Unsupervised Learning of Combination Features for Hierarchical Recognition Models

Heiko Wersing and Edgar Körner

Honda R&D Europe (Deutschland) GmbH
Carl-Legien-Str.30, 63073 Offenbach/Main, Germany
{*heiko.wersing,edgar.koerner*}*@hre-ftr.f.rd.honda.co.jp*

Abstract. We propose a cortically inspired hierarchical feedforward model for recognition and investigate a new method for learning optimal combination-coding cells in intermediate stages of the hierarchical network. The model architecture is characterized by weight-sharing, pooling, and Winner-Take-All nonlinearities. We show that an unsupervised sparse coding learning rule can be used to obtain a recognition architecture that is competitive with other more formally abstracted recognition approaches based on supervised learning. We evaluate the performance on object and face databases.

1 Introduction

The concept of convergent hierarchical coding assumes that sensory processing in the brain is organized in hierarchical stages of neural representations capturing increasingly complex feature combinations [1]. This concept has been criticized as leading to a combinatorial explosion of representative feature combinations under changing object views. Therefore, approaches were formulated to avoid this problem by combining hierarchical feature detection with pooling to achieve gradual invariance of response under transformations of the stimulus [2, 8, 11]. There is substantial experimental evidence in favor of the notion of hierarchical processing in the visual cortex, where an increase in receptive field size and stimulus complexity from initial to later processing stages can be observed [11]. Since rapid feedforward processing in recognition tasks has been shown experimentally [13], Körner et al. [5] proposed a bidirectional model for cortical processing, where an initial hypothesis on the stimulus is facilitated through a feed-forward latency encoding in relation to an oscillatory reference frame.

Methods of supervised feature optimisation which require class information were proposed such as greedy search [8] or gradient-based adaptation [7]. Nevertheless, especially unsupervised learning of features in higher hierarchical stages is still an issue of major interest. Redundancy reduction was proposed as an efficient hierarchical coding strategy [1] and has been applied to model the wavelet-like receptive fields of V1 cells by imposing sparse overcomplete representations [10]. Recently this principle and related concepts of independent component analysis were also applied to models of complex cells and higher-level contour

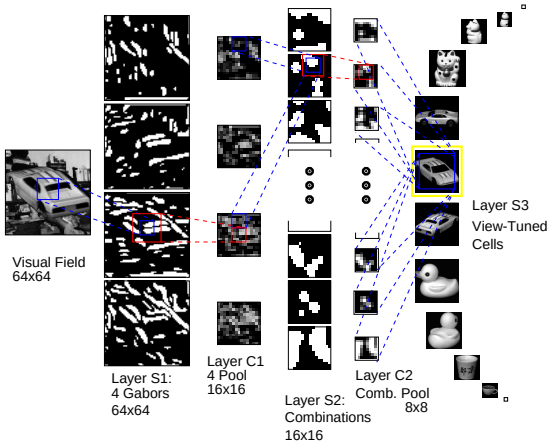

Fig. 1. Sketch of the hierarchical network. The input image is presented as a 64×64 pixel image. The S1 layer consists of 4 Gabor feature planes at 4 orientations with a dimension of 64×64 each. The C1 layer subsamples by pooling down to a resolution of 16×16 for each of the 4 S1 planes. The S2 layer contains combination coding cells with possible local connections to all of the C1 cells. The C2 layer pools the S2 planes down to a resolution of 8×8. The final S3 cells are tuned to particular views, which are represented as the activity pattern of the C2 planes for an input image.

coding combination cells (see [4] and references therein). In this contribution we show that a nonnegative sparse coding learning rule [4] can also be used to obtain optimized combination features in a recognition hierarchy. The resulting recognition architecture, with weights entirely based on unsupervised learning, is competitive with other recognition approaches trained by supervised learning. In Section 2 we describe our model setup with its feedforward nonlinearities and learning methods. The results on different recognition benchmarks are given in Section 3 and discussed in the concluding Section 4.

2 A Hierarchical Model of Invariant Recognition

Architecture. Our hierarchical model is based on a feedforward architecture with weight-sharing and a succession of feature-sensitive and pooling stages (see Fig. 1). The first feature-matching stage consists of an initial linear sign-insensitive receptive field summation, a Winner-Take-All mechanism between features at the same position and a final threshold function. We adopt the notation, that vector indices run over the set of neurons within a particular feature plane of a particular layer. To compute the response $q_1^l(x,y)$ of a simple cell in the first layer S1, responsive to feature type l at position (x,y) , first the image vector $\mathbf{I}$ is multiplied with a weight vector $\mathbf{w}_1^l(x,y)$ (e.g. Gabor filter) characterizing the receptive field profile:

$$q_1^l(x,y) = |\mathbf{w}_1^l(x,y) * \mathbf{I}|. \tag{1}$$

The inner product is denoted by $*$, i.e. for a 10×10 pixel image $\mathbf{I}$ and $\mathbf{w}_1^l(x,y)$ are 100-dimensional vectors. The weights $\mathbf{w}_1^l$ are normalized and characterize a localized receptive field in the visual field input layer. All cells in a feature plane l have the same receptive field structure, given by $\mathbf{w}_1^l(x,y)$, but shifted receptive field centers, like in a classical weight-sharing or convolutional architecture [2, 7]. In a second step a soft Winner-Take-All mechanism is performed with

$$r_1^l(x,y) = \begin{cases} 0 & \text{if } \frac{q_1^l(x,y)}{M} < \gamma_1 \text{ or } M = 0, \\ \frac{q_1^l(x,y) - M\gamma_1}{1-\gamma_1} & \text{else,} \end{cases} \tag{2}$$

where $M = \max_k q_1^k(x,y)$ and $r_1^l(x,y)$ is the response after the WTA mechanism which suppresses sub-maximal responses and provides a model of latency-based competition [5]. The parameter $0 < \gamma_1 < 1$ controls the strength of the competition. The activity is then passed through a simple threshold function with a common threshold θ_1 for all cells in layer S1:

$$s_1^l(x,y) = \mathrm{H}\big(r_1^l(x,y) - \theta_1\big), \tag{3}$$

where $\mathrm{H}(x) = 1$ if $x \geq 0$ and $\mathrm{H}(x) = 0$ else and $s_1^l(x,y)$ is the final activity of the neuron sensitive to feature l at position (x,y) in the S1 layer. The activities of the first layer of pooling C1-cells are given by

$$c_1^l(x,y) = \tanh\big(\mathbf{g}_1(x,y) * \mathbf{s}_1^l\big), \tag{4}$$

where $\mathbf{g}_1(x,y)$ is a normalized Gaussian pooling kernel with width σ_1, identical for all features l, and tanh is the hyperbolic tangent function. The features in the intermediate layer S2 are sensitive to local combinations of the features in the planes of the previous layer, and are thus called *combination cells* in the following. We introduce the layer activation vectors as $\bar{\mathbf{c}}_1 = (\mathbf{c}_1^1, \ldots, \mathbf{c}_1^K)$, $\bar{\mathbf{w}}_2^l = (\mathbf{w}_2^{l1}, \ldots, \mathbf{w}_2^{lK})$ with K=4. Here $\mathbf{w}_2^{lk}(x,y)$ is the receptive field vector of the S2 cell of feature l at position (x,y), describing connections to the plane k of the previous $C1$ cells. The combined linear summation over previous planes is then given by $q_2^l(x,y) = \bar{\mathbf{w}}_2^l(x,y) * \bar{\mathbf{c}}_1$. After the same WTA procedure with strength γ_2 as in (2), the activity in the S2 layer is given by $s_2^l(x,y) = H(r_2^l(x,y) - \theta_2)$ after thresholding with a common threshold θ_2. The step from S2 to C2 is identical to (4) and given by $c_2^l(x,y) = \tanh(\mathbf{g}_2(x,y) * \mathbf{s}_2^l)$, with Gaussian spatial pooling kernel $\mathbf{g}_2(x,y)$ with range σ_2.

Classification of an input image with C2 output $\bar{\mathbf{c}}_2$ is done by nearest neighbor match to previously stored template activations $\bar{\mathbf{c}}_2^v$ for each training view v. This can be realized e.g. by view-tuned units (VTU) [11] in an additional S3 layer with a radial basis function characteristics [11]according to $s_3^v = \exp(-||\bar{\mathbf{w}}_3^v - \mathbf{c}_2||^2)$ where $\bar{\mathbf{w}}_3^v = \mathbf{c}_2^v$ is tuned to the training C2 output of pattern v. Classification can then be performed by detecting the maximally activated VTU.

Parameter adjustment and learning. We adjust the nonlinearity parameters of the visual processing hierarchy in an incremental way. We first choose the processing nonlinearities in the initial layers to provide an optimal output

for a nearest neighbor classifier based on the C1 layer activations. We then keep the initial processing layer fixed and use a sparse coding learning rule to obtain optimized combination features.

The receptive fields of the S1 layer were set as first-order Gabor filters of 4 orientations. To adjust the WTA selectivity γ_1, threshold θ_1, and pooling range σ_1 of the initial layers S1 and C1, we considered a nearest neighbor classification setup of the COIL100 images [9] based on the C1 layer activations. For each of the 100 objects there are 72 views available at subsequent rotations of 5^o. We take four views at angles 0^o, 80^o, 160^o, and 240^o and store the corresponding C1 activation as a template. We can then classify a new test view by finding the template with lowest Euclidean distance in the C1 activation vector. By performing grid-like search over parameters we found an optimal classification performance at $\gamma_1 = 0.9$, $\theta_1 = 0.1$, and $\sigma_1 = 1.5$, giving a correct recognition rate of 72%. This particular parameter setting implies a certain coding strategy: The first layer of simple edge detectors combines a rather low threshold with a strong local competition between orientations. The result is a kind of "segmentation" of the input into one of the four different orientation categories (see also Figure 1). These features are pooled within a range that is comparable to the size of the Gabor S1 receptive fields.

To apply the sparse coding learning rule [4] we generated with the above S1,C1 setting an ensemble of C1 activity vectors for 1000 COIL images and extracted 10000 local 5×5 patches, each with a dimension $5 \times 5 \times 4 = 100$ and indexed by p. The learning rule is defined as the minimization of

$$E = \sum_p ||\bar{\mathbf{c}}_1^{(p)} - \sum_k s_k^{(p)} \bar{\mathbf{w}}_2^k||^2 + \lambda \sum_p \sum_k s_k^{(p)}, \tag{5}$$

jointly in the combination features $\bar{\mathbf{w}}_2^k$ and coefficients $s_k^{(p)}$, subject to the non-negativity of both the components of $\bar{\mathbf{w}}_2^k$ and the $s_k^{(p)}$. The left part of (5) measures the error of reconstructing the input patch $\bar{\mathbf{c}}_1^{(p)}$ from a set of (nonorthogonal) basis features $\bar{\mathbf{w}}_2^k$, while the right part enforces sparse activation of the coefficients $s_k^{(p)}$. After random initialization of the $\bar{\mathbf{w}}_2^k$, the optimization is performed as a two-stage gradient descent process [10, 4]: First the $\bar{\mathbf{w}}_2^k$ are fixed, and a local minimum of (5) is found in $s_k^{(p)}$ for each patch p, using an asynchronous fast fixed-point search as suggested in [14]. In the second step an average gradient step in $\bar{\mathbf{w}}^k$ is performed with $s_k^{(p)}$ set from the first step. Both steps are repeated till convergence. We chose a sparsity factor of $\lambda = 0.1$ and $k = 1, \ldots, 100$ basis features. Based on classification performance on the COIL dataset, nonlinearity parameters were set to $\gamma_2 = 0.0$, $\theta_2 = 1.7$, and $\sigma_2 = 1.0$.

For the setup of an initial quadrature pair nonlinearity without spatial pooling , Hoyer & Hyvärinen [4] obtained only collinear features of different lengths. Contrary to that, our feature set is more diverse and also contains local corner-like and more complex local combinations. This is caused by the different nature of both the input data, man-made objects here compared to texture-rich natural scenes [4], and the different initial processing nonlinearities.

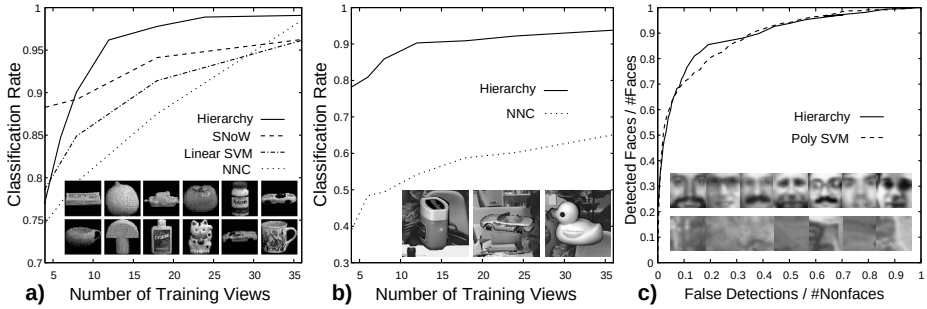

Fig. 2. Comparison of classification rates. a) compares the classification rates on the COIL100 dataset for our hierarchy to results obtained by Roth et al. using their SNoW model, a linear support vector machine, and direct image nearest neighbor classifier (NNC). In a wide regime of sufficient recognition task difficulty (compare NNC), our feature hierarchy achieves best results with high generalization. b) shows performance in a more difficult scenario of 20 objects with 8 pixel-wide position variance and cluttered surround for both training and testing data. The feature hierarchy offers strong robustness compared to NNC. c) shows an ROC plot comparison of face detection performance using a single optimized VTU, but with identical setting for the earlier hierarchical stages. The plot shows the combined rate of correctly identified faces over the rate of misclassifications as a fraction of all non-face images for the hierarchy and a polynomial kernel support vector machine classifier [3].

3 Results

In Figure 2a we compare the classification performance of our model, using C2 activity nearest neighbor matching, to the results published in [12] using the SNoW recognition approach and applying a linear support vector machine on the COIL-100 dataset. To show the application to another classification scenario using the same hierarchy, we used the ORL face image dataset (copyright AT & T Research Labs, Cambridge), which contains 10 images each of 40 people with variability in expression and pose. Without any parameter or feature modification we obtain a classification rate of 96% using 5 training views, compared to 96.5% [6] using gradient-based supervised learning on higher hierarchical stages.

Another central ability for visual recognition is the rejection of unknown stimuli. With an identical setting as described above, however, using a single sigmoidal output VTU with $s_3^1 = \tanh(\bar{\mathbf{w}}_3^1 * \bar{\mathbf{c}}_2)$, we performed gradient-based supervised optimization of $\bar{\mathbf{w}}_3^1$ with target outputs of -0.9 and 0.9 for nonfaces and faces respectively. The training ensemble (data from [3]) consists of 2429 19×19 pixel face images and 4548 non-face images. A threshold criterion was used to decide the presence or non-presence of a face for a different test set of 472 faces and 23573 non-faces. The non-face images consist of a subset of all non-face images that were found to be most difficult to reject for the support vector machine classifier considered by Heisele et al. As is shown in an ROC-plot in Figure 2b, which shows the performance depending on the variation of

the detection threshold, the architecture is competitive with a high performance nonlinear SVM classifier [3].

4 Discussion

We have shown that a sparse coding learning rule allows to derive efficient local combination features in a visual hierarchy in an unsupervised way, offering advantages to supervised optimization of features through greedy search [8] or gradient-based adaptation [7], which require class information. For the databases that we considered here, we could show that the resulting representation can also be applied to face recognition and classification with good results. This generalization across domains is a highly desirable property on the way to more general recognition architectures like the visual cortex.

Acknowledgments: We thank T. Poggio, C. Goerick, J. Eggert, T. Rodemann and U. Körner for stimulating discussions and B. Heisele for providing the face image data. This work was in part supported by BMBF grant 01IB001E (LOKI project).

References

1. H. B. Barlow. The twelfth Bartlett memorial lecture: The role of single neurons in the psychology of perception. *Quart. J. Exp. Psychol.*, 37:121–145, 1985.
2. K. Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1:119–130, 1988.
3. B. Heisele, T. Poggio, and M. Pontil. Face detection in still gray images. Technical report, MIT A.I. Memo 1687, 2000.
4. P. O. Hoyer and A. Hyvärinen. A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*, 2002. to appear.
5. E. Körner, M.-O. Gewaltig, U. Körner, A. Richter, and T. Rodemann. A model of computation in neocortical architecture. *Neural Networks*, 12(7-8):989–1005, 1999.
6. S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back. Face recognition: A convolutional neural-network approach. *IEEE Trans. Neur. Netw.*, 8(1):98–113, 1997.
7. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.
8. B. W. Mel and Jozsef Fiser. Minimizing binding errors using learned conjunctive features. *Neural Computation*, 12(4):731–762, 2000.
9. S. K. Nayar, S. A. Nene, and H. Murase. Real-time 100 object recognition system. In *Proc. of ARPA Image Understanding Workshop*, Palm Springs, 1996.
10. B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1 ? *Vision Research*, 37:3311–3325, 1997.
11. M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999.
12. D. Roth, M.-H. Yang, and N. Ahuja. Learning to recognize objects. In *Proc. of the Conf. on Pattern Recognition and Computer Vision*, 2000.
13. S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the visual system. *Nature*, 381:520–522, 1996.
14. H. Wersing, J. J. Steil, and H. Ritter. A competitive layer model for feature binding and sensory segmentation. *Neural Computation*, 13(2):357–387, 2001.