Visualization of higher-level receptive fields

Christian Hinze[1], Niko Wilbert[1,2], Laurenz Wiskott[1,2]
1 Institute for Theoretical Biology, Humboldt Universität zu Berlin, Berlin, Germany
2 Bernstein Center for Computational Neuroscience, Humboldt Universität zu Berlin, Berlin, Germany

Theory and experimental background

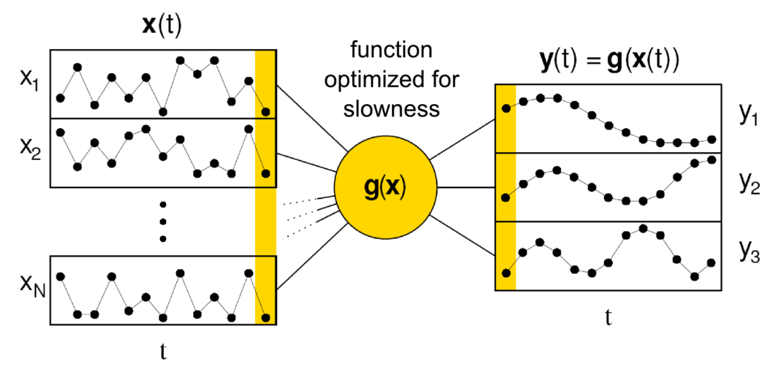

Fig. 1: The principle idea of SFA is extracting slowly varying information from fast varying input signals. The outputs are ordered by their invariance.

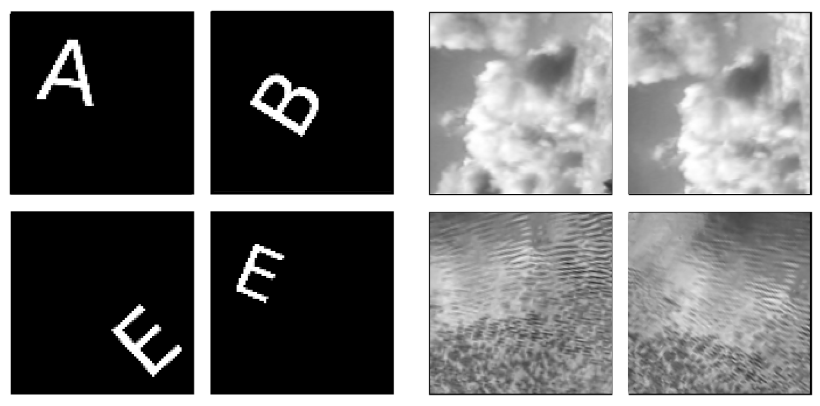

Fig. 2: Examples for the training data of the two networks.

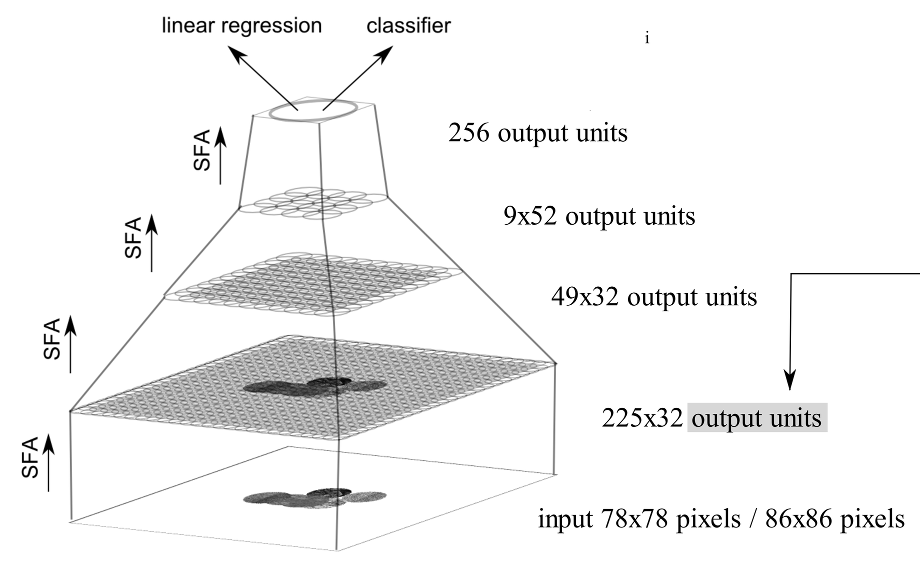

Fig. 3: Scheme of the hierarchical network used in the experiment.

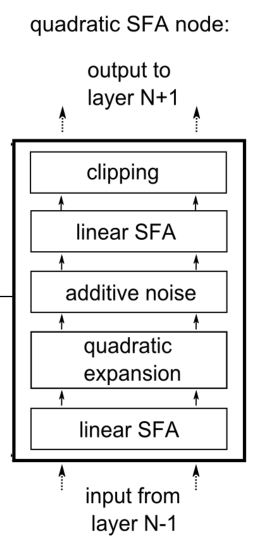

Fig. 4: Output unit design used in every layer.

Optimal stimuli

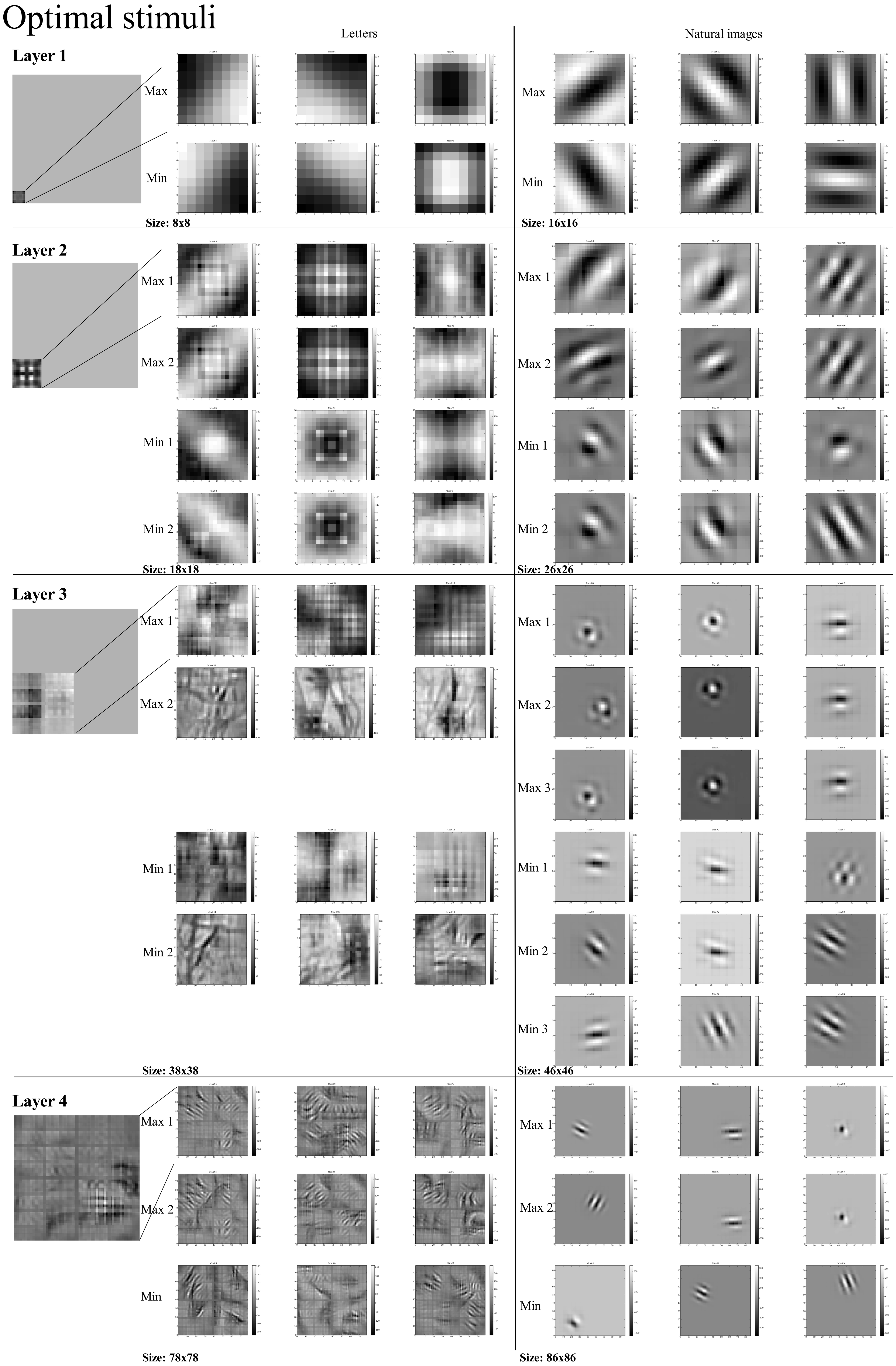

Fig. 5: Optimal stimuli (for a fixed vector norm) for selected units in different layers ("Letters" indicates the results for the letter network and "Natural Images" indicates the results for the network trained with sequences of natural images). The left picture shows the size of one receptive field compared to the input image. The rows on the right side show the maximal and the corresponding minimal stimuli for some units of the indicated layer. The starting point of the algorithm was a grey or natural image of the desired norm. "Size" specifies the receptive field size in each layer.

SFA (Slow Feature Analysis) is a method for learning slowly varying features from a (fast varying) input signal [1]. The different learned output functions are the optimal solutions (on a given set of functions) yielding the most slowly varying output signals on the training data (see fig. 1). In our case we took the polynomials of degree 2 in the input variables as a function set.

We used two differently trained versions of a 4-layer **hierarchical network** as a simple model for the visual system (see fig. 3). In this network SFA is performed hierarchically. The first version was trained with sequences of letters A-E in different positions, scales and rotations. The second network was trained with sequences of natural images (see fig. 2). After the training phase the letter network was able to distiguish between the learned objects and even to differentiate between new letters.

The principle architecture of **each output unit** of the network is shown in fig. 4. The first step is a linear SFA step in the letter network/a PCA (Principal Component Analysis) step in the natural images network followed by an optional addition of noise of very small amplitude and a consecutive linear SFA step on the quadratically expanded signal in both versions of the network. To avoid large output values a clipping is applied as the last step only in the letter network. The receptive fields of the output units overlap and become larger with increasing layer number. The highest layer's receptive field covers the whole input image. The receptive fields slightly differ in the two versions of the network as indicated in the figures (see fig. 5).

Receptive fields in higher areas of the visual cortex are very difficult to characterize, because it is not clear what they are tuned to and which stimuli to use to study them. Moreover the cells in those areas are highly non-linear. A unit in layer 1, 2, 3 or 4 of our network corresponds to a polynomials of degree 2, 4, 8 or 16, respectively, in the input variables.

It was shown that the units after one SFA step already have a wide range of complex cell properties, including optimal stimuli, phase and shift invariance, orientation and frequency selectivity. Most of their optimal stimuli look like Gabor patches [2].
These properties correspond very well with results from neurophysiology for cells in V1 and suggested that units of intermediate layers in our network might share properties with cells in V2 or V4.

This hope is underlined by the fact that units of higher layers are representing more complex information. They learn invariant object representations much like in IT [3]. Units in layer 4 of the letter network represent features like position, rotation and identity of the shown object.

To investigate the receptive fields of units in higher layers we performed a gradient descent/ascent in each layer in both versions of the network. We were looking for the optimal stimuli of the units on a sphere as a natural constraint. The radius was chosen to be the average norm of the training data. The algorithm started in a grey or natural image of the desired norm The results of our simulation for all layers and both versions of the network are shown in fig. 5.

References:

1. Wiskott, L. and Sejnowski, T.J.: Slow feature analysis: Unsupervised learning of invariances. Neural Computation 2002, 14(4):715-770.

2. Berkes, P. and Wiskott, L.: Slow feature analysis yields a rich repertoire of complex cell properties. Journal of Vision 2005, 5 (6):579-602,

3. Franzius, M., Wilbert, N. and Wiskott, L.: Invariant Object Recognition with Slow Feature Analysis. Proc. 18th Int'l Conf. on Artificial Neural Networks 2008, ICANN'08, Prague, September 3-6, eds. Vera Kurková, RomanNeruda and Jan Koutník, publ. Springer-Verlag, pp. 961-970.