# Simple-Cell-Like Receptive Fields Maximize Temporal Coherence in Natural Video

Jarmo Hurri and Aapo Hyvärinen
Neural Networks Research Centre
Helsinki University of Technology
P.O.Box 9800, 02015 HUT, Finland
{*jarmo.hurri,aapo.hyvarinen*}*@hut.fi*

August 26, 2002

**Abstract**

Recently, statistical models of natural images have shown emergence of several properties of the visual cortex. Most models have considered the non-Gaussian properties of static image patches, leading to sparse coding or independent component analysis. Here we consider the basic time dependencies of image sequences instead of their non-Gaussianity. We show that simple cell type receptive fields emerge when temporal response strength correlation is maximized for natural image sequences. Thus, temporal response strength correlation, which is a nonlinear measure of temporal coherence, provides an alternative to sparseness in modeling simple cell receptive field properties. Our results also suggest an interpretation of simple cells in terms of invariant coding principles, which have previously been used to explain complex cell receptive fields.

## 1 Introduction

The functional role of simple cells has puzzled scientists since the structure of their receptive fields was first mapped by Hubel and Wiesel in the 1950s (see, e.g., (Palmer, 1999)). The first hypothesis concerning their role was based on their visual appearance, that is, their similarity with edges and bars. The second major theory was the local spatial frequency analysis theory, which is based on generally applicable signal processing principles. The current view of the functionality of sensory neural networks emphasizes learning and the relationship between the structure of the cells and the statistical properties of the information they process (see, e.g., (Field, 1994; Simoncelli and Olshausen, 2001)). In 1996 a major advance was achieved when Olshausen and Field showed that simple-cell-like receptive fields emerge when sparse coding is applied to natural image data (Olshausen and Field, 1996). Similar results were obtained with independent component analysis (ICA) shortly thereafter (Bell and Sejnowski, 1997; van Hateren and van der Schaaf, 1998). In the case of image data, independent component analysis is closely related to sparse coding (Hyvärinen et al., 2001, pages 396–401) (Olshausen and Field, 1997).

In this paper we show that an alternative principle called *temporal coherence* (Becker, 1993; Földiák, 1991; Kayser et al., 2001; Mitchison, 1991; Stone, 1996; Wiskott and Sejnowski, 2002) leads to the emergence of simple cell receptive fields from natural image sequences. This finding is significant because it means that temporal coherence provides a complementary theory to sparse coding as a computational principle behind the formation of simple cell receptive fields. The results also link the theory of achieving invariance by temporal coherence (Földiák, 1991) to real-world visual data, and measured properties of the visual system. Whereas previous research has focused on establishing this link for complex cells, we show that such a connection exists even on the simple cell level.

Temporal coherence is based on the idea that when processing temporal input, the representation changes as little as possible over time. Földiák was one of the first authors to suggest the usefulness of temporal coherence in computational neuroscience (Földiák, 1991). He developed a two-layer network which was able to learn to identify a fixed feature, such as a line with a fixed orientation, even if the way the feature was expressed in the data changed, for example, if the line was translated. Földiák used temporal coherence as a tool to learn translation invariances: artificially generated input data was temporally coherent (consecutive input frames contained translated versions of a line with the same orientation), and by using competition and short-term memory, the output was also taught to be temporally coherent. This associated translated versions of a feature with each other.

Since Földiák several researchers have studied temporal coherence, and also other forms of coherence, such as coherence with respect to different views of the same scene. For example, in (Becker and Hinton, 1992; Stone, 1996), surface depth was discovered from stereograms by a multiple layer nonlinear network. In (Becker and Hinton, 1992), this was done by using stereograms representing different views of the same randomly generated scene, and maximizing mutual information between outputs. In (Stone, 1996), learning was achieved by using a temporal sequence of slightly different stereograms, and maximizing temporal smoothness of output while preserving variability in output. In these studies, the input data sets were generated so that there was an underlying, coherent parameter in the data, and the objective was to find that parameter by using coherence. Therefore the main result was the demonstration of the usefulness of coherence using simulated data.

The contribution of this paper is to show that when the input consists of *natural* image sequences, the linear filters that maximize *temporal response strength correlation* are similar to *simple cell receptive fields*. In what follows we first describe temporal response strength correlation, which is a measure of temporal coherence, and an algorithm which is capable of optimizing the measure. In Section 3 we apply this algorithm to natural image sequences. In addition to the main results we describe several control experiments, which were made to ensure the validity and novelty of our results. Finally, in Section 4 we give an intuitive explanation of why optimization of the objective function produces such results, discuss the spatiotemporal and nonlinear (non-negative) extensions of the model, and conclude the paper by discussing the implications of this work.

## 2 Temporal response strength correlation

In the basic model we restrict ourselves to consider linear spatial models of simple cells. Linear simple cell models are commonly used in studies concerning the connections between visual input statistics and simple cell receptive fields (Bell and Sejnowski, 1997; Olshausen and Field, 1996; van Hateren and van der Schaaf, 1998), because linearity seems to approximately characterize most simple cells (DeAngelis et al., 1993b). Extensions of this basic framework will be discussed in Sections 4.4 and 4.5.

The model uses a set of spatial filters (vectors) $\mathbf{w}_1, ..., \mathbf{w}_K$ to relate input to output. Let signal vector $\mathbf{x}(t)$ denote the input of the system at time $t$. A vectorization of image patches can be done by scanning images column-wise into vectors – for windows of size $N \times N$ this yields vectors with dimension $N^2$. The output of the $k$th filter at time $t$, denoted by signal $y_k(t)$, is given by the dot-product

$$y_k(t) = \mathbf{w}_k^T \mathbf{x}(t). \qquad (1)$$

Let matrix $\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_K]^T$ denote a matrix with all the filters as rows. Then the input-output relationship can be expressed in vector form by

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t), \qquad (2)$$

where signal vector $\mathbf{y}(t) = [y_1(t) \cdots y_K(t)]^T$.

Temporal response strength correlation, the objective function, is defined by

$$f(\mathbf{W}) = \sum_{k=1}^{K} \mathrm{E}_t \left\{ g(y_k(t)) g(y_k(t - \Delta t)) \right\}, \qquad (3)$$

where the nonlinearity $g$ is strictly convex, even (rectifying), and differentiable. The symbol $\Delta t$ denotes a delay in time. The nonlinearity $g$ measures the strength (amplitude) of the response of the filter, and emphasizes large responses over small ones (see Section 4). Examples of choices for this nonlinearity are $g_1(x) = x^2$, which measures the energy of the response, and $g_2(x) = \ln \cosh x$, which is a robustified version of $g_1$. A set of filters which has a large temporal response strength correlation is such that the same filters *often respond strongly at consecutive time points*, outputting large (either positive or negative) values. This means that the same filters will respond strongly over short periods of time, thereby expressing temporal coherence of a population code.

To keep the outputs of the filters bounded we enforce the unit variance constraint on each of the output signals $y_k(t)$, that is, we enforce the constraint $\mathrm{E}_t \left\{ y_k^2(t) \right\} = \mathbf{w}_k^T \mathbf{C_x} \mathbf{w}_k = 1$ for all $k$, where matrix $\mathbf{C_x} = \mathrm{E}_t \left\{ \mathbf{x}(t)\mathbf{x}^T(t) \right\}$. Additional constraints are needed to keep the filters from converging to the same solution. Standard methods (Hyvärinen et al., 2001) are either to force the set of filters to be orthogonal, or to force their outputs to be uncorrelated, from which we choose the latter. This introduces additional constraints $\mathbf{w}_i^T \mathbf{C_x} \mathbf{w}_j = 0$, $i = 1, ..., K$, $j = 1, ..., K$, $j \neq i$. These uncorrelatedness constraints limit the number of filters $K$ we can find so that $K \leq N^2$. The unit variance constraints and the uncorrelatedness constraints can be expressed by the single matrix equation

$$\mathbf{W} \mathbf{C_x} \mathbf{W}^T = \mathbf{I}. \qquad (4)$$

Note that if we use a nonlinearity $g(x) = x^2$, and $\Delta t = 0$, the objective function becomes $f(\mathbf{W}) = \sum_{k=1}^{K} \mathrm{E}_t \left\{ y_k^4(t) \right\}$. In this case the optimization of the objective function under the unit variance constraint is equivalent to optimizing the sum of kurtoses of the outputs. Kurtosis is a commonly used measure in sparse coding. Similarly, in the case of nonlinearity $g(x) = \ln \cosh x$ and $\Delta t = 0$, the objective function can be interpreted as a non-quadratic measure of the non-Gaussianity of filter outputs. We will return to this issue in Section 3.

Thus, the receptive fields are learned in our model by maximizing the objective function (3) under the constraint (4). The optimization algorithm used for this constrained optimization problem is a variant of the gradient projection method of Rosen (for the original algorithm see e.g. (Luenberger, 1969, pages 297–299)). The optimization approach employs whitening, that is, a temporary change of coordinates, to transform the constraint (4) into an orthonormality constraint. Then a gradient projection algorithm employing optimal symmetric orthogonalization can be used. See Appendix A for details.

# 3 Experiments on natural image sequences

## 3.1 Data collection

The natural image sequences used as data were a subset of those used in (van Hateren and Ruderman, 1998). The original data set consisted of 216 monochrome, non-calibrated video clips of 192 seconds each, taken from television broadcasts. More than half of the videos feature wildlife, the rest show various topics such as sports and movies. Sampling frequency was 25 frames per second, and each frame was block-averaged to a resolution of $128 \times 128$ pixels. For our experiments this data set was pruned to remove the effect of human-made objects and artifacts. First, many of the videos feature human-made objects, such as houses, furniture etc. Such videos were removed from the data set, leaving us with 129 videos. Some of these 129 videos had been grabbed from television broadcasts so that there was a wide black bar with height 15 pixels at the top of each image, probably because the original broadcast had been in wide screen format. Our sampling procedure never took samples from this topmost part of the videos. If these artifacts were not removed from the data set, the static ICA results computed for comparison showed longer horizontal and vertical receptive fields than results obtained from scenes without the artifacts. The final, preprocessed (see below) data set consisted of 200,000 pairs of consecutive $11 \times 11$ image windows (patches) at the same spatial position, but $\Delta t$ milliseconds apart from each other. Depending on the experiment, $\Delta t$ varied between 40 ms and 960 ms. However, because of the temporal filtering used in preprocessing, initially 200,000 longer image sequences with a duration of $\Delta t + 400$ ms, and the same spatial size $11 \times 11$, were sampled with the same sampling rate.

A second data set was needed for computing the corresponding (static) ICA solution for comparison. This data set consisted of 200,000 $11 \times 11$ images sampled from the same video data.

## 3.2 Preprocessing

The preprocessing in the main experiment consisted of three steps: temporal decorrelation, subtraction of local mean, and normalization. (The same preprocessing steps were applied in the control experiments. Whenever preprocessing was varied in control experiments it is explained separately below.) Temporal decorrelation can be motivated in two different ways. First, it can be motivated biologically as a model of temporal processing at the lateral geniculate nucleus (Dong and Atick, 1995). Second, as discussed above, for $\Delta t = 0$ the objective function can be interpreted as a measure of sparseness. Therefore it is important to rule out the possibility that there is hardly any change in short intervals in video data, since this would imply that our results could be explained in terms of sparse coding or ICA. To make the distinction between temporal response strength correlation and measures of sparseness clear, temporal decorrelation was applied because it enhances temporal changes. Note, however, that this still does not remove all of the static part in the video – this issue is addressed in the control experiments below.

To investigate the effect of temporal decorrelation, we first examined the histogram of the distances between subsequent image windows separated by $\Delta t = 40\,\text{ms}$, shown in Figure 1A. (This histogram also shows that there are indeed large changes between subsequent time points even without temporal decorrelation.) The local mean has been removed from these windows, and they have been normalized (see below); note that 2 is maximal distance because of normalization. Temporal decorrelation was performed with a temporal filter, shown in Figure 1B. The Fourier magnitude of the filter was computed by inverting the amplitude spectrum of input data. A Wiener filter approach was used in order to not amplify high frequency noise. (We determined the filter directly from data, see (Dong and Atick, 1995) for an analytic solution.) Noise power was estimated by assuming that it is equal to signal power at 5.5 Hz, a value also used in (Dong and Atick, 1995). Phases were determined by adding a minimum energy delay constraint on the filter (see (Oppenheim and Schafer, 1975, pages 504–511)). Filter length was originally 2500 ms, but was truncated to 400 ms because this part contains over 99% of filter energy. When short image sequences of length $\Delta t + 400\,\text{ms}$ are filtered with this temporal filter, the resulting sequences are of length $\Delta t$. From each of these temporally filtered short sequences, two windows separated by $\Delta t$ were taken. Figure 1C shows the distribution of the distances between such temporally decorrelated windows, after removal of local mean and normalization (see below), when $\Delta t = 40\,\text{ms}$. This histogram shows that temporal decorrelation enhances temporal changes in the data. Figure 1D shows the distribution of the distances between temporally decorrelated windows, after removal of local mean and normalization, when $\Delta t = 120\,\text{ms}$. In this case, consecutive windows are, on the average, quite far from each other when measured with the Euclidean norm. In fact, the peak of the histogram is approximately at 1.4, which for normalized vectors means that a large part of the consecutive windows are approximately orthogonal to each other.

After temporal decorrelation, the local mean (DC component) was subtracted from each window. This reduces the number of dimensions of the data by one, so with image window size $N = 11$, the number of filters $K \leq 120$. Finally, the sample vectors (vectorized windows) were normalized to have unit

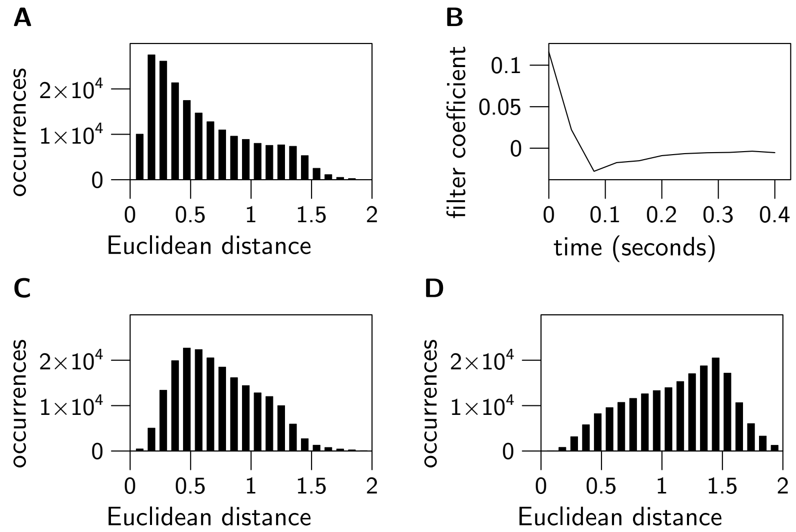


Figure 1: Temporal decorrelation enhances temporal changes. (A) Distribution of Euclidean distances between consecutive samples at the same spatial position, but 40 ms apart from each other, without temporal decorrelation. Note that two is the maximum value because of normalization. (B) The temporally decorrelating filter. (C) Distribution of Euclidean distances between consecutive samples at the same spatial position, but 40 ms apart from each other, *after* temporal decorrelation. (D) Distribution of Euclidean distances between consecutive samples at the same spatial position, but 120 ms *apart* from each other, after temporal decorrelation. Note that removal of DC component and normalization were also performed in all of these cases.

Euclidean norm, which can be considered as a form of contrast gain control (Carandini et al., 1997; Heeger, 1992). Note that no spatial low-pass filtering or dimensionality reduction was performed during preprocessing.

In the case of the second data set, which was used to compute the corresponding static ICA solution, preprocessing consisted of removal of the local mean, followed by normalization. Temporal decorrelation was not performed here, since it has no meaning in the case of static image data.

### 3.3 Results: temporally coherent filters of natural image sequences

The main experiment consisted of running the symmetric gradient projection algorithm 50 times, using different random initial values, followed by a quantitative analysis of these results, as well as results obtained with a corresponding ICA algorithm. In this experiment $\Delta t$ was 40 ms. The number of extracted filters was set at the maximum value $K = 120$. Nonlinearity $g$ in objective function (3) was chosen to be $g(x) = \ln \cosh x$ because of its robustness against outliers (Hyvärinen et al., 2001).

Figure 2 shows the resulting filters (i.e., rows of matrix $\mathbf{W}$) of the first run. The filters have been ordered according to $\mathrm{E}_t \{g(y_k(t))g(y_k(t - \Delta t))\}$, that is, according to their "contribution" into the final objective value (filters with largest values top left). The filters resemble Gabor filters. They are localized, oriented, and have different scales. These are the main features of simple-cell receptive fields (Palmer, 1999).

When compared qualitatively with earlier reported results, obtained with sparse coding and independent component analysis (Bell and Sejnowski, 1997; Olshausen and Field, 1996; van Hateren and van der Schaaf, 1998), our results show a larger variety of different spatial scales. To compare the results quantitatively, we extracted a corresponding set of 50 ICA separation matrices using the symmetric fixed-point ICA (FastICA) algorithm, with robust nonlinearity tanh (Hyvärinen et al., 2001, pages 183–185 and 196). This algorithm is a symmetric version of that used in (van Hateren and van der Schaaf, 1998) – the symmetric nature of the algorithm facilitates extracting a balanced set of filters (Hyvärinen et al., 2001, pages 194–195). Filters maximizing temporal response strength correlation were compared against the rows of the separating matrix (the ICA filters), since these filters are the natural counterparts in ICA (van Hateren and van der Schaaf, 1998). Figure 3 shows the ICA filters obtained from the first run.

In the quantitative comparison of the results, we measured the most important properties of the receptive fields. The results are shown in Figure 4. The measured properties were peak spatial frequency (Figures 4A and 4B, note logarithmic scale, units cycles/pixel), peak orientation (Figures 4C and 4D), spatial frequency bandwidth (Figures 4E and 4F), and orientation bandwidth (Figures 4G and 4H). See (van Hateren and van der Schaaf, 1998) for definitions of these measures. Although there are some differences, the most important observation here is the similarity of the histograms. This supports the idea that ICA / sparse coding and temporal coherence are complementary theories, in that they both result in the emergence of simple-cell-like receptive fields. As for the differences, the results obtained using temporal response strength correlation have a slightly smaller number of high-frequency receptive fields. Also,

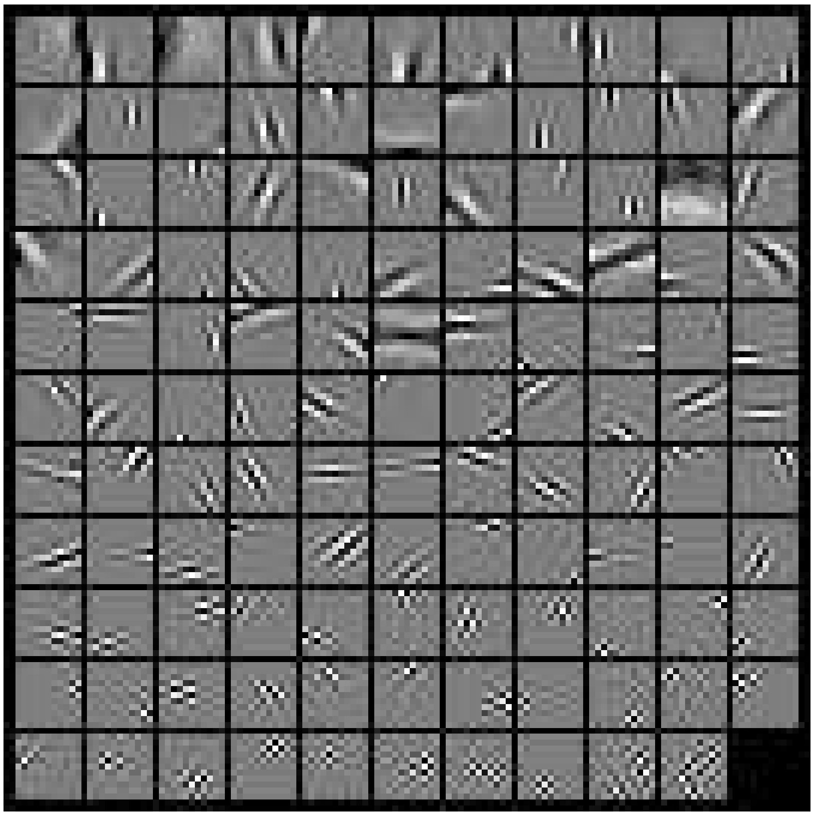

Figure 2: Temporally coherent filters of natural image sequences, given by the first run of the main experiment. The filters were estimated from natural image sequences by optimizing temporal response strength correlation with the symmetric gradient projection algorithm (here nonlinearity $g(x) = \ln \cosh x$). The filters have been ordered according to $\mathrm{E}_t \{g(y_k(t))g(y_k(t - \Delta t))\}$, that is, according to their "contribution" into the final objective value (filters with largest values top left).

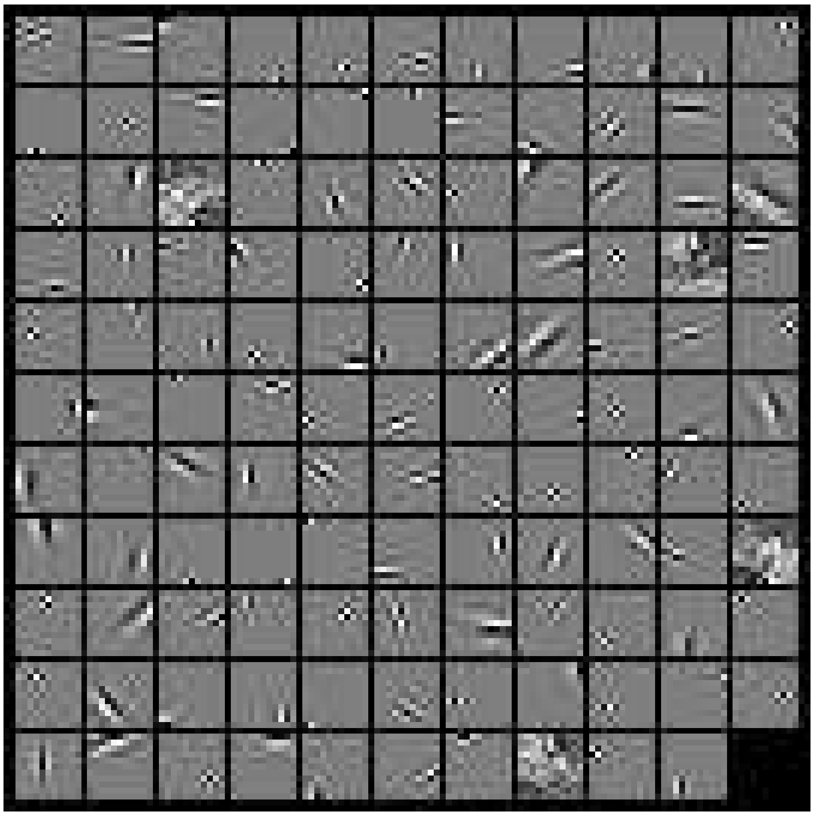

Figure 3: For comparison, ICA filters estimated from natural image sequences by using the symmetric fixed-point algorithm. Note that these filters have a much less smooth appearance than in most published ICA results; this is because for the sake of comparison, we show here the filters and not the basis vectors, and further, no low-pass filtering or dimension reduction was applied in the preprocessing.

temporal response strength correlation seems to produce receptive fields that are somewhat more localized with respect to both spatial frequency and orientation.

When the results are compared against the results in (van Hateren and van der Schaaf, 1998), the most important difference is the peak at zero bandwidth in Figures 4E and 4F. This difference is probably a result of the fact that no dimensionality reduction, antialiasing or noise reduction was performed here, which results in the appearance of very small, checkerboard-like receptive fields. This effect is more pronounced in ICA, which also explains the stronger peak at the 45° angle in Figure 4D.

## 3.4 Control experiments

To ensure the novelty and validity of our results, we made eight control experiments. All other aspects, except those specifically mentioned here, were similar in these experiments as in the main experiment.

### 3.4.1 Control experiment I: no temporal decorrelation

In the main experiment, we used temporally decorrelated data. However, as was seen in Figure 1, there is considerable temporal change in natural video data even without temporal decorrelation. Therefore it is reasonable to ask whether temporal decorrelation is necessary for achieving results like those shown in Figure 2. To answer this question, the algorithm was run for data which was *not* temporally decorrelated. The results are shown in Figure 5A. Although the filters with highest frequency components seem to be somewhat less localized, the results remain qualitatively very similar to those in Figure 2, in that they also resemble Gabor filters. This suggests that the simple-cell-like properties of the results of the main experiment are not a consequence of temporal decorrelation.

### 3.4.2 Control experiments II–IV: longer $\Delta t$

In the main experiment, consecutive samples were separated by 40 ms. Does the phenomenon found in the main experiment hold only for very small $\Delta t$? To answer this question, we examined the case $\Delta t = 120$ ms in control experiment II. As can be seen in Figure 1D, with this time separation preprocessed consecutive sample windows are typically much further from each other than when $\Delta t = 40$ ms. Figure 5B shows the results of applying the symmetric gradient projection algorithm to this data. Although the receptive fields are now larger than in the main experiment, the results are still qualitatively similar. This implies that the discovered relationship applies to short-time natural image sequences in general, not only for some specific value of $\Delta t$.

It seems natural that there would be an upper limit for which the results are no longer qualitatively similar to those in Figure 2. This is indeed the case. The degree of spatial localization decreases when $\Delta t$ increases. This can already be seen to some degree in Figure 5B, and is more pronounced in the results of control experiment III, in which $\Delta t = 480$ ms (Figure 5C). In the 480 ms case most of the filters are poorly localized, resembling Fourier basis vectors corresponding to different frequencies. As $\Delta t$ becomes large enough, spatial localization disappears, and filters also start to lose their orientation

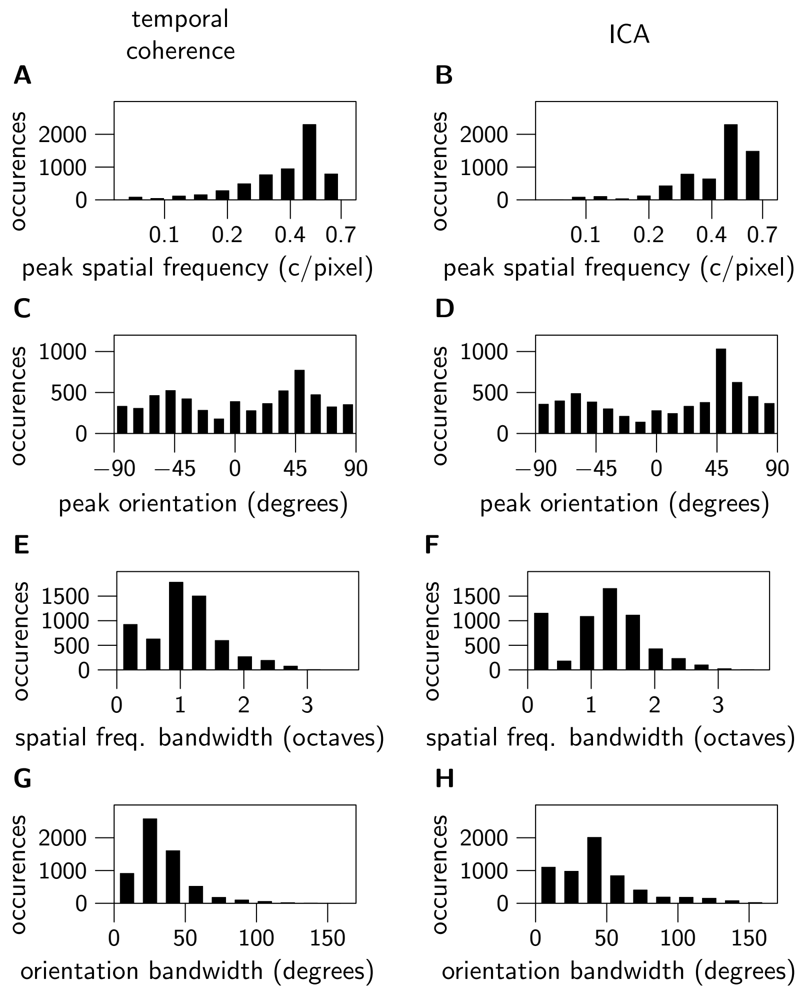


Figure 4: Comparison of properties of receptive fields obtained by optimizing temporal response strength correlation (left column, histograms A, C, E and G) and estimating ICA filters (right column, histograms B, D, F and H). See text for details.

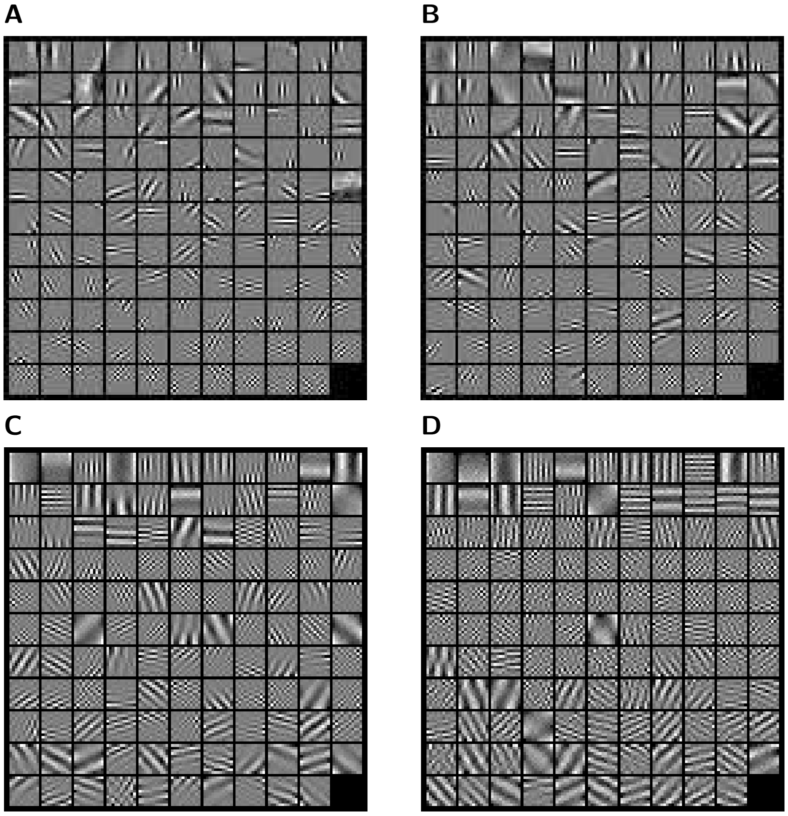


Figure 5: Results of control experiments I–IV. (A) Results of control experiment I in which no temporal decorrelation was performed. (B) Results of control experiment II in which $\Delta t = 120\,\mathrm{ms}$. (C) Results of control experiment III in which $\Delta t = 480\,\mathrm{ms}$. (D) Results of control experiment IV in which $\Delta t = 960\,\mathrm{ms}$.

selectivity. This is illustrated in Figure 5D for $\Delta t = 960\,\text{ms}$ (control experiment IV).

#### 3.4.3 Control experiment V: randomly selected consecutive windows

Control experiment V was made to ensure that the results reflect the dynamics of natural image sequences, and not just the relationship between any arbitrary image patches. Instead of using samples separated by $\Delta t$, random image samples were chosen as consecutive window pairs. No temporal decorrelation was done here, since random window pairs do not have a temporal relationship. Figure 6A shows the resulting spatial filters, which correspond to noise patterns, indicating that the original results in Figure 2 do reflect natural image sequence dynamics.

#### 3.4.4 Control experiment VI: complete removal of the static part of video

As was noted above, in most experiments the data was temporally decorrelated, leading to enhanced temporal change. For example, note from Figure 1D that in the case $\Delta t = 120\,\text{ms}$, after preprocessing (including the normalization step), the peak of the histogram of distances between preprocessed consecutive windows is approximately at 1.4. Remembering that the samples have been normalized, this indicates that a large part of the preprocessed sample consists of consecutive windows which are approximately orthogonal to each other. This supports the idea that our results are not a consequence of the static part of natural image sequences.

However, since preprocessing does not remove the static part of the video *completely*, we made another control experiment (control experiment VI), with $\Delta t = 120\,\text{ms}$, in which the static part was removed altogether. This was done by modifying the preprocessing step so that no temporal decorrelation was done; instead, after removal of the local mean, consecutive windows were *orthonormalized* using the Gram-Schmidt procedure. This removes the static part, the part present already at time $t - \Delta t$, from the window at time $t$ completely. No other form of temporal filtering was performed in this experiment.

Figure 6B shows the results of this control experiment. The resulting filters are still localized, oriented, and have different scales. This shows that the qualitative nature of our results is *not* a consequence of the static part of natural image sequences. The filters are more localized than in the corresponding experiment with temporal decorrelation, probably because removal of the static part is more likely to remove large features from the data (see Section 4). If the same orthogonalization procedure is applied in the case $\Delta t = 40\,\text{ms}$, the results (not shown) are even more localized. This suggests that here too, as in the case of temporally decorrelated data, the degree of spatial localization is a function of $\Delta t$.

#### 3.4.5 Control experiment VII: compensation of observer movement

Control experiment VII was made to study the role of observer (camera) movement. To compensate for this movement, a simple correlation-based tracking mechanism was implemented into the sampling procedure. Tracking was applied before temporal filtering (temporal decorrelation), so each 440 ms

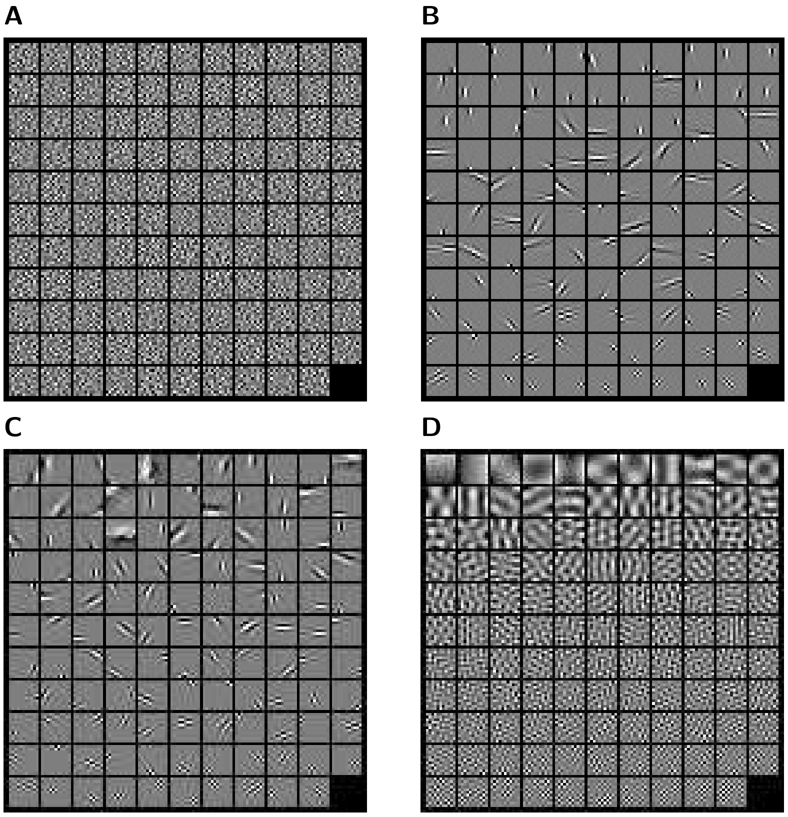


Figure 6: Results of control experiments V–VIII. (A) Results of control experiment V in which consecutive window pairs were chosen randomly. (B) Results of control experiment VI in which the static part of natural image sequences was removed altogether by employing Gram-Schmidt orthogonalization to consecutive windows. (C) Results of control experiment VII in which observer (camera) movement was compensated by using a tracking mechanism. (D) Results of control experiment VIII in which ordinary linear correlation between output values was maximized.

($= \Delta t + 400\,\mathrm{ms}$) sequence was tracked. Let $\mathbf{x}_1$ be the first vectorized sample window in a 440 ms sequence, and $\mathcal{X}_n$ be the set of candidate windows in the $n$th video frame of the sequence, differing at most 10 pixels from the spatial position of the first window $\mathbf{x}_1$. The $n$th window $\mathbf{x}_n$ was chosen by $\mathbf{x}_n = \arg\max_{\mathbf{x}\in\mathcal{X}_n} \frac{\mathbf{x}_{n-1}^T\mathbf{x}}{\|\mathbf{x}_{n-1}\|\|\mathbf{x}\|}$. The results are shown in Figure 6C. The results are qualitatively similar to the original results, showing that the main results are not a consequence of observer movement. The number of low-frequency receptive fields seems to be smaller in the control results. This change is probably caused by decreased large-scale movement (see Section 4).

### 3.4.6 Control experiment VIII: ordinary linear correlation

Finally, the purpose of control experiment VIII was to show that higher-order correlation is indeed needed for the emergence of simple-cell-like filters. To study this, we computed the optimal filter solutions for maximizing linear correlation $f_\ell(\mathbf{w}_k) = \mathrm{E}_t\left\{y_k(t)y_k(t-\Delta t)\right\}$ (see also (Mitchison, 1991)).[1] The unit variance constraint is used here again, so the problem is equivalent to minimizing $\mathrm{E}_t\left\{\left(y_k(t) - y_k(t-\Delta t)\right)^2\right\}$ with the same constraint. A closed-form solution can be derived from necessary Karush-Kuhn-Tucker (KKT) conditions and an additional zero DC constraint. Let $\mathbf{C}_\mathbf{x} = \mathbf{E}\mathbf{D}\mathbf{E}^T$ from which the DC eigenvector corresponding to zero eigenvalue has been dropped out. The necessary conditions are fulfilled by those eigenvectors of $\mathbf{E}\mathbf{D}^{-1}\mathbf{E}^T\mathrm{E}_t\left\{\left(\mathbf{x}(t) - \mathbf{x}(t-\Delta t)\right)\left(\mathbf{x}(t) - \mathbf{x}(t-\Delta t)\right)^T\right\}$ which correspond to non-zero eigenvalues. Analysis of sufficient KKT conditions reveals that the filters fulfilling the first-order conditions behave as in the case of principal components; that is, after the minimizing filter is found and removed from the set of eigenvectors, another eigenvector from the set will be the next minimum, assuming that the output of the next selected filter has to be uncorrelated with the outputs of the previously selected ones. The eigenvector corresponding to the smallest eigenvalue is the global minimum, and the next minimum is the one with the next smallest eigenvalue. Figure 6D shows the resulting filters, sorted according to the corresponding eigenvalue (smallest eigenvalue top left). These filters resemble Fourier basis vectors, and not simple cell receptive fields. Thus we see that emergence of localized receptive fields requires the use of a *nonlinear* temporal correlation measure.

# 4 Discussion

## 4.1 Temporal coherence of large responses

Temporal response strength correlation, as defined by equation (3), does not explicitly measure the rate of change of the output signal. Therefore, it is

[1]Note that this objective function is defined for single filters. A similar single-unit rule for optimizing $\mathrm{E}_t\left\{g(y_k(t))g(y_k(t-\Delta t))\right\}$ can be defined and optimized for temporal response strength correlation. The optima for this objective function are similar to those in Figure 2, but the receptive fields are more elongated. In addition, there are problems with obtaining a complete basis because a deflationary algorithm (Hyvärinen et al., 2001), in which first extracted solutions dominate, has to be used. In the case of linear correlation we do not have this problem since a closed-form solution can be found.

important to examine what type of temporal coherence the objective function measures. In order to do this, consider three different temporally uncorrelated signals, $y_1(t)$, $y_2(t)$, and $y_3(t)$, depicted in Figure 7. (Note that in the main experiment the output signals $y_k(t)$ are temporally uncorrelated because they are spatially linearly filtered from temporally decorrelated input data.) All of these signals have unit energy, and a Gaussian marginal distribution. Signals $y_2(t)$ and $y_3(t)$ have been created by reordering the samples of $y_1(t)$ so that they contain two intervals of high amplitude (see figure caption for details). Signal $y_3(t)$ has the largest temporal response strength correlation of these signals, as measured by equation (3). This is because the objective function *emphasizes the temporal coherence of large amplitudes.* This is *not* true for an arbitrary measure of amplitude correlation. For example, for $g(x) = \sqrt{|x|}$ (concave on interval $]0, \infty]$), signal $y_2(t)$ has a larger measure than $y_3(t)$ ($f(y_2(t)) \approx 0.73$, $f(y_3(t)) \approx 0.71$).

## 4.2 Temporal coherence vs. sparseness

We saw in the previous section that our objective function gives large values when $y_k(t)$ and $y_k(t - \Delta t)$ both have large amplitudes, thus emphasizing the correlation of large activations. This property must not, however, be confused with sparseness. Sparseness of $y_k(t)$ means that very large amplitudes, as well as very small ones, are relatively common. It is thus a property of the *marginal* distribution of $y_k(t)$. The temporal correlation of large amplitudes says nothing about their frequency, nor on any other aspect of the marginal distribution of $y_k(t)$. All the signals in Figure 7 (on the left) have a Gaussian marginal distribution, yet the signals vary considerably in their temporal coherence, as measured by our objective function.

Our measure of temporal coherence could indirectly measure the marginal distribution (sparseness) only in the case where there is little change in the data, that is, if the static part of the image sequence dominates. However, this possibility was ruled out by the use of temporal decorrelation, which ensures that there is quite a large amount of change in the data, as shown in Figure 1. Ultimately, control experiment VI showed that the use of temporal coherence produced similar results even if the static part was removed *completely*, thus proving that the principle of temporal coherence is distinct from sparse coding.

## 4.3 An intuitive explanation of results

### 4.3.1 The importance of translation in natural image sequences

The most universal local visual properties of objects are edges and lines, so in what follows we will limit the discussion to their dynamic properties. Objects can undergo a number of transformations in image sequences: translation, rotation, occlusion, and, for non-rigid objects, deformation. A transformation in the 3-D space can induce a different transformation in an image sequence. For example, a translation towards the camera induces a change of object size in the image sequence. Our hypothesis is that for local edges and lines, and during short time intervals, most 3-D object transformations result in *local translations* of these elements in image sequences. This is of course true for 3-D translations of objects. Figure 8 illustrates this phenomenon for two other transformations:

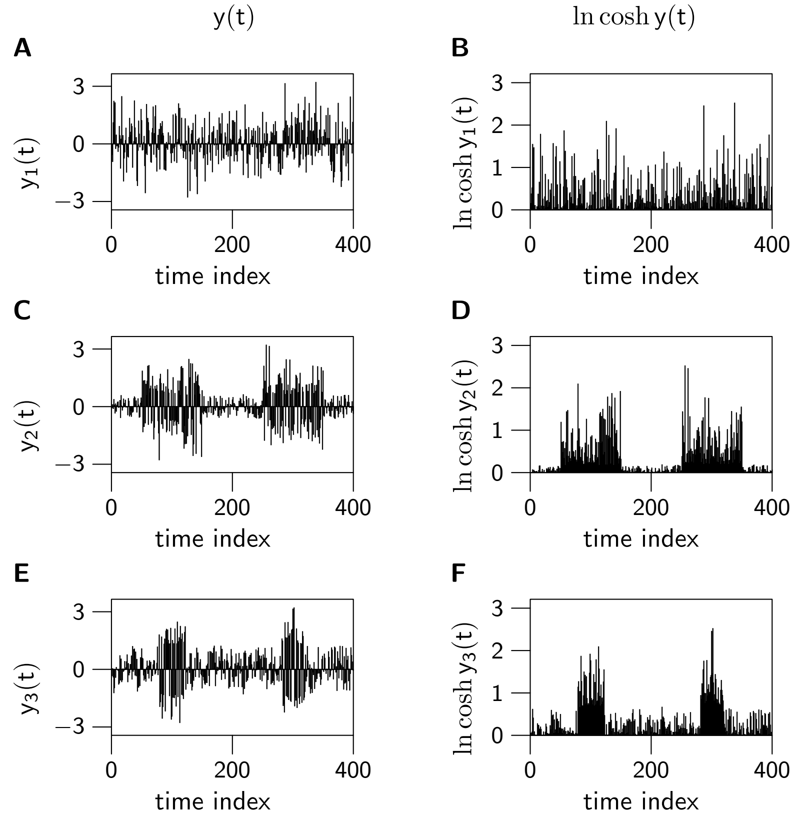


Figure 7: Temporal response strength correlation emphasizes temporal coherence of large amplitudes. Three temporally uncorrelated signals $y(t)$ with unit energy and Gaussian marginal distributions (A,C,E), and their rectified magnitudes $\ln\cosh y(t)$ (B,D,F). The signals in C and E were obtained by rearranging the time indices of the signal in A. This was done so that the two intervals of high amplitude in these signals contain the samples with the largest amplitudes, in random order. In C the total length of the intervals is half of the total signal length, in E this ratio is $\frac{1}{5}$. The signal in E has the largest temporal response strength correlation, as measured by eq. (3) ($f(y_1(t)) \approx 0.13$, $f(y_2(t)) \approx 0.23$, $f(y_3(t)) \approx 0.29$). This illustrates the fact that the objective function emphasizes the temporal correlation of large amplitudes. See text for discussion.

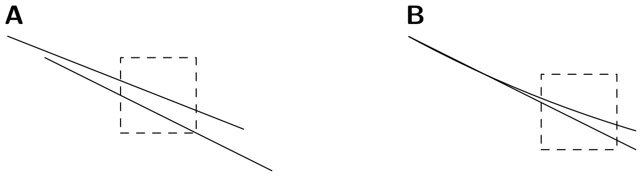


Figure 8: Examples of transformations inducing approximate local translations. Local area is marked with a dashed square. (A) Planar rotation. (B) Bending.

a planar rotation and bending of an object. Note that the effect illustrated in Figure 8A is even more pronounced if object rotation is not purely planar.

#### 4.3.2 Why Gabor-like filters maximize correlation of square-rectified responses

In order to demonstrate the correlation of square-rectified responses at consecutive time points, we will consider the interaction of features and filters in one dimension (orthogonal to the orientation of the filter). This allows us to consider the effect of local translations in a simplified setting. Figure 9 illustrates, in a simplified case, why the temporal response strengths of lines and edges correlate positively as a result of Gabor-like filter structure. Prototypes of two different types of image elements – the profiles of a line and an edge – which both have a zero DC component, are shown in the topmost row of the figure. The leftmost column shows the profiles of three different filters with unit norm and zero DC component: a Gabor-like filter, a sinusoidal (Fourier basis -like) filter, and an impulse filter. The rest of the figure shows the square rectified responses of the filters to the inputs as functions of spatial displacement of the input.

Consider the rectified response of the Gabor-like filter to the line and the edge. The squared response at time $t-\Delta t$ (spatial displacement zero) is strongly positively correlated with response at time $t$, even if the line or edge is displaced slightly. This shows how small local translations of basic image elements still yield large values of temporal response strength correlation for Gabor-like filters. If you compare the responses of the Gabor-like filter to the responses of the sinusoidal filter, you can see that the responses to the sinusoidal filter are typically much smaller. This leads to a lower value of our measure of temporal response strength correlation that emphasizes large values. Also, while the response of an impulse filter to an edge correlates quite strongly over small spatial displacements, when the input consists of a line even a very small displacement will take the correlation to almost zero.

Thus we can see that when considering three important classes of filters – filters which are maximally localized in space, maximally localized in frequency, or localized in both – the optimal filter is a Gabor-like filter, which is localized both in space and in frequency. If the filter is maximally localized in space, it fails to respond over small spatial displacements of very localized image elements. If the the filter is maximally localized in frequency, its responses to the localized image features are not strong enough.

Figure 10 shows why we need nonlinear correlations instead of linear ones: raw output values might correlate either positively or negatively, depending on

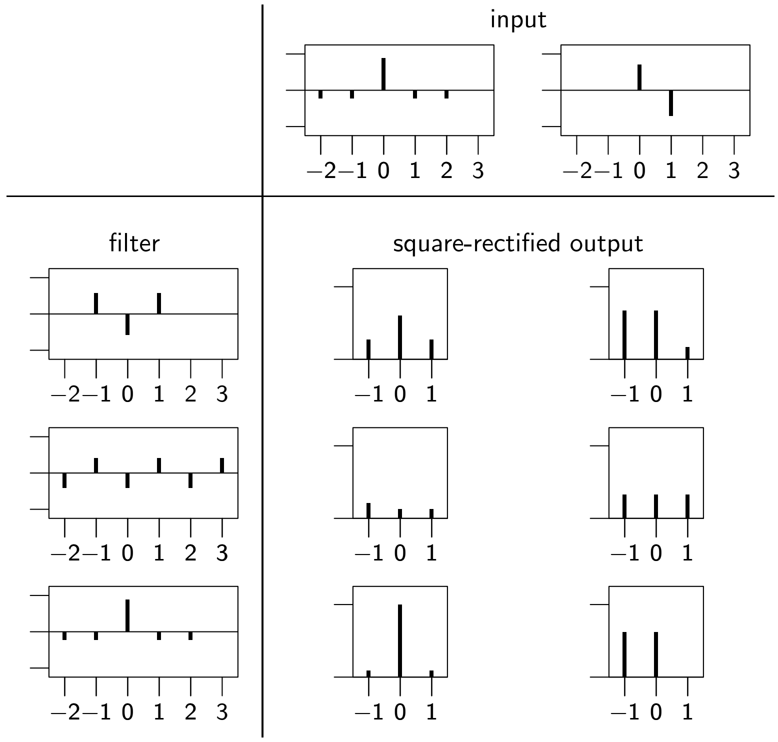


Figure 9: A simplified illustration of why a Gabor-like filter, localized in both space and frequency, yields larger values of temporal response strength correlation than a filter localized only in space or only in frequency. Top row: cross sections of a line (left) and an edge (right) as functions of spatial position. Leftmost column: cross sections of three filters with unit norm and zero DC component – a Gabor-like filter (top), a sinusoidal filter (middle), and an impulse filter (bottom). The other plots in the figure show the responses of the filters to the inputs as a function of spatial displacement of the input. The Gabor-like filter yields fairly large positively correlated values for both types of input. The sinusoidal filter yields small response values. The impulse filter yields fairly large positively correlated values when the input consists of an edge, but when the input consists of a line even a small displacement yields a correlation of almost zero.

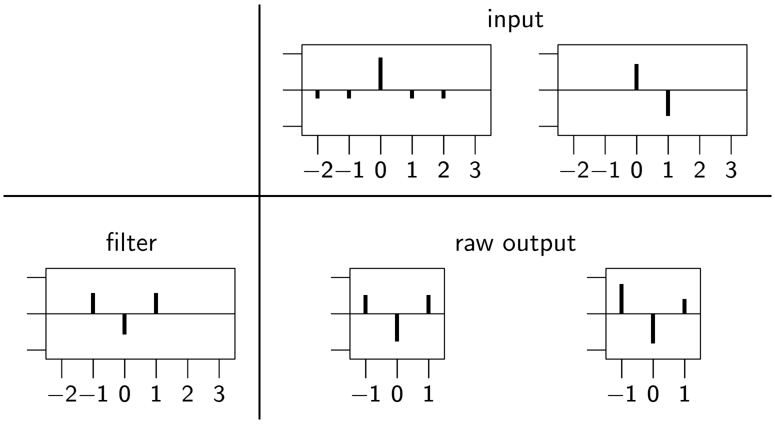


Figure 10: A simplified illustration of why nonlinear correlation is needed for the emergence of the phenomenon. Raw response values of the Gabor-like filter to the line and edge may correlate positively or negatively, depending on the displacement. (See Figure 9 for an explanation of the layout of the figure.)

the displacement. Thus we see why ordinary linear correlation is not maximized for Gabor-like filters, whereas the rectified (nonlinear) correlation is.

### 4.3.3 Emergence of simple-cell-like filters

Future research is needed to provide a detailed analysis of which properties of natural image sequence data are needed for the emergence of simple-cell-like filters. However, at this point we can provide a set of hypotheses as to why oriented, localized filters with multiple scales emerge from the data.

**Orientation** The filters are oriented because the contours (edges and lines) in image sequences are oriented: since our objective function emphasizes strong responses, the filters need to be matched to the dominant features in the data (see Figure 9).

**Localization** As illustrated in Figure 9, the features are limited in width because lines and edges are mostly limited in width as well, and because short-time translations are very local. This second point is supported by the results of control experiments II–IV (Figures 5B–5D), which showed that when $\Delta t$ is increased, localization decreases. The features are limited in length for the same reason that they are limited in ICA or sparse coding: contours in real images have some curvature. Even a weak curvature makes the match (dot-product) with an elongated Gabor quite weak, and gives poor temporal coherence (Hoyer and Hyvärinen, 2002).

**Multiple scales** The filters respond to multiple scales for two reasons. The first reason is the scale invariance of natural image sequences. This explanation is supported by the results of control experiment VI (Figure 6B), in which the static part of image sequences was removed completely. The results exhibit a narrower range of different scales, because removal of the static part of the sequences is more likely to remove large features. The second reason is that the features move with many different velocities in the data. This is supported by the results of control experiment VII (Figure 6C), in which tracking was used to compensate for observer movement. The results exhibit a narrower range of different scales, because tracking is likely to reduce the velocity of moving features in the data.

## 4.4 The case of spatiotemporal receptive fields

Due to limited computational resources, we are currently unable to estimate the most temporally coherent spatiotemporal receptive fields. However, argumentation similar to the intuitive explanation provided above can be given to illustrate a similar phenomenon in the spatiotemporal case. Figure 11 illustrates a case in which a vertical line is moving in the image sequence. The response of a simple-cell-like motion-selective spatiotemporal filter (DeAngelis et al., 1993a), whose spatiotemporal position and orientation match the initial position of the line and its direction of movement, is large in magnitude at consecutive time-points. This illustrates how large temporal response strength correlation could arise in the case of spatiotemporal receptive fields.

## 4.5 The case of a nonlinear (non-negative) cell model

Linear filters with negative coefficients and/or negative-valued data have signed outputs. Their widespread use as simple cell models is often motivated by an interpretation in which the term "simple cell" in fact refers to a unit of two cells. These two positive-output cells, with reversed polarities, are modeled using a single linear filter with signed output.[2] This two-cell approach will be explained in detail below. At this point notice, however, that this coupling is very different from complex cell pooling of simple cells. In complex cell pooling the coupled cells respond to similar features at *different* spatial positions, whereas two cells with opposite polarities respond to similar features at the same spatial position.

For single simple cells, a more realistic basic model of the mapping from input to output via a receptive field is a combination of linear filtering and a nonlinearity called *half-wave rectification* (see, e.g., (Heeger, 1992)). Using the same notation as above, the output of the cell, $y_k(t)$, is computed by

$$y_k(t) = \max\left\{0, \mathbf{w}_k^T \mathbf{x}(t)\right\}, \tag{5}$$

instead of the purely linear input-output relationship (1). In this model the output of a cell is never negative.

The purely linear model (1) combines the outputs of two such positive-output simple cells with reversed polarities. This is implemented so that the positive output values correspond to the output of one cell, and the negative values correspond the the output of another cell, with otherwise similar receptive field except for a change of the sign of all the connection weights. The exact way an input pattern is mapped into a response in such a model is as follows. Let $y_{k,1}(t)$ and $y_{k,2}(t)$ denote the outputs of two cells with reversed polarities. Their

[2] Another possibility would be to consider the negative and positive values as changes from maintained firing rate. However, simple cells have a low maintained firing rate, which makes this approach undesirable (Heeger, 1992).

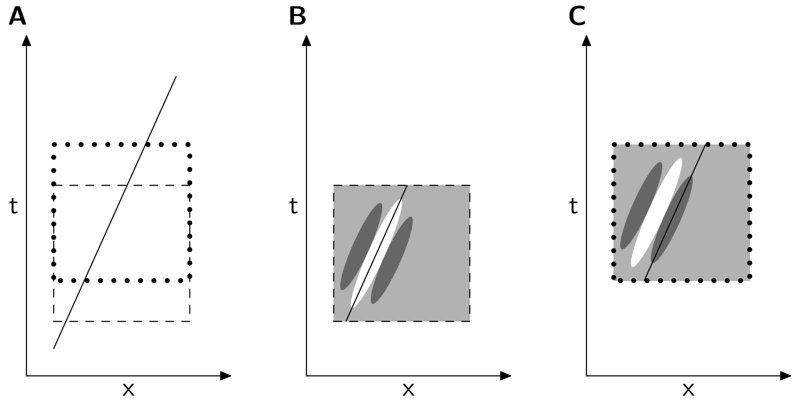


Figure 11: An illustration of how temporal response strength correlation could be exhibited by the outputs of simple-cell-like *spatiotemporal* receptive fields. A phenomenon analog to translation in the spatial case can be observed in the spatiotemporal case. Let $x$ and $y$ denote the horizontal and vertical spatial coordinates, respectively, and let $t$ denote the temporal coordinate. (A) The spatiotemporal trace (solid line) of a moving vertical line is shown here in the $x$–$t$ coordinate system. The plot is similar for all $y$-coordinates because the moving line is vertical. Two different overlapping spatiotemporal input windows, separated by a small time difference, are also marked, one with dashed line, and the other with dotted line. (B) A simple-cell-like spatiotemporal receptive field, with position and orientation that match the initial position of the line and its direction of movement, responds strongly to the moving line. Here the spatiotemporal filter has been superimposed over the dashed temporal window – white color indicates large positive values in the filter, dark color indicates large negative values, and middle gray indicates zero values. (C) When the same spatiotemporal receptive field, at the same spatial position, is applied to the same input a moment later (dotted spatiotemporal input window), the response is still strong, but the sign changes. Therefore the temporal response strength correlation of the outputs of the simple-cell-like spatiotemporal receptive field would be large for this kind of input.

outputs are given according to equation (5) by $y_{k,1}(t) = \max\left\{0, \mathbf{w}_k^T\mathbf{x}(t)\right\}$, and $y_{k,2}(t) = \max\left\{0, -\mathbf{w}_k^T\mathbf{x}(t)\right\}$. The overall output is defined as

$$y_k(t) = y_{k,1}(t) - y_{k,2}(t). \tag{6}$$

It is straightforward to show that this model is equivalent to a purely linear model, that is, that $y_{k,1}(t) - y_{k,2}(t) = \mathbf{w}_k\mathbf{x}(t)$.

Therefore, one might interpret our model as measuring the temporal coherence of this two-cell unit, where the cells have similar receptive fields with reversed polarities. A large value of the objective function indicates that either of the two cells responds strongly to the input. As was mentioned above, using the linear model is a common approach – the same model is used for example in (Bell and Sejnowski, 1997; Olshausen and Field, 1996; van Hateren and van der Schaaf, 1998), which makes the comparison of results obtained with these different models feasible.

However, a natural question is whether the same principle applies to the outputs of individual, half-rectified simple cells? To address this issue, we computed the optimal solution for half-wave rectified cell outputs, that is, replaced the original linear input-output relationship (1) with the half-wave rectified relationship (5), and computed the optimal solution for this model. The same constraint (4) was used in this case for simplicity. Instead of using the temporally decorrelated data set, we used the orthonormalized data set with $\Delta t = 120\,\text{ms}$, also used in control experiment VI (Section 3.4.4). This is because by using the orthonormalized data set we excluded the possibility of obtaining simple-cell-like receptive fields because of the static part of the video. This is important here since the static part might yield positive responses at consecutive time instances.

The results of this experiment are shown in Figure 12A. Although the features are not as well defined as before, the resulting filters still show orientation, localization, and different spatial scales.

What is the reason for the emergence of such features even in this case when the negative part of the linear filter response has been discarded, and the static part of the video has been removed completely? First, as in the purely linear case, the object function emphasizes the temporal correlation of large responses, and oriented and localized filters respond strongly to lines and edges. But how does temporal correlation arise? An illustration of a possible explanation is shown in Figure 12B. When a Gabor-like filter is applied to a line, the half-wave rectified output still correlates positively over time, but more weakly, and over longer spatial displacements.

Finally, let us note that some form of temporal coherence of simple cell outputs is *implicitly* assumed in many studies. The firing rate of a simple cell is typically assumed to code for the output of a linear filter, possibly after some simple nonlinear transformations, such as half-wave rectification. This requires that the output of the linear filter has some temporal coherence. If the output of the linear filter changes quite randomly, the firing rate cannot provide much information on the output, because then the firing rate is very noisy when computed over short time intervals. Thus, maximization of temporal coherence could have a relation to maximization of the efficiency of a code based on firing rates.

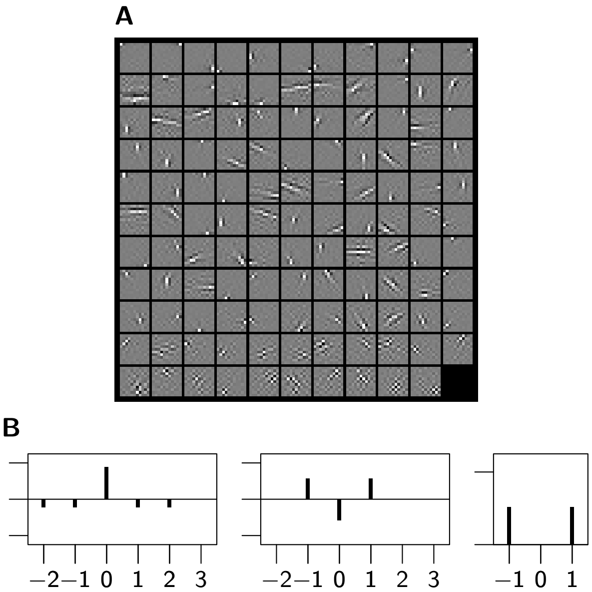


Figure 12: Temporal response strength correlation for a half-wave rectified (nonlinear, non-negative) cell model. (A) The results of running the algorithm for a nonlinear cell model with half-wave rectification of cell output. (B) When a simple-cell-like filter is applied to an input containing a moving line, the half-wave rectified output is still correlated positively over time, but the correlation is weaker, and the spatial displacement needed for the correlation is larger. Cross sections of a line (left) and a filter (middle), and the half-wave rectified output (right). See also Figure 9.

### 4.6 Implications of our results

In this paper we have shown that simple cell type receptive fields maximize temporal response strength correlation at cell output, when the input consists of natural image sequences. Temporal response strength correlation, or temporal correlation of rectified cell outputs, can be considered as a measure of temporal coherence. Our findings have several important implications. First, temporal coherence provides an alternative or complementary theory to sparse coding as a computational principle behind the formation of simple cell receptive fields. The application of either one of these principles results in the emergence of simple cell properties from natural data.

Second, Földiák and others (Földiák, 1991; Kayser et al., 2001; Kohonen et al., 1997; Wiskott and Sejnowski, 2002) have proposed that invariant visual representations, such as those found in complex cells, can be found by maximizing temporal coherence. (For an alternative complex-cell model using sparse coding see (Hyvärinen and Hoyer, 2001).) Our results show that this principle is applicable to the visual system even on the level of simple cells, which usually are not considered as invariant detectors. Although in some complex cell models (Kayser et al., 2001; Kohonen et al., 1997) simple cell receptive fields are obtained as by-products, the learning is strongly modulated by the complex cells, and therefore is very different from our model, which only considers the statistics of simple cell outputs. Moreover, the simple cell receptive fields learned in (Kayser et al., 2001; Kohonen et al., 1997) do not seem to have the important properties of spatial localization and multiresolution (different scales).

Furthermore, whereas most earlier research results linking temporal coherence and properties of visual system have been based on theoretical considerations and simulated data, the results published in this paper have been computed from natural image sequence data. To our knowledge this is the first time that localized and oriented receptive fields, with different scales, have been shown to emerge from natural data using the principle of temporal coherence. A step like this is important for any theory which tries to explain the structure and functionality of sensory neural networks using the statistical properties of natural input data.

#### Acknowledgements

We would like to thank Hans van Hateren for the permission to use the natural image sequences, and for providing us with the program used to measure receptive field properties. We would also like to thank Patrik Hoyer for helpful comments.

# A Details of the symmetric gradient projection algorithm

In this section we give a detailed description of the optimization algorithm. First, the maximization of objective function (3) under constraint (4) can be made easier by employing whitening, a temporary change of coordinates, so that constraint (4) is transformed into an orthonormality constraint. Let equation $\mathbf{C_x} = \mathbf{EDE}^T$ denote the eigenvalue decomposition of matrix $\mathbf{C_x}$. If an eigenvalue of $\mathbf{C_x}$ is zero, then that eigenvalue can be dropped out of eigenvalue matrix $\mathbf{D}$, and the corresponding eigenvector can be removed from eigenvector matrix $\mathbf{E}$. This was the case in our experiments because preprocessing reduced the dimensionality of input data $\mathbf{x}(t)$ by one when the DC component was removed from data. This dimensionality reduction by one also means that the number of filters that can be extracted is $K \leq N^2 - 1$. Defining matrix

$$\mathbf{U} = \mathbf{W}\mathbf{E}\mathbf{D}^{1/2}, \tag{7}$$

constraint (4) can be expressed as a an orthonormality constraint

$$\mathbf{U}\mathbf{U}^T = \mathbf{I}. \tag{8}$$

This simpler constraint will be easier to handle in the optimization algorithm below.

To express objective function (3) using the same transformed filter matrix $\mathbf{U}$, we have to solve equation (7) for matrix $\mathbf{W}$. This is equivalent to solving matrix equation $\mathbf{E}^T\mathbf{D}^{1/2}\mathbf{W}^T = \mathbf{U}^T$. In the case where the DC component has been removed this is an underdetermined set of linear equations. By imposing the additional zero DC constraint on the filters in rows of $\mathbf{W}$, the solution is given by

$$\mathbf{W} = \mathbf{U}\mathbf{D}^{-1/2}\mathbf{E}^T \tag{9}$$

(see (Hurri, 1997, pages 16–17) for details). Substituting this into equation (2) gives

$$\mathbf{y}(t) = \mathbf{U}\mathbf{D}^{-1/2}\mathbf{E}^T\mathbf{x}(t) = \mathbf{U}\mathbf{z}(t), \tag{10}$$

where signal vector $\mathbf{z}(t) = \mathbf{D}^{-1/2}\mathbf{E}^T\mathbf{x}(t)$. This transformation from input data $\mathbf{x}(t)$ to $\mathbf{z}(t)$ is called PCA whitening (Hyvärinen et al., 2001, pages 140–141).

The above shows that after input data $\mathbf{x}(t)$ is whitened, we can optimize

$$f(\mathbf{U}) = \sum_{k=1}^{K} \mathrm{E}_t \left\{ g(y_k(t)) g(y_k(t - \Delta t)) \right\}, \tag{11}$$

where output $\mathbf{y}(t)$ is given by equation (10), with respect to orthonormality constraint (8). When the solution to this problem is found, the solution to the original problem is given by equation (9).

The actual optimization algorithm used for this constrained optimization problem is a variant of the gradient projection method of Rosen (for the original algorithm see e.g. (Luenberger, 1969, pages 297–299)). Let $\alpha(m)$ be a non-negative, decreasing sequence of real numbers, which converges to zero (initial step length $\alpha(0)$ is changed adaptively to speed up convergence). Let $\mathbf{U}(0)$ be a random orthonormal matrix, and let $\mathbf{U}(n)$ be the value of matrix $\mathbf{U}$ at iteration $n$. The algorithm finds a new candidate point by projecting matrix $\mathbf{U}(n) + \alpha(m)\frac{df(\mathbf{U}(n))}{d\mathbf{U}}$ onto the constraint surface defined by (8). If the candidate point is not an improvement, $m$ is increased by one to find a new candidate point.

The critical step in the algorithm is the projection onto the constraint surface. This is achieved by optimal symmetric orthogonalization. Let $\mathbf{A}$ be a square matrix with full rank. Then $\mathbf{B} = \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1/2}$ is the nearest orthonormal matrix to $\mathbf{A}$ with respect to Frobenius matrix norm (see (Fan and Hoffman, 1955, Theorem 1) for a generalized proof of this). This is exactly the property required to achieve the projection step.

The resulting algorithm for maximizing temporal response strength correlation (TRSC) is shown below. The input arguments of the program are whitened data vectors $\mathbf{z}(t)$, convergence tolerance $\epsilon$, and initial step length $\alpha(0)$ (in this version $\alpha(m) = \frac{\alpha(0)}{2^m}$). Adaptation of initial step length $\alpha(0)$ is not included in this description. The algorithm assumes convergence when the objective function changes very little between successive steps. In our experiments convergence tolerance $\epsilon$ varied between $10^{-3}$ and $10^{-6}$. Note that, denoting $\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_K]^T$, the $k$th row of $\frac{df(\mathbf{U})}{d\mathbf{U}}$ is given by the transpose of

$$\begin{aligned} \frac{\partial f(\mathbf{U})}{\partial \mathbf{u}_k} &= \frac{\partial}{\partial \mathbf{u}_k} \mathrm{E}_t \left\{ g(y_k(t)) g(y_k(t-\Delta t)) \right\} \\ &= \mathrm{E}_t \left\{ g'(y_k(t)) g(y_k(t-\Delta t)) \mathbf{z}(t) + g(y_k(t)) g'(y_k(t-\Delta t)) \mathbf{z}(t-\Delta t) \right\}. \end{aligned}$$

<u>**funct**</u> $\mathbf{U} = \mathit{maxTRSC}(\mathbf{z}(t), \epsilon, \alpha(0))$
  $\mathbf{U}(0) := \mathit{rand}()$ <u>**comment**</u>: A random initial starting point.
  $\mathbf{U}(0) := \mathit{symmetricOrth}(\mathbf{U}(0))$
  $\mathit{fOld} := 0$
  $n := 0$
  <u>**while**</u> $f(\mathbf{U}(n)) - \mathit{fOld} > \epsilon$
    $\mathit{fOld} := f(\mathbf{U}(n))$
    $m := 0$
    <u>**while**</u> $f\left(\mathit{symmetricOrth}\left(\mathbf{U}(n) + \frac{\alpha(0)}{2^m}\frac{df(\mathbf{U}(n))}{d\mathbf{U}}\right)\right) \leq \mathit{fOld}$
      $m := m + 1$
    <u>**end**</u>
    $\mathbf{U}(n+1) := \mathit{symmetricOrth}\left(\mathbf{U}(n) + \frac{\alpha(0)}{2^m}\frac{df(\mathbf{U}(n))}{d\mathbf{U}}\right)$
    $n := n + 1$
  <u>**end**</u>
  $\mathbf{U} := \mathbf{U}(n)$
.
<u>**funct**</u> $\mathbf{B} = \mathit{symmetricOrth}(\mathbf{A})$
  $\mathbf{B} := \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1/2}$
.

# References


Becker, S. (1993). Learning to categorize objects using temporal coherence. In Hanson, S. J., Cowan, J. D., and Giles, C. L., editors, *Advances in Neural Information Processing Systems*, volume 5, pages 361–368. Morgan Kaufmann Publishers.

Becker, S. and Hinton, G. (1992). Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163.

Bell, A. and Sejnowski, T. J. (1997). The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338.

Carandini, M., Heeger, D. J., and Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, 17(21):8621–8644.

DeAngelis, G. C., Ohzawa, I., and Freeman, R. D. (1993a). Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. *Journal of Neurophysiology*, 69(4):1091–1117.

DeAngelis, G. C., Ohzawa, I., and Freeman, R. D. (1993b). Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation. *Journal of Neurophysiology*, 69(4):1118–1135.

Dong, D. W. and Atick, J. (1995). Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, 6(2):159–178.

Fan, K. and Hoffman, A. J. (1955). Some metric inequalities in the space of matrices. *Proceedings of the American Mathematical Society*, 6:111–116.

Field, D. (1994). What is the goal of sensory coding? *Neural Computation*, 6(4):559–601.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9:181–198.

Hoyer, P. O. and Hyvärinen, A. (2002). A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*, 42(12):1593–1605.

Hurri, J. (1997). Independent component analysis of image data. Master's thesis, Helsinki University of Technology. Electronic version available at `http://www.cis.hut.fi/jarmo/publications/`.

Hyvärinen, A. and Hoyer, P. O. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, 41(18):2413–2423.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons.

Kayser, C., Einhäuser, W., Dümmer, O., König, P., and Körding, K. (2001). Extracting slow subspaces from natural videos leads to complex cells. In Dorffner, G., Bischof, H., and Hornik, K., editors, *Artificial Neural Networks – ICANN 2001*, pages 1075–1080. Springer.

Kohonen, T., Kaski, S., and Lappalainen, H. (1997). Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM. *Neural Computation*, 9(6):1321–1344.

Luenberger, D. G. (1969). *Optimization by Vector Space Methods*. John Wiley & Sons.

Mitchison, G. (1991). Removing time variation with the anti-Hebbian differential synapse. *Neural Computation*, 3(3):312–320.

Olshausen, B. A. and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

Olshausen, B. A. and Field, D. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325.

Oppenheim, A. and Schafer, R. (1975). *Digital Signal Processing*. Prentice-Hall.

Palmer, S. E. (1999). *Vision Science – Photons to Phenomenology*. The MIT Press.

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216.

Stone, J. (1996). Learning visual parameters using spatiotemporal smoothness constraints. *Neural Computation*, 8(7):1463–1492.

van Hateren, J. H. and Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London B*, 265(1412):2315–2320.

van Hateren, J. H. and van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London B*, 265(1394):359–366.

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