performance and the error-bars represent the standard deviation for the current amount of sparseness. For comparison, the horizontal dotted line represents the average performance of the network using standard NMF and 15 views for training (see experiment 1). The result shows, that no stable significant improvement can be achieved for this case.

Figure 3, bottom, shows the results of the same experiment for sparseness constraints on the feature matrix. In this case, the result is more stable and a slight improvement of the classification rate can be observed for W_s values between 0.4 and 0.6.

From the results in this experiment, we conclude that imposing sparseness constraints does yield a slightly better performance in the current application, but that the constraints should be imposed on the feature matrix rather than on the latent variables.

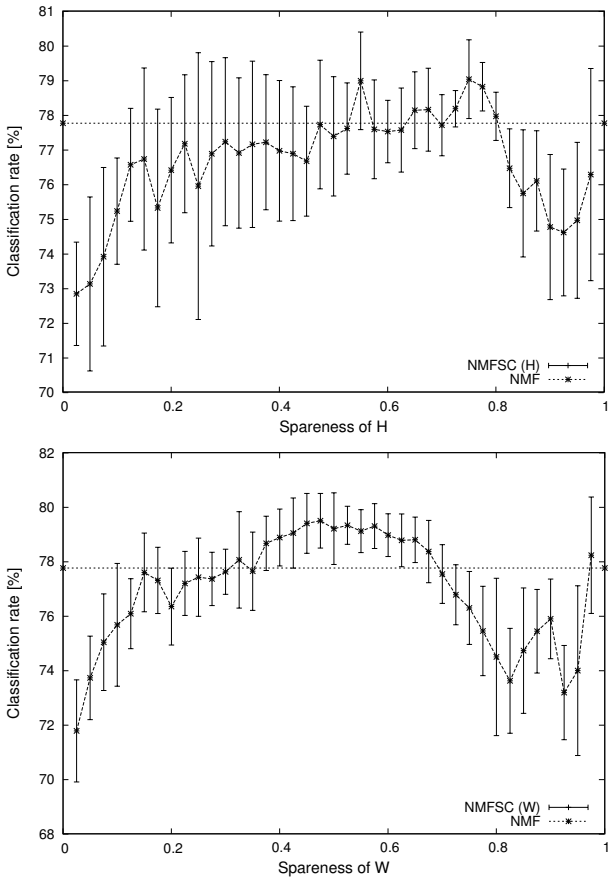

Figure 3: Results of experiment 2: An explicit sparseness constraint is imposed on the latent variable matrix H (top), and on the feature matrix W (bottom). See text.

5 Summary and Conclusion

In this contribution we utilize a recently proposed method for efficient coding called NMFSC for receptive field profile optimization in a hierarchical model of object recognition. We found that profiles, which are optimized using this method exhibit a certain "clutter-rejection" property when compared to random profiles. Moreover, the application allows us to analyze the effect of imposing different sparseness constraints on the feature matrix or on the latent variables in quantitative terms of classification performance. We found that a slightly better classification performance can be achieved by applying the constraints to the feature matrix, but this effect can not be observed for the latent variables.

Acknowledgments

This work was conducted within the scope of the project VAMPIRE (Visual Active Memory Processes and Interactive REtrieval) which is part of the IST program (IST-2001-34401).

References

[1] H. B. Barlow. Possible principles underlying the transformation of sensory messages. *Sensory Communication*, pages 217–234, 1961.

[2] A. J. Bell and T. J. Sejnowski. The independent components of natural images are edge filters. *Vision Research*, 37(27):3327–3338, 1997.

[3] C. Chennubhotla and A. Jepson. Sparse PCA: Extracting Multi-Scale Structure from Data. In *Proc. 8th Int'l Conf. Computer Vision*, 2001.

[4] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. In *Biol. Cybern.*, pages 36:193–202, 1980.

[5] P. O. Hoyer. Non-negative Matrix Factorization with Sparseness Constraints. *Machine Learning Research*, 5(37):1457–1469, 2004.

[6] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurons in the cat's striate cortex. *Journal of Physiology*, 148:574–591, 1959.

[7] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562, 2000.

[8] S. A. Nene, S. K. Nayar, and H. Murase. Columbia Object Image Library: COIL-100. Technical Report CUCS-006-96, Dept. Computer Science, Columbia Univ., 1996.

[9] B. A. Olshausen and D. J. Field. Emergence of simple cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[10] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in visual cortex. *Nature*, 2(11):1019–1025, 1999.

[11] H. Wersing and E. Körner. Learning optimized features for hierarchical models of invariant object recognition. *Neural Comp.*, 15(7):1559–1588, 2003.