HIERARCHICAL NEURAL LEARNING FOR OBJECT RECOGNITION

Daniel Oberhoff and Marina Kolesnik

Fraunhofer Institut für angewandte Informatik FIT
Schloss Birlinghoven
53754 St. Augustin
Germany

Marc M. Van Hulle

K.U.Leuven
Laboratorium voor Neuro- en Psychofysiologie
Campus Gasthuisberg, Herestraat 49
BE-3000 Leuven
Belgium

ABSTRACT

We present a neural-based learning system for object recognition in still gray-scale images. The system comprises several hierarchical levels of increasing complexity modeling the feed-forward path of the ventral stream in the visual cortex. The system learns typical shape patterns of objects as these appear in images from experience alone without any prior labeling. Ascending in the hierarchy, spatial information about the exact origin of parts of the stimulus is systematically discarded while the shape-related object identity information is preserved, resulting in strong compression of the original image data. On the highest level of the hierarchy, the decision on the class of an object is taken by a linear classifier depending solely on the object's shape. We train the system and the classifier on a publicly available natural image data set to test the learning capability and the influence of system parameters. The neural system performs respectably when recognizing objects in novel images.

1. INTRODUCTION

Humans easily learn object appearance through active visual perception of their typical shape patterns while for computers this task still poses a problem, which cannot be solved with sufficient robustness and accuracy.

Many object recognition systems have been suggested in the last 20-30 years and an increasing number of them have followed a biologically inspired design. Some of these studies were mainly concerned with modeling human visual perception, but many exhibit performance comparable or superior to non-biological state of the art systems [1, 2, 3, 4, 5].

This work presents a neural system, which reflects the processing in the ventral stream of human visual cortex. The system is built upon a hierarchical architecture of interleaved (S) and (C) layers. Their naming as well as their functionality of these layers is reminiscent of simple and complex/hypercomplex cells identified in the visual cortex by Hubel and Wiesel in the 1960s[6]. S-layers, which are the central functional components of the system, are driven by a stack of simple cells with the capability to learn patterns present in the visual input. The organization of these stacks is comparable to the columnar structure found throughout areas of the ventral stream in the visual cortex [6, 7, 8].

A series of models with a similar architecture of interleaved S- and C- layers has been suggested in the literature. The first such network called "Neocognitron", suggested by Kunihiko Fukushima in 1980 [1], consisted of a series of S- and C-layers with shared weights for a set of local receptive fields and separate inhibitory and excitatory sub-populations of cells. It autonomously forms classes for presented characters and correctly classifies slightly distorted and noisy versions of these characters. In 1989 Yan LeCun et al. [9], presented a similar network, "LeNet", for written character recognition that generated local feature descriptors through back-propagation. A later version of this network has been shown to act as an efficient framework for nonlinear dimensionality reduction of image-sets from real-world objects [2]. The network, however, does not learn autonomously and requires an error signal on the output to perform the back-propagation. The latter is not biologically justified. In 2003 Riesenhuber and Poggio [5] presented a framework for understanding object recognition in the visual cortex with a similar layout. Their work concentrated mostly on the biological plausibility of its architecture and the correspondence of model parts with areas of visual cortex. They introduced Gaussian pattern-tuned units as a model for the simple cells, and a max-function pooling input from a local population of S-layer cells to model functionality of complex cells. Learning in their model is constrained to the tuning of simple cells to random snapshots of local input activity generated by presentations of objects of interest. The model, however, was successfully applied to the modeling of of V4 and IT neuron responses and also as an input stage to a classifier for object and face recognition [10, 5].

Supported by the European Commission (Contract no.:12963 NEST, project MCCOOP)

The approach we present here inherits the adaptive nature of the Neocognitron and some details of the S-layer activation, and the technique of maximum pooling in the C-layers from the model of Riesenhuber and Poggio. The novelty lies in the way that adaptive S-layer units extract essential shape features from the input.

2. NEURAL MECHANISMS OF OBJECT RECOGNITION

The neural system is set up along cortical mechanisms of object recognition, which are thought to be mediated by neurons of the ventral visual pathway [11]. According to a widely accepted consensus based on neurophysiological studies, the ventral pathway begins in V1, where simple cells with small receptive fields respond preferably to oriented bars [6]; neural signals from V1 are then projected onto areas V2 and further on to V4, where neurons show an increase both in receptive field size and the complexity of their preferred stimuli [12]. At the top of the ventral stream are neurons of inferotemporal cortex (IT), which consists of two areas TEO and TE representing form, color and texture of objects. TE is the last exclusively visual area in the ventral stream for object recognition [7].
Progressing up the ventral hierarchy, there is a gradual shift in the properties of neurons and their connectivity:

- receptive field size increases, as does the tolerance to scale and position changes in preferred stimuli;
- receptivity to complex patterns successively increases;
- patterns of cortical projections become successively less topographic;

Extensive neurophysiological evidence shows that inputs to area TE retain no obvious retinotopic organization, which means that single neurons in TE respond to objects anywhere in the visual field, whereas explicit information about the spatial location of an object is not retained at this highest level of the ventral stream. In addition, activation of neurons in TE is largely invariant to shift and scale transformation of their "preferred stimuli".
It is undisputed that the capability to recognize objects is gained through visual experience. Because the onset of visual stimuli on the retina is a 2-dimensional projection of an otherwise 3-dimensional world, the function of the ventral stream is to encode those invariant features of objects that are useful for their recognition across a wide variety of object appearances. Tuning of IT neurons also depends on visual experience. Studies have shown that the majority of IT neurons are view-tuned, showing selectivity to objects in certain orientations [8]. At the same time, these view-tuned neurons exhibit translation invariance of 4 degrees (about twice the stimulus size) and scale invariance of two octaves [5]. It has been concluded that whereas view-invariant recognition requires experiencing novel objects in many views [7], significant position and scale invariance seems to be readily present in view-tuned neurons [5]. Object recognition is possible for rapid visual presentations [13, 14].
Following this experimental evidence as well as a successful example of modeling the feed-forward architecture [5], our neural system is restricted to forward processing of incoming visual input, following the layered organization of visual cortex from V1 to IT in a series of interleaved layers with properties reflecting the typical changes in the properties of corresponding visual neurons. Learning is similarly performed in a bottom-up manner, driven by experience alone, without the use of any prior knowledge, following above mentioned evidence that this kind of learning is prevalent in the corresponding neural areas. Final decision on object recognition in the system is taken by linear classification, following evidence that the encoding of object shapes by the neurons in IT allows for object identification and classification using a simple linear classifier [10].

3. THE LEARNING SYSTEM

3.1. Hierarchical architecture

In a hierarchy of S-/C-layer pairs (Figure 1). The functional role of the S-cells is to form a sparse representation of the experienced input activity while C-cells discard information irrelevant to the recognition process, such as the exact spatial location or size of a stimulus, but preserve information about the retrieved pattern's identity. In the hierarchy each following S-layer receives input from the C-layer of the preceding pair. Through this stepwise process increasingly complex and potentially large input patterns are encoded . The first S-layer (S1) is modeled by a Gabor filter bank to establish orientation selective cells with properties similar to cortical simple cells [15].

3.2. Shift invariant C-layers

The C-layers are implemented by pooling over afferent S-cells from the previous S-layer with the same selectivity but at slightly different positions. For instance, each complex cell in (C1) pools outputs over a neighborhood of the corresponding simple cell (S1) with the same preferred orientation. Thus a stack of C-cells contains the same number of cells as each of the afferent stacks of S-cells. The pooling is done via non-linear MAX operation analogous to [5], where a plausible biological implementation has also been suggested. The MAX-operation picks the strongest input activity from a set of S1-inputs. Similarly, in every following C-layer (C2) through (C4) units relay the maximum activities

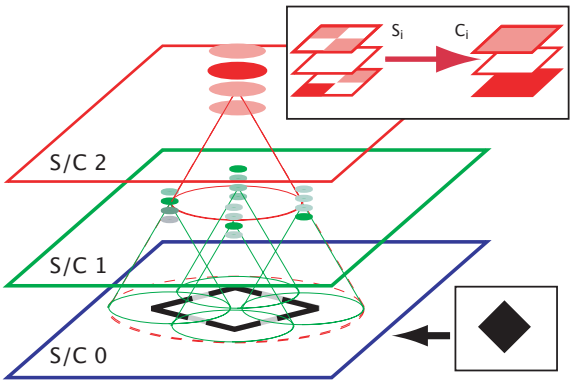

Fig. 1. Schematic illustration of the system functionality: The image of an oblique square is presented and decomposed into oriented contours by the Gabor-filter driven S1-cells. (C1) and following C-layers MAX-pool over local regions (*inset*) to allowing for slight distortions and size changes and reduce the resolution of the representation. Cells in the S2-layer are tuned to local collections of C1 activations, such as the corners of the square. Cells in the S3-layer are tuned to local collections of such features effectively "binding" different parts together. The S3-layer finally contains a unit tuned to the arrangement of the four corner-cells defining the square. This tuning is, due to the C-layers, inherently tolerant to small distortions of the shape. In principle, an arbitrary number of S/C-layers can be stacked in this way limited only by the reduction in resolution.

of a spatial sub-area of the preceding S-layer. Neighboring C-cells draw their inputs from non-overlapping neighborhoods, thus effectively performing sub-sampling.

3.3. Pattern selective S-layers

(S2) through (S4) consist of S-cells with Gaussian tuning to local patterns. Each S-cell possesses a preferred pattern of local activity p_i, the size of which is fixed for each layer, and a tuning bandwidth σ_i determining the sharpness of tuning to the preferred pattern. Only the structure of input patterns is considered and not their actual strength, which is only used for scaling the cell's response after the Gaussian tuning. Each cell's pattern, p_i, is scanned across the whole spatial domain of the input to yield the response of the corresponding S-cell, a_i, at each spatial position:

$$a_i(x,y) = |I_r(x,y)| \times e^{\frac{(\hat{I}_r(x,y)-p_i)^2}{2\sigma_i}} \qquad (1)$$

where $I_r(x,y)$ is a square shaped patch of the input patch with side-length $2r+1$ centered on (x,y). The norm and distance are Euclidean. $\hat{I}_r$ denotes a normalized value of I_r.

1. Select candidate patches for learning.
2. Compare each candidate to all stored patterns. The most similar pattern (i) is selected and an updated tuning width is computed using equations (2).
   IF $\sigma_i' < \sigma_0$ (this pattern is not *new*) *or* no pattern is stored yet
   - the existing pattern and associated tuning bandwidth are updated according to eq. (3).

   OTHERWISE:
   - the pattern is stored as a novel pattern and the associated tuning bandwidth is set to $\sigma_i = \frac{\sigma_0}{2}$.

   σ_0 is updated according to (4) to control the number of patterns.
3. Pairwise distances between stored patterns are computed and if two patterns are more similar than $\sigma_0/2$, the pattern with the smaller tuning bandwidth is removed. Significance measures s_i are updated and patterns with $s_i < s_0$ are removed.

Fig. 2. Learning steps for a layer of pattern selective S-cells

3.4. S-layer learning process

The local patterns p_i are generated and updated in a learning process summarized in figure 2. Candidate patches for learning are generated by starting with a regular grid with grid spacing $2r$, and subsequently shifting to the locus of strongest activity within a distance r. Patches with a norm below the threshold are discarded. Next, the S-cell, i', with the strongest response, evaluated according to eq. (1), is selected within a distance r, and an updated tuning bandwidth is calculated as follows:

$$\sigma_i' = (1-\alpha_{eff})\sigma_i + \alpha_{eff}\,|p_i - I_r(x+\Delta x, y+\Delta y)| \qquad (2)$$

where $\alpha_{eff} = \alpha|I_r(x,y)|$ denotes the effective learning rate which depends on a preset learning rate α and the local input intensity, thereby suppressing the learning of weak background noise patterns. Displacements, $\Delta x, y < r$, in (2) denote the offsets at which the winning S-cell, responds best. If σ_i' exceeds a threshold, σ_0, a new S-cell is generated with a copy of the experienced pattern and $\sigma_0/2$ as an initial tuning bandwidth. Otherwise the old S-cell's preferred pat-

tern and the corresponding tuning bandwidth are updated:

$$\begin{aligned} \sigma_i &\leftarrow \sigma_i' \\ p_i &\leftarrow \frac{\alpha_{eff}\left|\hat{I}_r(x,y) - p_i\right|}{|p_i'|} \end{aligned} \tag{3}$$

where $\leftarrow$ denotes updating. To control the number of stored patterns the threshold σ_0 is updated continuously using:

$$\sigma_0(t+1) = -\frac{\sigma_0(t)}{\tau}\alpha_{eff}\frac{n-n_0}{n_0} \tag{4}$$

where n_0 is the desired number of S-cells and τ is a time constant based on the training set size. Initially, $\sigma_0 = 1$.
To respond to changes in σ_0, and to filter out insignificant patterns two quantities are monitored on a per image basis. The first is the minimum distance, d_i, of the cell's preferred pattern to those of other cells, evaluated within a range of offsets, $\Delta x, y < r$, to also capture shifted duplicates:

$$d_i = \min_{\Delta x, \Delta y, j}\left(\left|p_i|_{\Delta x,\Delta y}^{2r+1,2r+1} - p_j|_{0,0}^{2r+1-\Delta x,2r+1-\Delta y}\right|\right) \tag{5}$$

The second is a running average of the significance s_i expressed as:

$$s_i(t+1) = e^{\frac{-1}{\tau}}\left(\max_{x,y}\left(a_i(x,y)\right) - \bar{a}_i^{max}(t)\right)\Big) \tag{6}$$

Initially $s_i = 1$. Cells are discarded if $d_{ij} < \sigma_0$ or $s_i < s_0$ where s_0 is a fixed significance threshold.
Through this sequence of learning steps S-cells become tuned to frequently occurring patterns, without reliance on any form of *a priori* knowledge and with as little as two iterations over the test set. This is to be compared with the effort needed for the implementation of similar systems as Bayesian probabilistic networks [16] or by employing energy minimization of an objective function to generate the preferred patterns [17]. Generally, the above scheme is similar to the ART family of learning algorithms [18], but simpler and with a more intuitive parametrization. Furthermore we do not know of an implementation of ART in a hierarchical system like this. The complete hierarchy is trained by repeating these steps for the layers (S2) through (S4) making S-cells receptive to input patterns of increasing size and complexity (see also figure 2).

3.5. Object classification

To infer the class of an object presented to the neural system, linear classifiers were trained and evaluated on the output of the C-layers. Training was performed by fitting a multivariate normal distribution to the C-layer outputs for each object class, using the same images used to train the S-layers. Unseen images were then assigned to the class whose learned distribution yielded the highest posterior. Similar techniques have previously been used to infer object identity on the output of a population of IT-neurons [10].

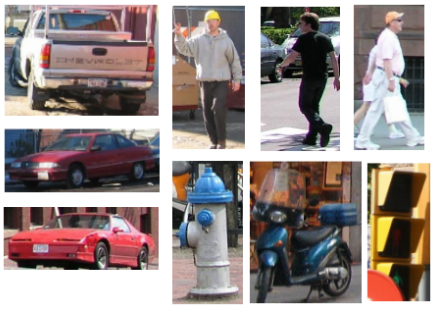

Fig. 3. Examples of images used for training and testing: cars (left), people (top right), and other objects (bottom right)

4. SYSTEM TRAINING AND RECOGNITION RESULTS

For the results presented here the Gabor filters in (S1) where instantiated with spatial frequencies of 0.3 and $0.6 pixels^{-1}$, four different orientations, and a bandwidth of one octave. C-layers units pooled over 5x5 pixel regions (non-overlapping) in (C1) and over 3x3 regions in (C2) through (C4). Receptive field sizes of the S-layers were 5x5 for the layer (S2) and 3x3 for the layers (S3) and (S4). Each layer (S3) through (S4) learned twice more preferred patterns than the previous S-layer, leaving the number of patterns to be learned by (S2) as a parameter. Learning parameters were globally set to $\tau = 100 frames$ and $\alpha = 0.1$.
The pattern selective cells in the layers (S2) through (S4) were tuned by learning in a bottom up manner: first, (S2) was tuned on the output of (C1); next, (S3) was tuned on the output of (C2); and, last, (S4) on the output of (C3). For tuning each layer, all images from the training set were presented three times in random order, so that the layer could converge on a stable state of equilibrium. After the tuning had been completed, learning was turned off and the training set was presented once more to train the classifier. Separate classifiers were trained on the output of each C-layer to monitor the increase of object-specific information through the hierarchy. For this the output from (C1) through (C3) was sub-sampled (using max-pooling) to match the resolution of the last layer (C4). Input vectors to the classifier for each object presentation were the response of a single C-unit stack at the spatial location with the largest sum of activities over the stack.
Real world object views (Figure 3) were obtained by extracting their bounding boxes in images from the publicly

layer	C1	C2	C3	C4
cars	63.3 %	68.7 %	74.7 %	90.3 %
people	46.4 %	57.6 %	76.5 %	84.7 %
other	48.9 %	50.4 %	61.3 %	72.2 %

Table 1. Best recognition rates at each layer of the hierarchy

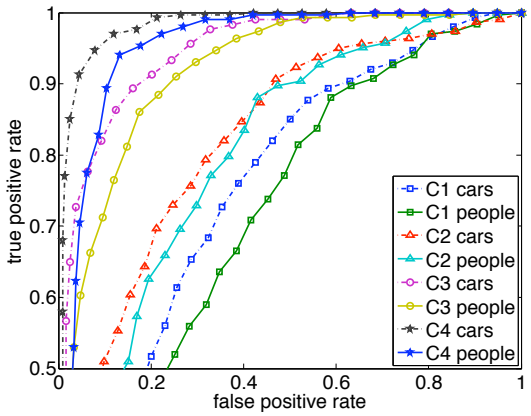

Fig. 4. ROC curves for the layers (C1) through (C3).

available "LabelMe"[1] database. Our target objects for recognition were cars and people. Based on the available labeling data, crops of cars, people and clutter, as negative examples, where extracted from images featuring street situations. The size of the crops was restricted to the range from 80x80 pixels to 320x320 pixels corresponding to two octaves. Smaller crops were discarded whereas larger crops where scaled down linearly. The system had to take a multiple-choice decision signaling whether a car or a pedestrian, or neither of them was present.

Classification performance of separately trained classifiers operating on the output of subsequent hierarchical C-layers was evaluated by means of receiver operator characteristics (ROC) curves, which quantify the trade-off between sensitivity and (1-specificity), shown in Figure 4. Two distinct sets with 900 image examples each (300 per object class plus 300 negative examples) were used for training and testing respectively. The capacity of layer (S2) was set to 30 patterns. The ROC curves show steady gain in recognition performance for an increasingly deep hierarchy with the strongest increase taking place between (C2) and (C3). This can be assigned to the fact that (S3) is the first layer that can encode patterns more complex than simple lines and corners, which are found to be encoded by (S2). Best recognition rates are summarized in Table 1. Recognition rates for cars are highest as these exhibit the least variations in shape. Recognition rates for pedestrians are still comparable with state of the art results for pedestrian recognition (see [4]) even though we only used a simple linear classifier and a relatively small training set.

[1] labelme.csail.mit.edu

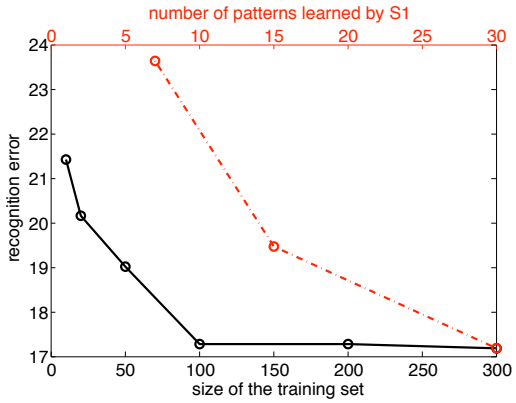

Fig. 5. Recognition error as a function of the training set size (solid line) and the number of learned patterns (dotted line). The number of training images used to train the classifier remained constant.

Figure 5 shows the sensitivity of the recognition performance on the number of training examples used to train the S-layers and on the learning capacity of the S-layers. The quick saturation of system performance at 100 training examples exhibits the strong ability of the system to generalize. The dependence on the network capacity also shows some saturation, but indicates that the system benefits from the increasing number of stored patterns.

5. CONCLUSIONS AND FUTURE WORK

We have presented a neural hierarchical system for shift invariant object recognition employing unsupervised learning to generate features to classify object shapes. The architecture is largely inspired by findings on the ventral stream in visual cortex.

The system is mostly self tuned with only a few remaining parameters, each one having a direct intuitive meaning: storage capacity (n_0), temporal memory (τ and $\bar{a}_0$), and learning rate (α). Additionally, the number of S/C-layer pairs is also a system parameter, since it is not strictly fixed. Our results indicate, that four S/C-pairs are sufficient for the complexity of the recognition problem. Biologically one can associate the S1/C1-pair with area V1, S2/C2 with V2, S3/C3 with V4, and S4/C4 with IT.

Although our neural system feeds on pretty generalized shape descriptors, it shows respectable recognition performance of 91 % for cars and 85 % for people despite of their variable shape appearance. The performance is also comparable with classification results on readouts of IT-neurons on

macaque monkeys (see [10], for comparison: our best overall classification performance was 82.4%, cf. figure 4), and state of the art results given in [4]. The latter is remarkable considering we use a very simple linear classifier.

The presented neural system is inherently general because it uses unsupervised learning with no *a priori* information about the nature of objects it is trained upon. The layered design of the system allows its further extension by a virtually unlimited amount of additional connections and by additional processing streams. For instance, incorporation of feedback projections and lateral interactions could act as a temporal and spatial stabilization mechanism where an input configuration would activate a recognition of some previously experienced configuration triggering the enhancement of those features that support it [19, 20], which could also act as a precursor for attention to a particular shape configuration, possibly enhancing classification or identification and figure-ground-separation. Furthermore some mechanism to automatically determine the necessary depth and capacity of the system would probably be beneficial.

6. REFERENCES

[1] Kunihiko Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cyb.*, vol. V36, no. 4, pp. 193–202, Apr. 1980.

[2] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. CVPR*. 2006, IEEE Press.

[3] Bernd Heisele, Thomas Serre, Massimiliano Pontil, and Tomaso Poggio, "Component-based face detection.," in *Proc. CVPR*, 2001, pp. 657–662.

[4] S. Munder and D. M. Gavrila, "An experimental study on pedestrian classification," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1863–1868, 2006.

[5] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, and T. Poggio G. Kreiman, "A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex," CBCL Memo 259, 2005.

[6] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular ineraction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, pp. 106–154, 1962.

[7] R. Desimone and C. G. Gross, "Visual areas in the temporal cortex of the macaque," *Brain Research*, vol. 178, pp. 363–380, 1979.

[8] N. K. Logothesis and J. Pauls, "Psychophysical and physiological evidence for viewer-centered object representation in the primate," *Cerebral Cortex*, 1995.

[9] Y. LeCun, L. D. Jackel, B. Boser, J. S. Denker, H. P. Graf, I. Guyon, D. Henderson, R. E. Howard, and W. Hubbard, "Handwritten digit recognition: Applications of neural net chips and automatic learning," *IEEE Comm.*, pp. 41–46, November 1989, invited paper.

[10] Chou P. Hung, Gabriel Kreiman, Tomaso Poggio, and James J. DiCarlo, "Fast readout of object identity from macaque inferior temporal cortex," *Science Reports*, 2005.

[11] L.G. Ungerleider and J.V. Haxby, ""what" and "where" in the human brain," *Curr. Opin. Neurobiol.*, vol. 4, pp. 157–165, 1994.

[12] E. Kobatake and K. Tanaka, "Neuronal selectivities to complex object features in the ventral pathway of the macaque cerebral cortex," *J. Neurophysiol*, vol. 71, pp. 856–867, 1994.

[13] Simon Thorpe, Denis Fize, and Catherine Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, no. 6582, pp. 520–522, June 1996.

[14] MC Potter, "Meaning in visual search," *Science*, vol. 187, no. 4180, pp. 965–966, Mar. 1975.

[15] J.G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Amer. A*, vol. 2, pp. 1160–1169, 1985.

[16] Thomas Dean, "Scalable inference in hierarchical generative models," in *Proc. 9th Int. Symp. Art. Int. & Math.*, 2006.

[17] Heiko Wersing and Edgar Korner, "Learning Optimized Features for Hierarchical Models of Invariant Object Recognition," *Neural Comp.*, vol. 15, no. 7, pp. 1559–1588, 2003.

[18] G. A. Carpenter and S. Grossberg, *The Handbook of Brain Theory and Neural Networks, Second Edition*, chapter Adaptive resonance theory, pp. 87–90, MIT Press, 2003.

[19] Stephen Grossberg, "How does the cerebral cortex work? development, learning, attention and 3d vision by laminar circuits of visual cortex," Tech. Rep. 005, Boston University, Boston, MA, 2003.

[20] S. W. Zucker, *Problems in Systems Neuroscience*, chapter "Which Computation Runs in Visual Cortical Columns?", Oxford University Press, 2001.