Local Feature Analysis: A general statistical theory for object representation

Penio S Penev† and Joseph J Atick‡ §
Computational Neuroscience Laboratory
The Rockefeller University
1230 York Avenue
New York, NY 10021-6399
http://venezia.rockefeller.edu
†penev@venezia.rockefeller.edu
‡atick@venezia.rockefeller.edu

§ To whom all correspondence should be addressed.

Short title: Local Feature Analysis

October 9, 1996

Abstract.

Low-dimensional representations of sensory signals are key to solving many of the computational problems encountered in high-level vision. Principal Component Analysis has been used in the past to derive practically useful compact representations for different classes of objects. One major objection to the applicability of PCA is that it invariably leads to global, nontopographic representations that are not amenable to further processing and are not biologically plausible. In this paper we present a new mathematical construction—Local Feature Analysis (LFA)—for deriving local topographic representations for any class of objects. The LFA representations are sparse-distributed and, hence, are effectively low-dimensional and retain all the advantages of the compact representations of the PCA. But unlike the global eigenmodes, they give a description of objects in terms of statistically derived local features and their positions. We illustrate the theory by using it to extract local features for three ensembles—2D images of faces without background, 3D surfaces of human heads, and finally 2D faces on a background. The resulting local representations have powerful applications in head segmentation and face recognition.

1. Introduction

In most evolved animals the representation of sensory signals formed by the peripheral receptors is very high-dimensional. For example, in the human retina there are more than six million cones, each capable of discriminating about a hundred shades of light. From the activity of this huge array of receptors the brain has to discover where and what objects there are in the field of view and recover in detail their attributes such as color, texture, and 3D nature.

One can argue that the goal of sensory processing should be to reduce the dimensionality of the input space. All high-level vision problems become more tractable when formulated in a low-dimensional space, e.g., shape-from-shading (Atick et al. 1996) and face recognition (Atick et al. 1995). Also, good generalization critically depends on finding the correct low-dimensional representation (in context of neural networks see the review by Geman et al. 1992).

In representing natural signals one expects to be able to lower the dimensionality considerably because these signals possess significant statistical regularities, or redundancies (for experimental measurements of various regularities see Burton and Moorhead 1987; Field 1987; Tolhurst et al. 1992; Hancock et al. 1992; Ruderman and Bialek 1994; Ruderman 1994b; Dong and Atick 1995a; Atick et al. 1996). These are manifested in the fact that the ensemble of actual activation of sensory receptors in response to natural stimuli occupies a small fraction of the total allowed phase space—the space of all possible receptor activations. Thus, one can lower the dimensionality by finding a suitable parameterization of the subspace occupied by natural stimulation.

Furthermore, if one is concerned not with the translation- and scale-invariant ensemble of all natural signals, but only with a limited class of objects, for example correctly aligned and scaled human faces, then there is additional expectation for finding a low-dimensional representation. Indeed, not every natural signal is a human face, so even the limited subspace of the receptor space, occupied by natural signals, is not entirely populated with faces. Intuitively, one would expect that there is a small number of variables that a face should be described with—much lower than the number of pixels needed to represent it, and practical measurements have confirmed that (Sirovich and Kirby 1987).

Currently, there are many algorithms of varying complexity for attempting to discover low-dimensional representations of signals by relying on their statistical regularities†. So far, however, the most practical and systematic method has been Principal Component Analysis. PCA assumes that the probability density of the input ensemble in the space of receptor activation patterns is significantly nonzero only in a low-dimensional linear subspace, which is subsequently parameterized with a linear expansion in the eigenvectors of the correlation matrix of the ensemble. The power of PCA stems from its ease of computability and its general applicability, and so far it has been used in many real-world problems. For example, it has been used to produce a representation of 2D faces—eigenfaces (Sirovich and Kirby 1987)—and

† These include algorithms for Principal Component Analysis (Linsker 1988; Oja 1989; Sanger 1989; Földiak 1990; Plumbey 1991), Gaussian Component Analysis (Goodall 1960; Atick et al. 1993), Independent Component Analysis (Jutten and Herault 1991; Comon 1994; Bell and Sejnowski 1995), Factorial Learning (Barlow et al. 1989; Hentschel and Barlow 1991; Schmidhuber 1992; Redlich 1993a; Redlich 1993b; Atick et al. 1993), Infomax (Linsker 1988), Imax (Becker and Hinton 1992), Projection Pursuit (Intrator 1992), Matching Pursuit (Phillips and Vardi 1995), and symplectic maps (Deco et al. 1995), etc. For a recent review of all these techniques see the book (Deco and Obradovic 1996).

of 3D heads—eigenheads (Atick et al. 1995)—which are powerful representations for face recognition and for shape-from-shading.

Undoubtedly, PCA has some significant limitations. For example, PCA is not capable of extracting local feature-like structures in objects. Also, in general, PCA produces global nontopographic linear filters whose output is not amenable to subsequent processing very naturally. Local representations are desirable since they offer robustness against variability due to changes in localized regions of the objects.

Can we rectify these shortcomings of PCA without resorting to complex, practically noncomputable algorithms? In this paper we show that the answer is yes. More precisely, for any input ensemble of objects, we show how to construct, from the global PCA modes, a *local topographic* representation of objects in terms of local features. The procedure—which we call Local Feature Analysis (LFA)—derives a dense set of local feed-forward receptive fields, defined at each point of the receptor grid and different from each other, that is optimally matched to the input ensemble, and whose outputs are as decorrelated as possible. Since the objects from the input ensemble span only a very-low-dimensional subspace, the dense set of outputs necessarily contains residual correlations. We use these residual correlations to define lateral inhibition which acts to sparsify the output. Thus, the final representation is a *local sparse-distributed* representation in which only a small number of outputs are active for any given input. The number of active units is on the order of the dimensionality of the PCA subspace, but their subset changes from one input example to another, providing valuable additional information about the locations of the features in the currently represented example. We give a practical implementation for finding approximations to the stable states of this network dynamics that is computationally very efficient on a serial machine.

We illustrate LFA by using it to derive local features in three different object ensembles that are similar to ensembles for which global representations have been derived in the past (Sirovich and Kirby 1987; Atick et al. 1996). The first ensemble—2D images of faces without background—serves to illustrate the ability of the method to derive local features intrinsic to objects; it yields receptive fields for noses, mouths, eyes, cheeks, etc. The second ensemble—3D surfaces of human heads—while producing the same conceptual features, has the added advantage of being much more controlled since it contains no variability due to scale, light, or albedo in the data. In this case we derive high quality 3D filters that are matched to the 3D features. The third ensemble comprises of 2D images of faces on a background. Besides the regular "face" features, the most prominent additional ones that LFA derives, are those signaling transitions from the background to the head. The positions of activation provide an accurate outline of the head that could subsequently be used for segmentation.

The organization of this paper is as follows. In Section 2 we present the LFA formalism and illustrate it on two different ensembles of objects. In Section 3 we show how a local sparse-distributed representation can be derived from the representation of Section 2. There we give a neural network architecture as well as an efficient serial algorithm for producing output sparsification and illustrate it on the previously introduced ensembles. Section 4 is discussion. In Appendix A we give some information about the databases used to derive the results in this paper. In Appendix B we argue on the basis of minimal mean square error (m.s.e.) of the reconstruction, that the ultimate representation is a hybrid between the global PCA and the local LFA representations. We show that it can be obtained by using the

first few global modes together with the local receptive fields derived in Section 2. A comparison between different strategies of sparsification is given in Appendix C.

2. Local features from global modes

The mathematical construction that follows is of broad applicability and is initially presented in the most general terms. Let a sensory signal be given by $\phi(\mathbf{x})$ where $\{\mathbf{x}\}$ is a sampling, or receptor grid which needs not be regular, with V total sampling points that possess some topography. The index $\mathbf{x}$ can be a spatial, temporal or any other modality index or combination thereof. For images $\phi(\mathbf{x}) = I(\mathbf{x})$ with $\mathbf{x}$ the 2D grid of photoreceptors; for surfaces $\phi(\mathbf{x})$ is given by the radial map $r(\theta, \ell)$ in cylindrical coordinates. An ensemble of sensory signals will be denoted by $\{\phi^t(\mathbf{x}), t = 1, \cdots, T\}$ where T is the total number of examples in the ensemble.†

We use PCA to extract a hierarchical orthonormal basis of the linear subspace that the input ensemble spans. This is done by diagonalizing the correlation matrix of the ensemble

$$R(\mathbf{x}, \mathbf{y}) \equiv \langle \phi^t(\mathbf{x}) \phi^t(\mathbf{y}) \rangle \equiv \frac{1}{T} \sum_{t=1}^{T} \phi^t(\mathbf{x}) \phi^t(\mathbf{y}) = \sum_{r=1}^{T} \Psi_r(\mathbf{x}) \lambda_r \Psi_r(\mathbf{y}) \qquad (1)$$

to produce the orthonormal set of eigenmodes $\Psi_r(\mathbf{x}), r = 1, \cdots, T$ and their respective eigenvalues λ_r, ordered in the natural hierarchy of decreasing magnitude.‡

The PCA representation§

$$\phi(\mathbf{x}) = \sum_{r=1}^{T} A_r \Psi_r(\mathbf{x}) \text{ with } A_r = \int \Psi_r(\mathbf{x}) \phi(\mathbf{x}) \equiv \int K_r(\mathbf{x}) \phi(\mathbf{x}) \qquad (2)$$

is decorrelated in the sense that

$$\langle A_r A_q \rangle = \lambda_r \delta_{rq}. \qquad (3)$$

It has the property of best reconstruction—truncations in the expansion (eq. 2) with $N < T$ have a minimum mean square error (m.s.e.) $\langle \int |\phi(\mathbf{x}) - \phi^{rec}(\mathbf{x}))|^2 \rangle \equiv \langle \|\phi - \phi^{rec}\|^2 \rangle$ where

$$\phi^{rec}(\mathbf{x}) \equiv \sum_{r=1}^{N} A_r \Psi_r(\mathbf{x}). \qquad (4)$$

It has been shown in (Sirovich and Kirby 1987; Atick et al. 1996) and confirmed by us in the current work that the PCA representation (eq. 2) of human faces and heads generalizes well: given a certain m.s.e. tolerance, only a small number of modes

† In general, the best description of the ensemble is through the probability density function $\mathcal{P}[\phi(\mathbf{x})]$. In the context of PCA, one is trying to discover some of the properties of the ensemble from a limited set of examples $\{\phi^t(\mathbf{x})\}$, without the full knowledge of $\mathcal{P}[\phi(\mathbf{x})]$. The only information PCA uses is the second-order correlation function $R(\mathbf{x}, \mathbf{y})$ (eq. 1), a byproduct of $\mathcal{P}[\phi(\mathbf{x})]$, which itself is not used anywhere in this analysis.

‡ In the case $T << V$, $R(\mathbf{x}, \mathbf{y})$ is highly degenerate (rank deficient), and one can use the snapshot method (Sirovich 1987) to carry out the diagonalization of a prohibitively large $V \times V$ matrix through a diagonalization of a $T \times T$ matrix instead.

§ Throughout the paper we use the integral over the input space $\mathbf{x}$ to signify the dot product of two input patterns. Implicit in our definition of the integral is a normalization by the volume V, i.e., $\int \equiv \frac{1}{V} \int_V d\mathbf{x}$

$N < T << V$ is needed to represent out-of-sample examples. In Appendix C we also show that it has the feature of object constancy in the sense that it suppresses input noise (see Fig. A3).

The PCA representation (eq. 2) is compact—it has a greatly reduced dimensionality but is typically nonlocal—the supports of the kernels K_r extend over the entire range of $\mathbf{x}$. Also, it is not topographic—nearby values in the r index do not possess any relationship among each other in contrast to nearby values of the grid variable $\mathbf{x}$ which obey topography. Locality and topography are desirable features in certain segmentation and pattern analysis tasks, and they seem to be properties of neural processing, at least at the early to intermediate stages in the visual pathway. So, it would be of interest to discover representations that possess these properties.

Topography means that the kernels of the representation should be labeled with the grid variable $\mathbf{x}$ instead of the PCA eigenmode index r. The most general topographic kernel that projects signals to the subspace spanned by the eigenmodes is given by

$$K(\mathbf{x},\mathbf{y}) = \sum_{r,s=1}^{N} \Psi_r(\mathbf{x}) Q_{rs} \Psi_s(\mathbf{y}) \tag{5}$$

where Q_{rs} is a priori an arbitrary matrix. The space of the outputs

$$O(\mathbf{x}) \equiv \int K(\mathbf{x},\mathbf{y})\phi(\mathbf{y}) = \sum_{r,s=1}^{N} \Psi_r(\mathbf{x}) Q_{rs} A_s \tag{6}$$

thus inherits the same topography as the input space.

By insisting on topography, we arrive at V outputs $O(\mathbf{x})$ described by $N << V$ linearly independent variables A_r. This means that we can no longer satisfy the desirable condition of output decorrelation $\langle O(\mathbf{x})O(\mathbf{y})\rangle = \delta(\mathbf{x},\mathbf{y})$ (Atick and Redlich 1992). We can nevertheless impose and satisfy the condition of minimum correlation of the output. By minimizing

$$E = \int d\mathbf{x}d\mathbf{y} |\langle O(\mathbf{x})O(\mathbf{y})\rangle - \delta(\mathbf{x},\mathbf{y})|^2 \tag{7}$$

with respect to the matrix $\mathbf{Q}$, it is not difficult to show that $\mathbf{Q}$ must be given by

$$Q_{rs} = \frac{1}{\sqrt{\lambda_r}} U_{rs} \tag{8}$$

where U_{rs} is any orthogonal matrix satisfying $\mathbf{U}^T\mathbf{U} = 1$ and λ_r is the eigenvalue, corresponding to $\Psi_r(\mathbf{x})$.

This transformation is familiar from previous work on the retina (Atick and Redlich 1992). The whitening factor $1/\sqrt{\lambda_r}$ normalizes the PCA output variance (eq. 3) to unity.† The resulting PCA outputs can then be mixed by any orthogonal

† In practice one should also include noise filtering, since the process of whitening by multiplying by $1/\sqrt{\lambda_r}$ amplifies both signal and noise. Whereas the power of the signal decreases with r, the power of the noise remains constant. The whitening factor is most significant in the regime of small values of λ which is precisely where the signal to noise ratio is small. In order to fight noise amplification, one should multiply additionally by a low-pass noise filter, and then the resulting factor will be effectively a band-pass filter in the eigenmode number r, i.e., it attenuates the power for small r (high λ) as well as high r (small λ) and amplifies it for intermediate values of r—very much like the contrast sensitivity curves encountered in earlier work (Atick and Redlich 1992). The exact form of the optimal bandpass filter can be derived only after a specific model of the noise is adopted.

transformation **U** without affecting the degree of their decorrelation. In fact, this symmetry was exploited in previous work to produce representations that, without destroying decorrelation, possess other desirable properties—for example, scale invariance, which leads to a multi-scale representation (Li and Atick 1994). We will keep the existence of this degree of freedom in mind for future reference, but in the current analysis we will make the simplest choice $U_{rs} = \delta_{rs}$. This fixing of the unitary symmetry was derived in a previous work on the retina (Atick et al. 1993) by applying the criterion of minimal distortion from input to output, where it was shown to generate local receptive fields in the case $N = V$.

With this choice for **Q** (eq. 8), the LFA outputs become

$$O(\mathbf{x}) = \int \sum_{r=1}^{N} \Psi_r(\mathbf{x}) \frac{1}{\sqrt{\lambda_r}} \Psi_r(\mathbf{y}) \phi(\mathbf{y}) = \sum_{r=1}^{N} \frac{A_r}{\sqrt{\lambda_r}} \Psi_r(\mathbf{x}) \tag{9}$$

and their residual correlation can be computed easily using the orthonormality of the modes $\Psi_r(\mathbf{x})$:

$$\langle O(\mathbf{x}) O(\mathbf{y}) \rangle = \sum_{r=1}^{N} \Psi_r(\mathbf{x}) \Psi_r(\mathbf{y}) \equiv P(\mathbf{x}, \mathbf{y}). \tag{10}$$

The function $P(\mathbf{x}, \mathbf{y})$ is an interesting object; it can be readily recognized as the projection operator onto the subspace spanned by the PCA eigenmodes or (in the case $T = N$) the subspace spanned by the examples used to derive the modes. On that subspace $P(\mathbf{x}, \mathbf{y})$ acts as the identity operator, i.e., $\int P(\mathbf{x}, \mathbf{y}) \phi^t(\mathbf{y}) = \phi^t(\mathbf{x})$. In the extreme limit that the modes $\Psi_r(\mathbf{x})$ span the complete input space ($N \to V$), any input can be accurately represented as an expansion in them. Then $P(\mathbf{x}, \mathbf{y}) \to \delta(\mathbf{x}, \mathbf{y})$ and complete decorrelation of the outputs is achieved, as in the case of the retinal (Atick and Redlich 1992) and the LGN (Dong and Atick 1995b; Dan et al. 1996) analysis.

To summarize, given the eigenmodes $\Psi_r(\mathbf{x})$ with eigenvalues λ_r we can construct the following two functions:

$$\begin{aligned} K(\mathbf{x}, \mathbf{y}) &= \sum_{r=1}^{N} \Psi_r(\mathbf{x}) \frac{1}{\sqrt{\lambda_r}} \Psi_r(\mathbf{y}) \\ P(\mathbf{x}, \mathbf{y}) &= \sum_{r=1}^{N} \Psi_r(\mathbf{x}) \Psi_r(\mathbf{y}). \end{aligned} \tag{11}$$

$K(\mathbf{x}, \mathbf{y})$ is the kernel of the representation, and $P(\mathbf{x}, \mathbf{y})$ turns out to be the residual correlation of the outputs.

The output array $\{O(\mathbf{x})\}$ preserves all the information in the global modes A_r. In fact, by acting with $\Psi_r(\mathbf{x})$ on (eq. 9), one can derive the reconstruction:

$$A_r = \int \sqrt{\lambda_r} \Psi_r(\mathbf{x}) O(\mathbf{x}). \tag{12}$$

To reconstruct the example directly from the output $\{O(\mathbf{x})\}$ we substitute (eq. 12) in the reconstruction formula (eq. 4)

$$\phi^{rec}(\mathbf{x}) = \sum_{r=1}^{N} \int \sqrt{\lambda_r} \Psi_r(\mathbf{y}) O(\mathbf{y}) \Psi_r(\mathbf{x}) = \int K^{(-1)}(\mathbf{x}, \mathbf{y}) O(\mathbf{y}) \tag{13}$$

where the "inverse" kernel is†

$$K^{(-1)}(\mathbf{x},\mathbf{y}) = \sum_{r=1}^{N} \Psi_r(\mathbf{x})\sqrt{\lambda_r}\Psi_r(\mathbf{y}). \tag{14}$$

The reconstruction error $\langle\|\phi - \phi^{rec}\|^2\rangle$ for the LFA representation is exactly equal to that for the PCA representation, and it is given by $\langle\|\phi_-(\mathbf{x})\|^2\rangle$ where $\phi_-(\mathbf{x})$ is the part of $\phi(\mathbf{x})$ that is orthogonal to the subspace spanned by the eigenmodes. This means that the topographic LFA representation has the same best reconstruction, generalization, and object constancy properties as the global nontopographic PCA one.

The $O(\mathbf{x})$ representation (eq. 9) is a general case of the previous work on optimal retinal coding (Atick and Redlich 1992). In the special case of the retina, the input ensemble is assumed to be translationally and rotationally invariant. The eigenfunctions of its correlation matrix are then the Fourier modes $\Psi_{\mathbf{f}}(\mathbf{x}) = \exp(i\mathbf{f}\cdot\mathbf{x})$, and the eigenvalues are given by the measured power spectrum of natural scenes $1/|\mathbf{f}|^2$. In that case $K(\mathbf{x},\mathbf{y}) = \int d\mathbf{f}|\mathbf{f}|\exp(i\mathbf{f}\cdot(\mathbf{x}-\mathbf{y}))$, which, it is easy to verify, is a local center-surround type kernel, with $P(\mathbf{x},\mathbf{y}) = \delta(\mathbf{x},\mathbf{y})$.

Next we explore what $K(\mathbf{x},\mathbf{y})$ and $P(\mathbf{x},\mathbf{y})$ turn out to be for the more interesting case of object ensembles. The results in Figure 1 show the $\mathbf{K}$ and $\mathbf{P}$ derived for Ensemble 1—2D images of human heads. Analogous results for Ensemble 2—3D surfaces of human heads—are shown on Figure 2. Since the results exhibit the same properties, we will discuss them together.

As we can see, the receptive fields develop compact support and are local. They are also strongly matched to the local features of the face. For example, a receptive field matched to a mouth develops at position $1a$, a nose receptive field—at position $1b$, and eyebrow, jaw-line, and cheek-bone receptive fields—at positions $1c$, $1d$ and $1e$, respectively. The same results—local feature receptive fields (for nose, forehead, eye, jaw-line, and cheekbone in $2a$, $2b$, $2c$, $2d$, and $2e$, respectively)—are observed for input Ensemble 2. We should note that these are two very different input receptor spaces; the first is intensity samplings of (logarithmically gain controlled) photographic images of naturally rendered heads, the second is the radii (in millimeters) of surfaces of heads before rendering and with no albedo. The fact that they develop conceptually similar receptive fields supports the theoretical understanding that LFA captures the underlying structure of the input ensemble probability density, regardless of what the ensemble or the receptor space happens to be.

Note that the receptive fields that develop are not edge detectors in general; they are feature detectors, different from each other, and matched to the feature that is expected near their respective centers. Note also that the receptive fields have captured a correct symmetry—strong at the eyes, eyebrows and cheeks—which reflects the bilateral symmetry of human faces, and nonexistent at the outlines, which reflects the pose variability in the input ensembles. The symmetry is greater and the receptive fields are more sharply defined in Figure 2, because the input ensemble is better aligned and has less extrinsic variability.

† One can define a family of functions $K^{(n)}(\mathbf{x},\mathbf{y}) = \sum_{r=1}^{N} \Psi_r(\mathbf{x})\left(\frac{1}{\sqrt{\lambda_r}}\right)^n \Psi_r(\mathbf{y})$ labeled by the integer n. $K^{(1)}(\mathbf{x},\mathbf{y}) \equiv K(\mathbf{x},\mathbf{y})$ is the kernel of the LFA representation; $K^{(0)}(\mathbf{x},\mathbf{y}) \equiv P(\mathbf{x},\mathbf{y})$ is the residual correlation of the output array $\{O(\mathbf{x})\}$; $K^{(-1)}(\mathbf{x},\mathbf{y})$ is the reconstructor to the original example, while everything is derived from $K^{(-2)}(\mathbf{x},\mathbf{y}) \equiv R(\mathbf{x},\mathbf{y})$, which is the correlation function of the input ensemble.

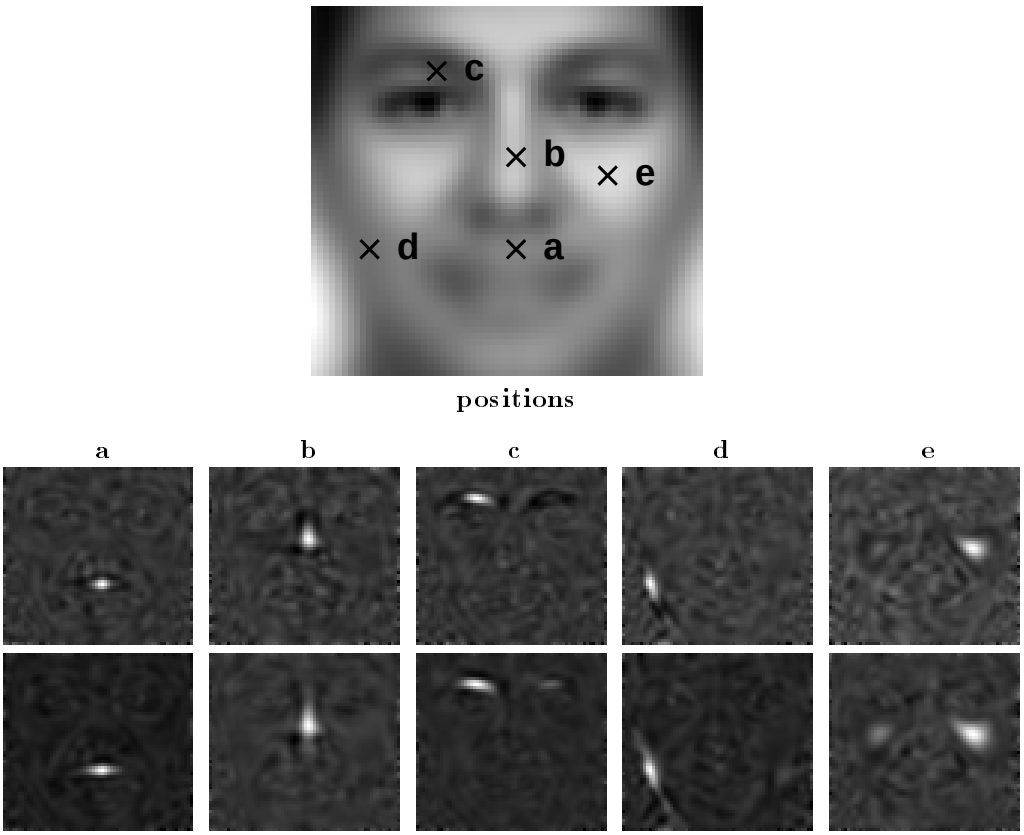


Figure 1. $K(\mathbf{x},\mathbf{y})$ and $P(\mathbf{x},\mathbf{y})$ for images of faces
positions: The average of the examples in Ensemble 1, described in Appendix A, is marked with the respective positions $\mathbf{x}_0$ of the distributions shown on a—e.
a—e: Five kernels $K(\mathbf{x}_0,\mathbf{y})$ (top row) and correlators $P(\mathbf{x}_0,\mathbf{y})$ (bottom row) for the same ensemble at five choices for $\mathbf{x}_0$. Lowpass noise filtering is performed with the $F_r = \frac{\lambda_r}{(\lambda_r+n^2)}$ so that $K(\mathbf{x},\mathbf{y}) = \sum_{r=1}^{N} \Psi_r(\mathbf{x}) \frac{F_r}{\sqrt{\lambda_r}} \Psi_r(\mathbf{y})$ The parameters are $V = 3840$, $T = 1039$, $N = 400$, and $n = 0.25$ (which results in the peak of the bandpass filter being at around $r = 400$).

The receptive fields for the examined ensembles happen to be mostly local, although locality was not imposed; the only imposed condition was topography, which, along with a simple fixing of the unitary symmetry, allowed the correlation function to manifest its local structure in the locality of the kernels. Indeed, wherever the correlations are not local—as in the places of partial bilateral symmetry—the receptive fields turn out to be nonlocal as well.

To illustrate the reconstruction power of LFA we have applied it to the representation of an out-of-sample example (Example 1 of Appendix A) shown in Fig. 3. This image was captured with a camcorder (linear gain control) as opposed to the photographic images (logarithmic gain control) of the ensemble used in deriving the representation. Also, the lighting conditions and the backgrounds are very different. We note that the representation $O(\mathbf{x})$, shown in (b), is very different from that given

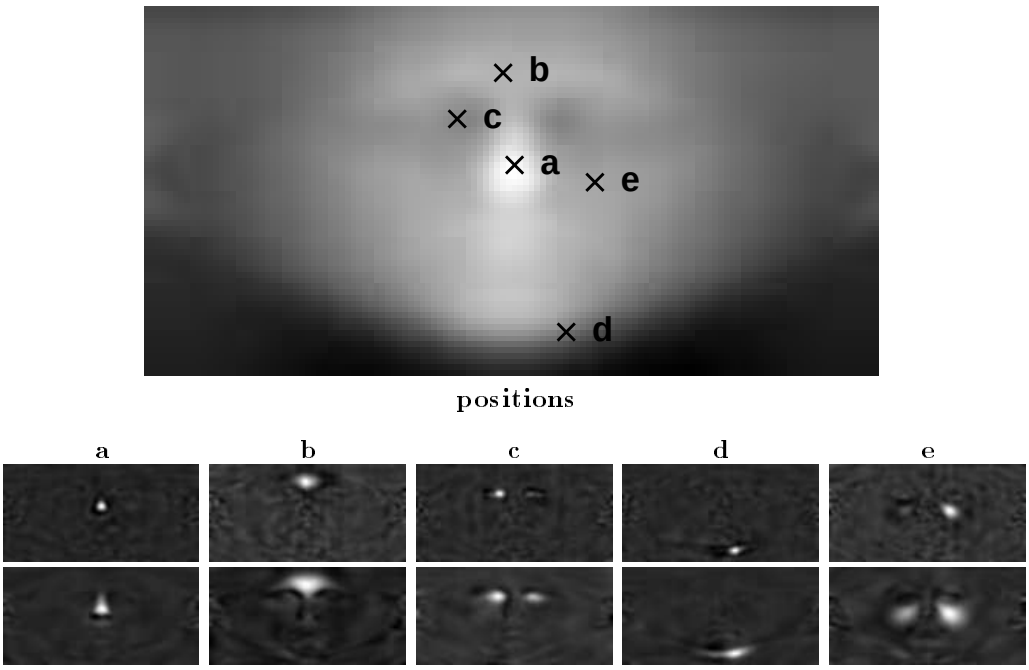


Figure 2. $K(\mathbf{x},\mathbf{y})$ and $P(\mathbf{x},\mathbf{y})$ for surfaces of heads positions: The average of the examples in Ensemble 2, described in Appendix A, is marked with the respective positions $\mathbf{x}_0$ of the distributions shown on a—e.
a—e: Five kernels $K(\mathbf{x}_0,\mathbf{y})$ (top row) and correlators $P(\mathbf{x}_0,\mathbf{y})$ (bottom row) for the same ensemble at five choices for $\mathbf{x}_0$. The parameters are $V = 8192$, $T = N = 348$, and $n = 0$ (see Fig. 1).

by edge detectors (for example, Canny 1986). It is active at all places where the image deviates from our expectation of faces, not only at the edges, thus revealing the face-specific features of the example. $O(\mathbf{x})$ is a representation in the sense that we can reconstruct the original example using (eq. 13). When we examine the reconstruction $\phi^{rec}(\mathbf{x})$ in (c), we note that it is a filtered version of the original $\phi(\mathbf{x})$ in (a) that preserves the identity of the subject, which makes $O(\mathbf{x})$ very interesting for applications like face recognition. By looking at the error $\phi_-(\mathbf{x})$ in (d) we observe that it is small and is distributed roughly evenly throughout the image, without any specific structure, which supports the theoretical expectation for object constancy.

3. Sparse-distributed from topographic representations

In the previous section we overcame the main shortcoming of the PCA representation (eq. 2) by developing LFA (eq. 11). In the process, we obscured the low-dimensionality of the output and introduced residual correlations. In this section we show how to use those residual correlations to define lateral inhibition, which sparsifies the output, recovering the low-dimensionality of the representation.

One might be tempted to apply LFA again, this time on $O(\mathbf{x})$ instead of on $\phi(\mathbf{x})$. By noting that the correlation function of the output $P(\mathbf{x},\mathbf{y}) = \sum_{r=1}^{N} \Psi_r(\mathbf{x}) \cdot 1 \cdot \Psi_r(\mathbf{y})$

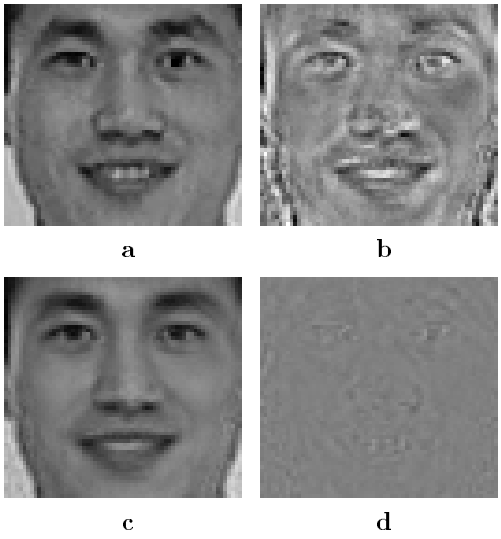


Figure 3. Reconstruction with the LFA representation $O(\mathbf{x})$
a: $\phi(\mathbf{x})$ of Example 1. b: its $O(\mathbf{x})$ calculated with the LFA kernel, derived from Ensemble 1, $N = 400$, $n = 0.25$ (see Fig. 1). c: the reconstruction $\phi^{rec}(\mathbf{x})$, using the shown $O(\mathbf{x})$. d: the reconstruction error $\phi_{-}(\mathbf{x})$, shown on the scale of $\phi(\mathbf{x})$, with the 0 shifted to gray.

(eq. 10, compare with eq. 1) is already in diagonalized form and all its eigenvalues are unity, one concludes that everything possible was done on the first step, and subsequent applications would be trivial.

One characteristic difference between the correlation functions of the input and the output $R(\mathbf{x}, \mathbf{y})$ and $P(\mathbf{x}, \mathbf{y})$, respectively, is that the former is global, in contrast with the latter, which is local (see bottom rows of Fig. 1 and Fig. 2). Thus, the outputs in a local region are correlated, signaling a feature redundantly, so the one that best describes it can be chosen to remain active and the rest—suppressed. Suppose $O(\mathbf{x}_m)$ is active. Since it is correlated with other outputs via $P(\mathbf{x}, \mathbf{x}_m) \equiv P_m(\mathbf{x})$ we can predict them to some extent and transmit only the error. The optimal predictor is $O^{pred}(\mathbf{x}) = \frac{P_m(\mathbf{x})}{P(\mathbf{x}_m, \mathbf{x}_m)} O(\mathbf{x}_m)$† which shows that each output $O(\mathbf{x}_m)$ predicts a small neighborhood to an extent governed by the support of $P_m(\mathbf{x})$. One possible strategy for sparsification is to represent $O(\mathbf{x})$ with only a small subset of values $\{O(\mathbf{x}_m)\}_{\mathbf{x}_m \in \mathcal{M}}$ where the set of active units $\mathcal{M}$ is chosen so that the supports of the predictors $\propto P_m(\mathbf{x})$ cover the $\mathbf{x}$ space reasonably well. Then the positions of activation will signal the locations of the strongest features in the example.

There are many ways to achieve this type of sparsification—for example, competitive learning or a winner-take-all strategy (Malsburg 1973; Grossberg 1987; Rumelhart and McClelland 1982; McClelland and Rumelhart 1981; Kohonen 1984; Touretzky 1989), as well as including explicit terms in the error function of a neural network (Olshausen and Field 1996). Here we propose a neural network for doing so with feed-forward connections given by the filter $K(\mathbf{x}, \mathbf{y})$, lateral inhibitory

† This is a special case of eq. 18, derived later in this section. With only one selected point $\mathbf{x}_m$, P' is the number $P(\mathbf{x}_m, \mathbf{x}_m)$.

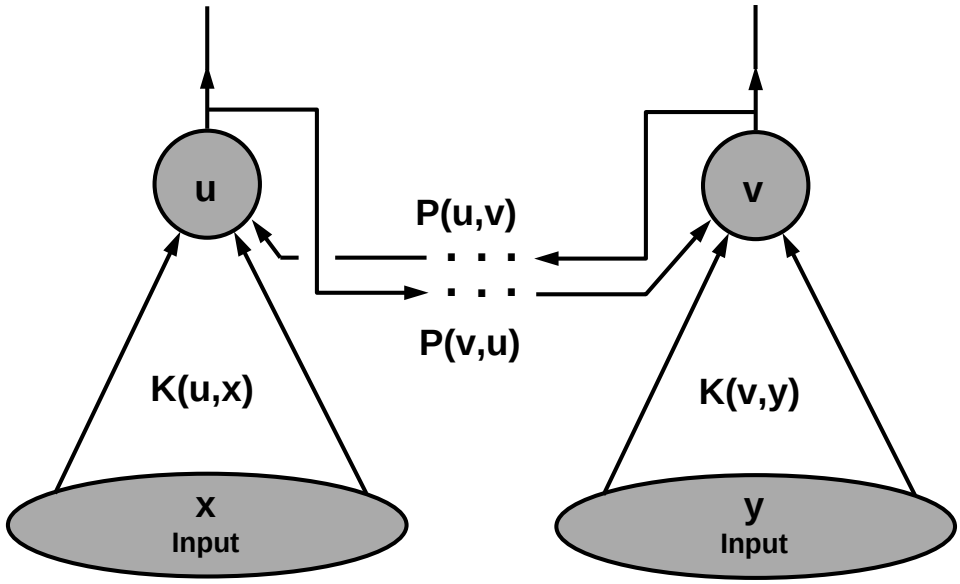


Figure 4. The sparsifying LFA neural network
The units have feedforward receptive fields $K(\mathbf{x}, \mathbf{u})$ that are optimally matched to the object class and decorrelate as much as possible, while obeying the topography. The residual correlations are suppressed through lateral inhibition, proportional to $P(\mathbf{u}, \mathbf{v})$. The output of the network is a sparse-distributed representation of the object, built by the statistically derived local features.

connections given by $P(\mathbf{x}, \mathbf{y})$ (Figure 4), with sigmoidal nonlinear units and standard Hopfield dynamics (Hopfield 1982):

$$\mathcal{O}(\mathbf{x}) = \mathrm{g}\Big(\int K(\mathbf{x}, \mathbf{y})\phi(\mathbf{y}) - \alpha \int P(\mathbf{x}, \mathbf{y})\mathcal{O}(\mathbf{y})\Big). \tag{15}$$

α is the strength of the lateral inhibition and g is an appropriate nonlinear function. Since the lateral connections are symmetric, $P(\mathbf{x}, \mathbf{y}) = P(\mathbf{y}, \mathbf{x})$, this network is guaranteed to converge to a stable state. Such a network was proposed in (Földiak 1990) for forming sparse representations. What we are adding here is the explicit specification of the feed-forward and lateral connections as the $K(\mathbf{x}, \mathbf{y})$ and $P(\mathbf{x}, \mathbf{y})$ functions given in eq. 11.

The interest in networks of this type goes beyond their ability to produce sparsification. Their underlying architecture resembles that of the prototypical cortical circuitry and, hence, they could be a biologically plausible model for cortical coding. It would be interesting to try to see if the relationships predicted by the theory between lateral connections, feed-forward connections and the statistics of the input are realized in cortical circuitry.

In general, the relative strength of the inhibition α and the nonlinear function g should be chosen according to some optimality principle (e.g., histogram equalization) given some transmission criteria or constraints such as bandwidth and fidelity. For our purposes this will not be required. It is sufficient to note that in the limit of hard-core rectification the steady state $\mathcal{O}(\mathbf{x})$ of the network (eq. 15) is ON at the locations,

where the output is unsuppressed and *OFF*, where it is suppressed. We interpret the set of *ON* units as the set of active units $\mathcal{M}$ defined earlier and choose to represent $O(\mathbf{x})$ (eq. 9) with the values at these locations $\{O(\mathbf{x}_m)\}_{\mathbf{x}_m \in \mathcal{M}}$. The relative amount of the active units is governed by the strength of the lateral inhibition α. We choose it so that their number is $|\mathcal{M}| \approx N$.

Instead of simulating the dynamics of the proposed network on serial computers, we give a deterministic algorithm for incrementally producing (in no more than N steps) a set $\mathcal{M}$, that we believe is close to the steady state solution for the network (eq. 15).

We start with the empty set $\mathcal{M}^{(0)} = \emptyset$ and at each step add a point to $\mathcal{M}$, chosen according to the criterion below. At the m-th step we do two things. First, given the current set $\mathcal{M}^{(m)}$, we attempt to reconstruct $O(\mathbf{x})$ by:

$$O^{rec}(\mathbf{x}) = \sum_{m=1}^{|\mathcal{M}|} a_m(\mathbf{x}) O(\mathbf{x}_m). \tag{16}$$

If we succeed in reconstructing $O(\mathbf{x})$, then by eq. 13 we can reconstruct the example $\phi(\mathbf{x})$. The optimal linear prediction coefficients $a_m(\mathbf{x})$, chosen to minimize the average reconstruction m.s.e. on $O(\mathbf{x})$

$$E = \langle \|O^{err}(\mathbf{x})\|^2 \rangle \equiv \langle \|O(\mathbf{x}) - O^{rec}(\mathbf{x})\|^2 \rangle \tag{17}$$

are

$$a_m(\mathbf{x}) = \sum_{l=1}^{|\mathcal{M}|} P(\mathbf{x}, \mathbf{x}_l)(P'^{-1})_{lm} \text{ with } P(\mathbf{x}_l, \mathbf{x}_m) \equiv P'_{lm}. \tag{18}$$

Then, on the $m+1$-st step we choose the point that has the maximum reconstruction error $O^{err}(\mathbf{x}_{m+1})$ and add it to $\mathcal{M}$.† We keep adding points to $\mathcal{M}$ until the reconstruction error goes below some acceptable level, or until we choose N of them.‡

In trying to reconstruct $O(\mathbf{x})$, and therefore A_r (eq. 12), from a limited set of values $O_m \equiv O(\mathbf{x}_m)$, one is solving the set of $|\mathcal{M}|$ equations $\sum_{r=1}^{N} \mathbf{\Psi}_{rm} A_r = O_m$, where $\mathbf{\Psi}_{rm} \equiv \Psi_r(\mathbf{x}_m)$ (cf. eq. 9). When $N = |\mathcal{M}|$, $\mathbf{\Psi}_{rm}$ is a square matrix and is, in general, invertible. Therefore, after picking N points, the entire $O(\mathbf{x})$ should be reconstructed without error.

Even though in principle any N points should be sufficient to recover $O(\mathbf{x})$, there are practical considerations in favor of their judicious choice. For example, if one chooses points that are too close to each other—in the sense that their $P(\mathbf{x}, \mathbf{y})$ have significantly overlapping supports—the values $\{O(\mathbf{x}_m)\}_{\mathbf{x}_m \in \mathcal{M}}$ will be correlated and the representation will be redundant. Moreover, the useful information in them will be carried by the least-significant digits in their representation, requiring it to have extremely high precision. This is wasteful in principle, and impractical in biology,

† There are various criteria one could employ in choosing the next point to add to the mask $\mathcal{M}$. For example, instead of picking the point with the biggest current error, one could pick the point that would achieve the lowest total error on the next step. Unfortunately, this needs a computation on the order of V on each step, and in the light of the speed and the robustness of the simpler criterion, we chose not to implement it.

‡ Note that since the algorithm is incremental we never need to compute the inverse of the matrix P' (eq. 18) explicitly. The inverse at the m-th step is related to the inverse at the $m-1$-st step through a simple algebraic formula. The algorithm for inversion through partitioning is available in many books on numerical methods, for example, (Noble 1992; Press et al. 1992).

where the available S/N ratio of the noisy neurons cannot be pushed very far. The described sparsification algorithm chooses points whose outputs are not predicted well by the already chosen ones, which creates a bias towards allocating resources efficiently.

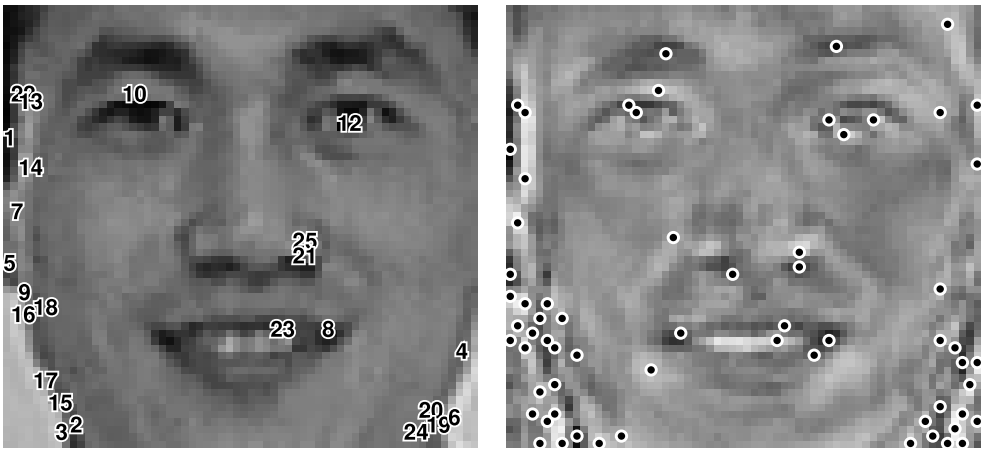


Figure 5. Resource allocation by the sparsification algorithm
The building of the mask $\mathcal{M}$ for Example 1, reconstructed with the kernel from Ensemble 1. $N = 220$, $n = 0.25$ (see Fig. 1). a: the first 25 points—$\mathcal{M}^{(25)}$, overlayed on $\phi(\mathbf{x})$ and numbered sequentially. b: the points in $\mathcal{M}^{(64)}$. The reconstruction with these points is shown on the top row of Fig. 6*b*.

An example of such resource allocation—the result of applying the algorithm to Example 1—is shown on Figure 5. The location and the order of first 25 points—those in $\mathcal{M}^{(25)}$ (*a*) show that resources are allocated first at the places with biggest deviations from the expectation—the outlines of the face and the most unusual features—which is a desirable property. The points in $\mathcal{M}^{(64)}$—enough for acceptable reconstruction of Example 1 (see the top row of Fig. 6*b*)—are shown with dots in (*b*). We observe that only a few values are needed to represent each individual feature, which is a result of the generalization properties and the locality of the representation. On the other hand, the fact that we are trying to describe the small amount of present background with our knowledge of faces, leads to the expected elevated density of the allocated resources in those regions. In the discussion section we will give an idea of how to take advantage of this by cascading at least two such representations, each with knowledge of a different correlation function.

In Figure 6 we compare the quality of reconstruction using the sparse topographic representation (*b*) with two other strategies—the global PCA representation (*a*) and a uniform subsampling (*c*), when all of them use the same number of values—64. We calculated the PCA coefficients A_r (eq. 2) and the LFA ones $O(\mathbf{x})$ (eq. 9) for $N = 220$, then sparsified to get $\{O(\mathbf{x}_m)\}_{\mathbf{x}_m \in \mathcal{M}}$. We calculated the reconstruction $\phi^{rec}(\mathbf{x})$ either (*a*), using eq. 4 with the first 64 PCA coefficients, or (*b*), using eq. 13 with $O^{rec}(\mathbf{x})$ derived from eq. 16 with the points in $\mathcal{M}^{(64)}$. For comparison, we subsampled $\phi(\mathbf{x})$ on a uniform 8×8 grid and reconstructed it with those 64 points (*c*). In all cases we reconstructed about Ψ_1—the average of the ensemble; we kept

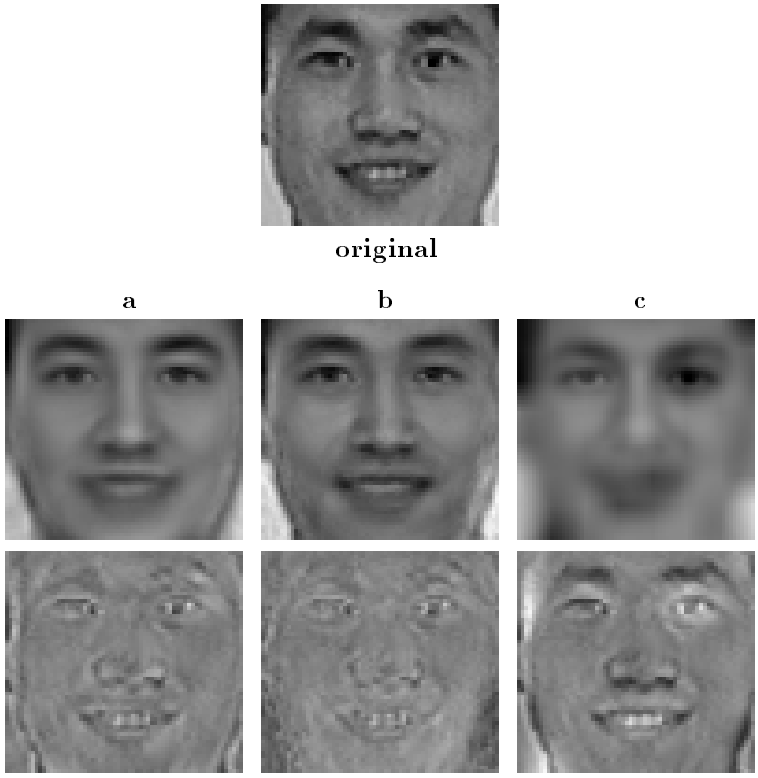


Figure 6. Reconstruction with a fixed number of values
The reconstruction (top row) and the 2× magnified error (bottom row) for $\phi(\mathbf{x})$ of Example 1 (*original*) in the context of Ensemble 1. Reconstruction in all cases is about the average head Ψ_1. *a*: using the first 64 PCA coefficients. *b*: using an approximation of $O(\mathbf{x})$ ($N = 220$, $n = 0.25$) with $\mathcal{M}^{(64)}$, shown on fig. 5*b*. *c*: using uniform subsampling on a 8×8 grid (64 points).

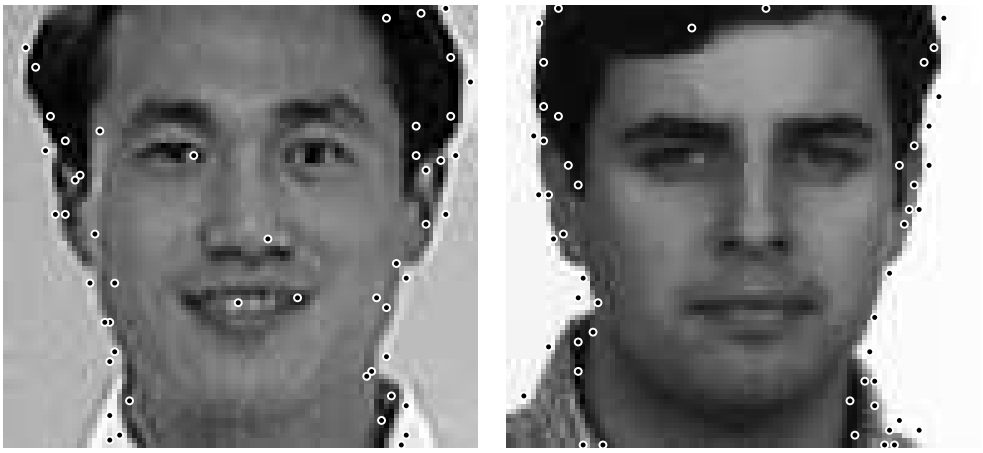
A_1 fixed. Looking at the results, we can see that the perceptual quality of the reconstruction $\phi^{rec}(\mathbf{x})$ (top row) is the best, and the error $\phi(\mathbf{x}) - \phi^{rec}(\mathbf{x})$ (bottom row) contains the least identity information for the LFA representation (*b*).† This makes the sparse-distributed representation a promising candidate for practical applications like compression and object recognition. Indeed, we have found this type of representation very powerful in our related work on face recognition.

The representation produced by the described sparsification algorithm based on reconstructing $O(\mathbf{x})$ using $P(\mathbf{x}, \mathbf{y})$, contains all of the information in $O(\mathbf{x})$ and so possesses all of its desirable properties—best reconstruction, generalization and object

† One would like to know how the perceptually better sparsified LFA representation compares in terms of m.s.e. with the global PCA one. The m.s.e. in Fig. 6 is 184, 227, and 508 for (*a*), (*b*), and (*c*), respectively, out of 560 total power in $\left(\phi(\mathbf{x}) - A_1\Psi_1\right)$. If we look at the error on (*b*), we see that there is a global component in it—positive in the center and negative near the borders; it is very much like an error in a strong global mode, which we identify as Ψ_2 on Fig. A2. Indeed, the m.s.e. due to the error in A_2 in (*b*) is 75. If one fixes not only A_1, but also A_2, the m.s.e. of the sparsified LFA is 152, compared to 184 for PCA. This suggests that from m.s.e. point of view the optimal representation is a hybrid between PCA and LFA—an idea which we exploit in Appendix B.

constancy. In addition, it is sparse distributed, instead of dense, which reveals the low dimensionality of the object space.

The algorithm works extremely robustly in practice; we have produced sparse-distributed representations for all ensembles described previously in the paper with great success. In all cases $O(\mathbf{x})$ was recovered practically to machine precision after choosing N points. In Appendix C we speculate where the observed numerical stability could be coming from.



a **b**

Figure 7. Segmentation by sparsification
$\mathcal{M}^{(50)}$ for Example 1 (a) and Example 2 (b), for the kernel of Ensemble 3 with $V = 8640$, $T = 1039$, $N = 400$, $n = 0.25$ (see Fig. 1).

The application of the sparsification algorithm to Ensemble 3 (see Appendix A) is shown on Fig. 7. Since the ensemble consists of 2D images of heads on a background, the most prominent "features" are the transitions from the background to the head. By looking at the points in $\mathcal{M}^{(50)}$ we see that they are almost exclusively at the boundaries of the heads, and that they pinpoint them with great precision. This suggests using roughly aligned ensembles for fine segmentation, and subsequently feeding the resulting separated objects to a high-fidelity LFA module for feature extraction.

Finally, we should point out that sparse representations have been argued to have some desirable properties, and that sparseness has been previously postulated as a design principle for visual coding by many groups (Barlow 1972; Palm 1980; Barlow 1985; Baum et al. 1988; Zetzcshe 1990; Field 1994; Olshausen and Field 1996). In the current work we show that sparsification is required to unveil the low-dimensionality of the object ensemble, temporarily obscured in the dense output of the localized receptive fields.

4. Discussion

In this paper we have shown how to overcome the main limitation of the PCA representation (eq. 2) by developing LFA (eq. 11)—a method for deriving a dense set of *local topographic feed-forward* receptive fields, defined at each point of the receptor

grid $\mathbf{x}$, different from each other, *optimally matched* to the second-order statistics of the input ensemble, and having outputs as decorrelated as possible. Then we used the residual correlations to define lateral inhibition which *sparsifies* the output, recovering the *low-dimensionality*, and further *decorrelating* the representation. The resulting receptive fields are matched to the structures that are expected at their respective centers, thus giving a representation of the objects in terms of their local *features*.

It has been generally believed that second-order statistics cannot capture the local spatio-temporal correlations, which we think of as "features;" therefore, higher-order correlations should be used to derive them. Nevertheless, LFA—a purely second-order method—discovers a description of any class of objects in terms of statistically derived local features. So how is this possible?

Here we need to dispel some common confusions. In thinking about the relative importance of second-order versus higher-order statistics, one should not neglect the fact that implicit in the definition of the correlation function $R(\mathbf{x}, \mathbf{y}) \equiv \langle \phi^t(\mathbf{x}) \phi^t(\mathbf{y}) \rangle$ is the choice of ensemble over which the averaging is performed—R is an ensemble property. The selection of the members of an ensemble—the so-called sampling, or categorization process—affects what R is and what information it captures.

For example, if the ensemble contains all natural images without any special alignment or selection, then the averaging will include images with arbitrary numbers of objects at different positions and distances. The correlation function R in this case will be translation as well as scale invariant, and for the most part will capture the variance due to the object distribution in the natural world (cf. Ruderman 1996) and will not carry any information about particular objects and their features. This is the autocorrelation function of natural scenes that has been measured by several groups (Burton and Moorhead 1987; Field 1987; Tolhurst et al. 1992; Hancock et al. 1992; Ruderman and Bialek 1994; Ruderman 1994b; Dong and Atick 1995a), and it is the quantity that seems to dictate the coding of the retina (Atick and Redlich 1992) and the LGN (Dong and Atick 1995b; Dan et al. 1996).

On the other hand, suppose we accept that natural images, at some level, can be classified into different categories depending on what objects they contain.† Then the correlation function for every category will be different from the one computed by averaging over all images. In this case second-order statistics can convey significant information about objects in the given class, carrying what would have been carried by higher-order statistics in the full ensemble. Breaking any symmetry of the input ensemble, or categorization (including correct alignment), is the nonlinear step that, among other things, shifts the information from higher-order order statistics in the full ensemble to second-order in the restricted one.

Of course, one may say that it is precisely the higher-order statistics that dictate how to categorize. While this may be true in principle, in practice explicit knowledge of high-order statistics may not be required to achieve categorization. It may be that through evolution the brain has discovered more efficient ways—supervised or unsupervised—for organizing objects into categories. Causal associations such as reward and punishment, feedback from the environment or from success and failure, and other signals not intrinsic to the objects themselves, could be useful in categorizing

† There are several lines of evidence that there are object-class specific functional areas, face-specific among others, in the cortex of primates (Nachson 1995)—MRI of cortex activity in humans (Allison et al. 1994), extra-striate cortex neurophysiology in macaque (Desimone 1991; Gross 1992; Rolls 1992; Perrett et al. 1992), and face recognition impairment, prosopagnosia among others, in humans (Young 1992).

signals.

Furthermore, any categorization algorithm needs not be a final one; it suffices to be a hypothesis generator. Then a powerful measure for hypothesis testing is the pixel entropy of the resulting representation. This is a completely local quantity, trivially computed from the output distribution of single units. As is well known, minimal pixel entropy for the output units in an information preserving representation, guarantees that the code is close to factorial—one in which the elements are statistically independent and, hence, constitute a compact vocabulary for representing objects in that class. A correct categorization is one with a lower pixel entropy at a given reconstruction fidelity level.

We should note that a system that decorrelates on the basis of the total correlation function† will have output units that redevelop correlations when their activity is monitored over a restricted class of examples. Instead of trying to apply LFA on the pixel representation of these examples as we did here ‡, in general one should work directly on the output of the lower level of processing. This output may serve as a blackboard, through which LFA modules for different categories compete for the current example. Each module would send its reconstruction to the blackboard, which would calculate the residual error and present it for analysis to the other modules. The computational loop would continue until every aspect of the image is analyzed and represented using the existing LFA categories. This architecture would be a concrete implementation of the suggestion for the computational architecture of the thalamo-cortical complex of Mumford (1991, 1992).

One of the main objectives of this paper was to find a representation that is easily amenable to further processing. The sparse-distributed LFA representation has that property, since it highlights the location of the features in the objects. This information can be subsequently used to align the features and treat them as new object ensembles with their own correlation functions, which leads to a nontrivial cascading of LFA modules. We propose this as a basis for a multi-stage hierarchical information processing system, not entirely unlike the primate cortex.

Acknowledgments

This work is supported in part by a grant from the Office of Naval Research contract number N00014-95-1-0381 and by a grant from the W. M. Keck Foundation. We wish to thank Paul Griffin and Norman Redlich for useful conversations.

Appendix A. The databases

This Appendix gives details of the three object ensembles, as well as the test samples not in them, used to produce the results in the paper.

Ensemble 1 The examples in this ensemble are part of the ARPA/ARL FERET database, which consists of grayscale photographic images of a racially diverse set of human subjects in natural conditions on a plain background. The lighting is roughly

† There is good evidence that the visual system is conserved only with the most general correlation function at the initial stages of the visual processing—the retina (Atick and Redlich 1992) and the LGN (Dong and Atick 1995b; Dan et al. 1996).

‡ In this paper we wanted to illustrate the idea of LFA so it was not strictly necessary to concern ourselves with any other correlation function but that of the ensemble we were studying.

diffuse with a single Lambertian source. No attempts have been made to control either the direction or the strength of the source or the expression and the facial hair of the subjects. The photographs have been taken on different days over a six-month period and at greatly varying distances. Most of the subjects appear four times in the database with a couple of months between the pairs of photographs, although there are some subjects that appear only once. The photographs have been scanned (and some unknown to us gain control has been applied in the process) to produce an 8-bit grayscale format with 256×384 samples.

We selected from the FERET database $T = 1039$ examples without glasses and in relatively frontal poses as part of our data set and we kept the remaining 7 for out-of-sample but in-database testing.

For each example the locations of both eyes were selected manually. The examples were then rotated so that the inter-eye line is horizontal and scaled down (with smoothing based on the scale factor) so that the inter-eye distance is 28 pixels. The fixed point of the examples was then set to be the middle of that line. Finally, the examples in the data set were cropped through a 64×60 window centered horizontally about the fixed point and starting 15 rows above it.

Ensemble 2 The examples of this ensemble are part of the U.S. Air Force *Mini Survey* database (Robinette and Whitestone 1992) which consists of 348 laser scans of the 3D surfaces of heads of human subjects. The data samples consist of a representation of the surface in uniformly sampled polar coordinates $r(\theta, \ell)$ with 512 samples spanning a full revolution in the θ direction and 256 samples in the ℓ direction. Anthropological landmarks on the surfaces were selected manually.

The examples were aligned by us, so that the fixed point $(\theta_{fixed}, \ell_{fixed})$ was at the *sellion*—the deepest depression of the vertical center of the nose bone between the eyes. Spikes and missing data points were filled through linear interpolation of known good samples around the patch by an automatic algorithm.

The examples were resampled in a new set of polar coordinates by shifting the vertical axis to pass through the center of masses of five layers up and five layers down the cross-section through the fixed point. The examples were smoothed with a 3×3 Gaussian filter and undersampled twice to produce a 256×128 representation. All examples in the *Mini Survey study* database were chosen to produce a $T = 348$ data set.

Finally, the examples in the data set were cropped through a 128×64 window centered horizontally about the fixed point and starting 17 rows above it.

Ensemble 3: The examples in this ensemble were produced from the data set of Ensemble 1, the only difference being the cropping. The examples were cropped through a 96×90 window centered horizontally about the fixed point and starting 30 rows above it. The images in this ensemble include not only faces but the background as well.

Example 1: This image was taken with a video Hi8 camcorder in almost completely diffuse lighting (there is a weak Lambertian light coming exactly from the left side). The image was captured on a Silicon Graphics *IRIS Indigo 3000* workstation with the *SVideo* capture board. Automatic gain control has been used both on the camcorder and on the capture board.

The image was converted to grayscale, oversampled twice, and then underwent the same procedure of alignment, rotation, and scaling as the examples in Ensembles 1 and 3. Two croppings were produced as described above to be used with the two ensembles respectively.

Example 2: This example is one of the 7 left aside in preparing the data set for Ensembles 1 and 3.

Appendix B. The Hybrid Representation

In this appendix we construct a new representation—a hybrid between PCA and LFA—that could be optimized for balance between a small data rate and good reconstruction fidelity. We offer a qualitative discussion of the merits of the PCA and LFA limits and estimate the position of the crossover regime between the two. We argue that any system operating in practice should be built along the lines of this balance.

In the comparison of the representational power of PCA and LFA with a fixed number of values (Fig. 6), we found that in order to make the perceptually superior LFA have lower m.s.e.† than the PCA, we needed to fix not only A_1, but also A_2. In general, when making measurements, inevitably there are errors, and those induce errors in the reconstruction of the example. Based on the sources of errors, it is sometimes better to measure a PCA coefficient directly, by a convolution with a global receptive field. In the rest of the appendix we show what are the intrinsic factors guiding this choice and estimate where the crossover regime occurs. A rigorous treatment is possible only after a specific model of the sources of errors (noise) is adopted and is out of the scope of this work.

Given a fixed number of eigenmodes N, the best approximation is achieved by ϕ^{rec} from eq. 4. Then the minimum r.m.s. error is achieved by exact knowledge of the coefficients A_r in eq. 4. In practice, we can never estimate the coefficients A_r with infinite accuracy either because of discretization in the coefficients themselves or because of uncertainty in their calculation. The uncertainty Δ_{A_r} in our knowledge of A_r induces an error in the reconstruction which, by eq. 4, is given for each example by

$$\|\Delta_\phi(\mathbf{x})\|^2 = \sum_{r=1}^{N} |\Delta_{A_r}|^2. \tag{B1}$$

Depending on how we estimate the A_r coefficients, the error Δ_{A_r} is different. We can estimate A_r entirely from knowledge of $O(\mathbf{x})$ or from $\phi(\mathbf{x})$. If we estimate A_r entirely from $O(\mathbf{x})$, by using eq. 12 we get $\Delta_{A_r} = \sqrt{\lambda_r}\int_V \Psi_r(\mathbf{x})\Delta_O(\mathbf{x})$. If we make the reasonable assumption that the V values of $\Psi_r(\mathbf{x})$ and $\Delta_O(\mathbf{x})$ are uncorrelated then $\Delta_{A_r} \propto \sqrt{\lambda_r}\sigma(\Delta_O(\mathbf{x}))$, where σ^2 is the variance of the distribution of the error $\Delta_O(\mathbf{x})$. If we define the relative error $\epsilon_o \equiv \frac{\sigma(\Delta_O(\mathbf{x}))}{\sigma(O(\mathbf{x}))}$ to be dimensionless and note that $\sigma_o \equiv \sigma(O(\mathbf{x}))$ depends neither on r nor on ϵ_o, we arrive finally at the estimate of the power of the error in the individual modes:

$$|\Delta_{A_r}|^2 \propto \lambda_r \epsilon_o^2 \sigma_o^2. \tag{B2}$$

† Fidelity of reconstruction in the sense of minimal m.s.e. has been previously used as a design principle by Ruderman to derive linear filters in the context of noisy environments (Ruderman 1994a).

In a real system, especially in noisy neural wetware, ϵ_o cannot be pushed arbitrarily low due to errors from discretization (noise), integration of eq. 9, and ignorance of the true eigenmodes $\Psi_r(\mathbf{x})$ and the true eigenvalues λ_r participating in the calculation. Thus, if we estimate A_r from $O(\mathbf{x})$ only we are bound to make huge errors in the first modes, where λ_r is big (see Fig. A1).

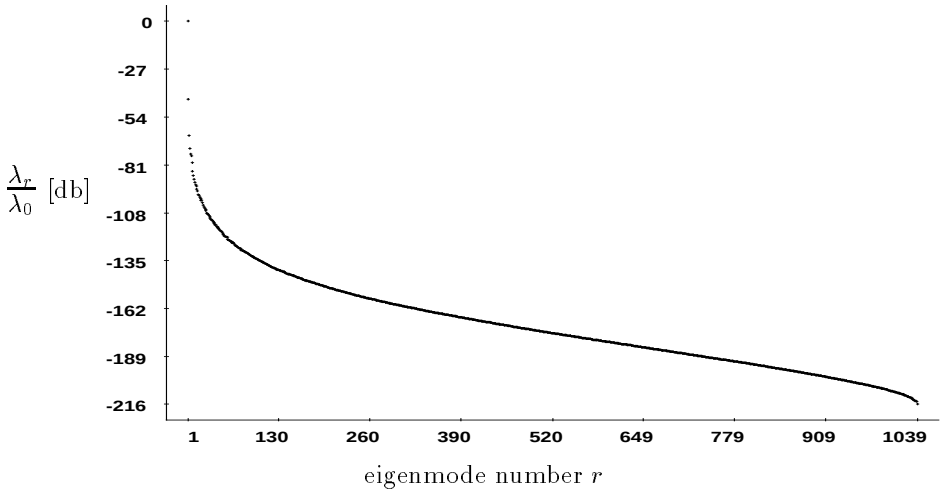


Figure A1. The spectum of the eigenvalues λ_r of an object correlation matrix $R(\mathbf{x}, \mathbf{y})$

λ_r/λ_1 for the correlation matrix $R(\mathbf{x}, \mathbf{y})$ of Ensemble 1 is shown in [db]—i.e. $10 \times \log_2(\lambda_r/\lambda_1)$—versus the eigenmode index r.

If, on the other hand, we estimate the coefficients A_r not from the LFA representation but from the PCA representation (eq. 2), then the error is:

$$|\Delta_{A_r}|^2 \propto \epsilon_\phi^2 \sigma_\phi^2. \tag{B3}$$

Note the absence of the λ_r factor in eq. B3 that is present in eq. B2. As before, ϵ_ϕ cannot be pushed very low, this time mainly due to input noise and ignorance of $\Psi_r(\mathbf{x})$. The difference in the λ_r dependence of eq. B3 and eq. B2 means that it is a better strategy—more robust against noise—to estimate A_r for modes with large λ_r (the first few ones) using PCA, while for higher r, where λ_r is small, it is better to carry out the calculation from the LFA output.

This makes perfect sense; if we look at the first few eigenmodes on Fig. A2 we observe that they are global, integrating, or smoothing filters. These are efficient in suppressing noise both in the input and in themselves (i.e., one can dispense with a low resolution version of them). On the other hand, the higher modes start to become ripply, or differentiating filters, which are likely to amplify noise both in the example and in themselves, and estimating A_r using them quickly becomes disadvantageous to estimating A_r using $O(\mathbf{x})$, because of the noise suppression factor λ_r in eq. B2.

Thus, we expect in general, that a better fidelity representation to require the use of global PCA in one regime and localized LFA in another. This is intuitively correct since there are a few global object attributes like overall scale, global shape, and lighting conditions, which are best captured by a global calculation, and there

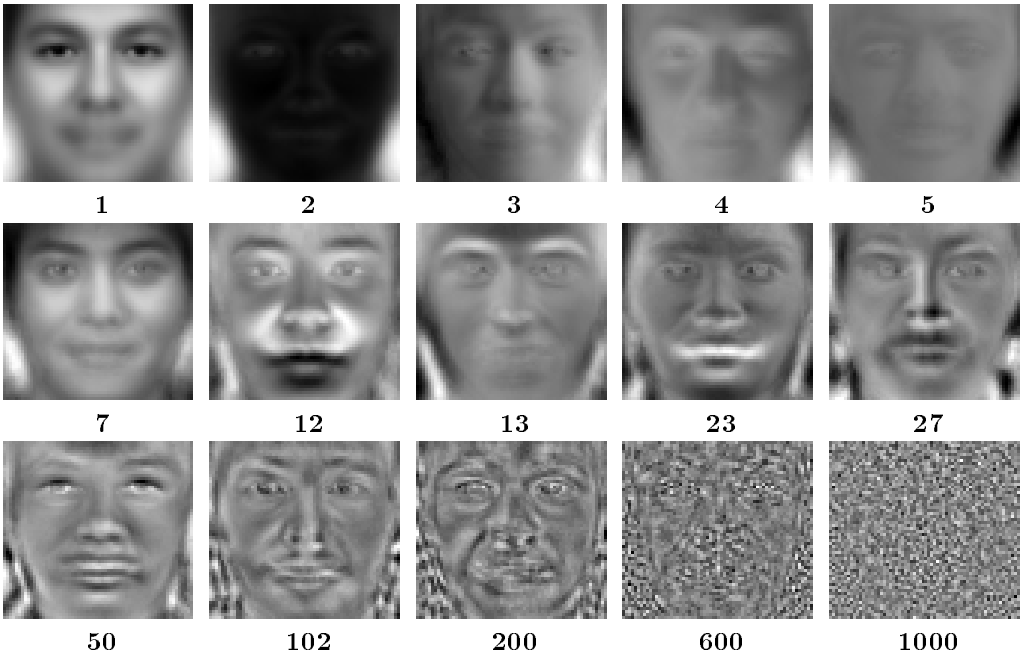


Figure A2. The eigenmodes $\Psi_r(\mathbf{x})$ of an object correlation matrix $R(\mathbf{x},\mathbf{y})$
Some of the eigenmodes $\Psi_r(\mathbf{x})$ of the correlation matrix $R(\mathbf{x},\mathbf{y})$ of Ensemble 1 ($T = 1039$). Each mode is labeled with its index r.

are local features which are best captured by LFA. One can ask when should one stop extracting global attributes and start looking for local features? To answer this question we equate the right sides of eq. B2 and eq. B3. This gives for the transition between PCA and LFA:

$$\frac{\lambda_r \sigma_o^2}{\sigma_\phi^2}\left(\frac{\epsilon_o}{\epsilon_\phi}\right)^2 \propto 1. \tag{B4}$$

This expression can be made more informative if we estimate the rest of the terms above. Since ϵ_o and ϵ_ϕ are dimensionless ratios, we expect both of them to be the same order, so $\left(\frac{\epsilon_o}{\epsilon_\phi}\right)^2 \propto 1$. The variance in $\phi(\mathbf{x})$ itself is of the order of the variance in the average so that $\sigma_\phi^2 \propto \lambda_1$. Next we note that $\sigma_o^2 = \langle O(\mathbf{x})O(\mathbf{x})\rangle$. Using eq. 10, with $\mathbf{x} = \mathbf{y}$, which now is a sum of N positive terms $\Psi_r(\mathbf{x})^2$, we conclude that $\sigma_o^2 \propto N\sigma_\Psi^2$. Remembering that $\Psi_r(\mathbf{x})$ are normalized to unity ($\sigma_\Psi^2 = 1$), and substituting everything in eq. B4, we get

$$\frac{\lambda_r}{\lambda_1} \propto \frac{1}{N}\left(\frac{\epsilon_\phi}{\epsilon_o}\right)^2 \tag{B5}$$

Equation B5 provides an estimate of the mode number r_{cross} where the crossover regime between the PCA and LFA representations occurs. Indeed, when we have a relatively reliable output $O(\mathbf{x})$ (low ϵ_o/ϵ_ϕ, large $\lambda_{r_{cross}}$, low r_{cross}) we should calculate a small number of global PCA coefficients and use the LFA representation for the rest

and vice versa. If we have a small dimensionality N of the object subspace, the PCA coefficients A_r are coded very redundantly by the LFA representation $O(\mathbf{x})$ so we can tolerate bigger errors ϵ_o before we resort to the global modes, and vice versa.

To illustrate the strategy, we show a typical power spectrum—λ_r as a function of r—in Fig. A1. If we choose $N = 400$, as we did in preparing figures 1 and 3, then the transition point happens for $\frac{\lambda_r}{\lambda_1} \propto \frac{1}{400} \approx -85$db. Looking at the figure we find that $r_{cross} \approx 10$. Thus in this case the best strategy is to calculate global coefficients from the first 10 or so modes and then use the LFA modes for the rest. Analogously, if we choose $N = 220$ as in figures 6 and 5, then $r_{cross} \approx 7$. We note by looking at Fig. A1, that the slope of $\log \lambda_r/\lambda_1$ in that regime is big, so that a factor of 2 on either side of eq. B5 leads to inclusion or exclusion of only a very small number of modes.

In general, there should be a smooth transition in the way the A_r coefficients are estimated, pooling the knowledge of the PCA and the LFA representations. Given a set of measured PCA coefficients $\{A_r\}$, a set of LFA values $\{O(\mathbf{x}_m)\}_{\mathbf{x}_m \in \mathcal{M}}$, and a set of relative uncertainties in those values $\{\alpha_r\}$, one could construct a cost function of the type $\langle \|O^{rec}(\mathbf{x}) - O(\mathbf{x})\|^2 \rangle - \sum_{r=1}^{N} \alpha_r \langle |A_r^{rec} - A_r|^2 \rangle$ and derive the optimal reconstructor by varying it with respect to the reconstructing coefficients.

Appendix C. Sparsification before vs. after decorrelation

In the derivation of the sparsification algorithm we didn't use the fact that $P(\mathbf{x}, \mathbf{y})$ is a projector, so one might think that the algorithm is applicable to any input, provided we know its correlation function. More precisely, one might be tempted to sparsify $\phi(\mathbf{x})$, whose correlation function is $R(\mathbf{x}, \mathbf{y})$, without decorrelation first. A priori this may sound like an attractive idea; in practice, however, it does not work. Apart from the computational problems associated with inverting a densely populated matrix $R(\mathbf{x}, \mathbf{y})$ (since it is nonlocal as opposed to the diagonally dominated $P(\mathbf{x}, \mathbf{y})$), sparsification of $\phi(\mathbf{x})$ directly suffers from a severe input noise instability. This noise instability is not encountered in the sparsification of the $O(\mathbf{x})$ since $K(\mathbf{x}, \mathbf{y})$ has noise suppression properties.

To illustrate this we we performed the following experiment. For any given example in the database, we systematically added Gaussian pixel noise with increasing power and studied the reconstruction error for three different strategies: PCA (*diamonds*), sparsification of the LFA output (*crosses*), and sparsification of the input before any preprocessing (*squares*). In Fig. A3 we show the error $\langle \|\phi_0 - \phi_{noise}^{rec}\|^2 \rangle$ for the three different strategies averaged over the database and plotted as a function of the noise power (*solid line*) with ϕ_0—the PCA reconstruction at zero noise.

The first thing to note is that the error for the sparsified $O(\mathbf{x})$ is always about two decades below the actual noise power, while the error for the sparsified $\phi(\mathbf{x})$ is always above the noise. This means that LFA has a powerful object constancy property—the representation changes very little under substantial amount of input noise. On the other hand, the sparsification of $\phi(\mathbf{x})$ behaves extremely poorly; it actually produces an error substantially greater than the added noise. This is understandable, because sparsification is in fact interpolation, which is known to behave poorly on noisy data (as opposed to interpolating a smoothed version of the data, as in the LFA case).

The results for the PCA representation are theoretically equivalent to those for the LFA one and this is seen in the high noise limit, where the two error curves coincide. In the low noise regime, however, we found in practice the noise filtering capabilities of PCA are not as good as those for LFA, which can be seen by the saturation of

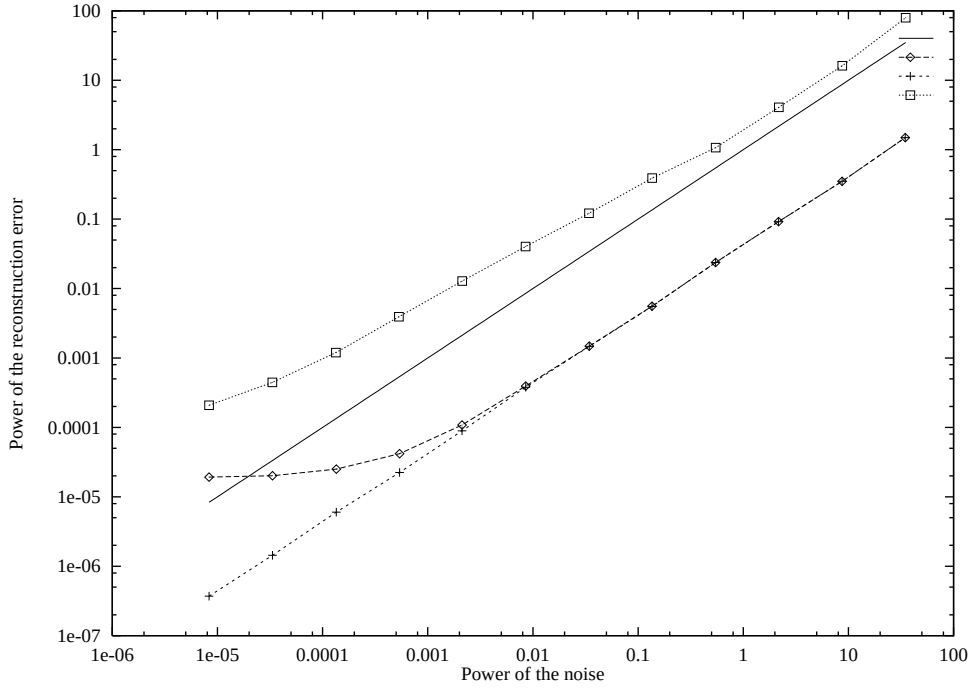


Figure A3. Object constancy in the presence of noise
The power of the object reconstruction error $\langle \|\phi_0 - \phi^{rec}_{noise}\|^2 \rangle$ as a function of the power of the input noise $\langle \|\phi_0 - \phi_{noise}\|^2 \rangle$ under three different strategies. *solid line* : f$(x) = x$, representing the noise itself. *diamonds* : global PCA representation. *crosses* : LFA represenation with sparsification. *squares* : sparcification on the example. See text for details and discussion.

the PCA curve. This might be due to the bigger error stability of the LFA, discussed earlier in Appendix B (eq. B2).

We conclude, that sparseness is only one of the many desirable properties of the LFA representation, and it only works in concert with the others—decorrelation, generalizability, and object constancy.

References

Allison, T., G. McCarthy, A. Nobre, A. Puce, and A. Belger (1994). Human extrastriate visual cortex and the perception of faces, words, numbers, and colors. *Cerebral Cortex 4*(5), 544–54.

Atick, J. J., P. A. Griffin, and N. A. Redlich (1995). Face recognition from live video: Now for real-world applications. *Advanced Imaging 10*(5), 58–62.

Atick, J. J., P. A. Griffin, and N. A. Redlich (1996). Statistical approach to shape-from-shading: deriving 3D face surfaces from single 2D images. *Neural Computation 8*, 1321–40.

Atick, J. J., Z. Li, and A. N. Redlich (1993). What does post-adaptation color appearance reveal about cortical color representation? *Vision Research 33*(1), 123–

9.

Atick, J. J. and N. A. Redlich (1992). What does the retina know about natural scenes? *Neural Computation 4*(2), 196–210.

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual phycology? *Perception 1*(4), 371–394.

Barlow, H. B. (1985). The twelfth Bartlett memorial lecture: The role of single neurons in the phycology of perception. *Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology 37*(2), 121–145.

Barlow, H. B., T. P. Kaushal, and G. J. Mitchison (1989). Finding minimum entropy codes. *Neural Computation 1*(3), 412–423.

Baum, E. B., J. Moody, and F. Wilczek (1988). Internal representation for associative memory. *Biological Cybernetics 59*, 217–228.

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

Bell, A. J. and T. J. Sejnowski (1995). An information maximization approach to blind separation and blind deconvolution. *Neural Computation 7*(6), 1129–1159.

Burton, G. and I. Moorhead (1987). Color and spatial structure in natural scenes. *Applied Optics 26*, 157–170.

Canny, J. F. (1986). A computational approach to edge detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence PAMI-8*(6), 679–698.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing 36*(3), 11–20.

Dan, Y., J. J. Atick, and C. R. Reid (1996). Efficient coding of natural scenes in the lateral geniculate nucleus: Experimental test of a computational theory. *The Journal of Neuroscience 16*(10), 3351–3362.

Deco, G. and D. Obradovic (1996). *An information-theoretic approach to neural computing*. New York: Springer-Verlag.

Deco, G., L. Parra, and S. Miesbach (1995). Redundancy reduction with information-preserving nonlinear maps. *Network: Computation in Neural Systems 6*, 61–72.

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *Journal of cognitive neuroscience 3*(1), 1–8.

Dong, D. W. and J. J. Atick (1995a). Statistics of natural time-varying images. *Network: Computation in Neural Systems 6*(3), 345–358.

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

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A4*, 2379–2394.

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

Földiak, P. (1990). Forming sparse representations by local anti-hebbian learning. *Journal of the Optical Society of America. 64*, 165–170.

Geman, S., E. Bienenstock, and R. Doursat (1992). Neural networks and the bias / variance dilemma. *Neural Computation 4*, 1–58.

Goodall, M. C. (1960). Performance of a stochastic net. *Nature 185*, 557–558.

Gross, C. G. (1992). Representation of visual stimuli in inferior temporal cortex. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences 335*(1273), 3–10.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science 11*, 23–63.

Hancock, P., R. Baddeley, and L. S. Smith (1992). The principal components of natural images. *Network: Computation in Neural Systems 3*, 61–70.

Hentschel, H. and H. H. Barlow (1991). Minimum-entropy coding with hopfield networks. *Network: Computation in Neural Systems 2*(2), 135–148.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America 79*(8), 2554–8.

Intrator, N. (1992). Feature extraction using an unsupervised neural network. *Neural Computation 4*, 98–107.

Jutten, C. and J. Herault (1991). Blind separation of sources. part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing 24*, 1–10.

Kohonen, T. (1984). *Self-Organization and Associative Memory*. Berlin, Heidelberg: Springer-Verlag.

Li, Z. and J. Atick (1994). Towards a theory of the striate cortex. *Neural Computation 6*, 127–146.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer 21*, 105–117.

Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik 14*, 85–100.

McClelland, J. L. and D. E. Rumelhart (1981). An interactive activation model of context effects in letter perception: Part 1. an account of basic findings. *Psychological Review 88*, 375–407.

Mumford, D. (1991). On the computational architecture of the neocortex. I: The role of the thalamo-cortical loop. *Biological Cybernetics 65*(2), 135–145.

Mumford, D. (1992). On the computational architecture of the neocortex. II: The role of cortico-cortical loops. *Biological Cybernetics 66*(3), 241–251.

Nachson, I. (1995). On the modularity of face recognition: The riddle of domain specificity. *Journal of Clinical & Experimental Neuropsychology 17*(2), 256–75.

Noble, J. V. (1992). *Scientific FORTH* (1st ed.)., pp. 86–90. Charlottesville, VA.: Mechum Banks Publishing.

Oja, E. (1989). Principal components and linear neural networks. *Neural networks 5*, 927–935.

Olshausen, B. A. and D. J. Field (1996). Natural image statistics and efficient coding. *Network : Computation in Neural Systems 7*, 333–339.

Palm, G. (1980). On associative memory. *Biological Cybernetics 36*(1), 19–31.

Perrett, D. I., J. K. Hietanen, M. W. Oram, and P. J. Benson (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences 335*(1273), 23–30.

Phillips, P. J. and Y. Vardi (1995). Data-driven methods in face recognition. In M. Bichsel (Ed.), *International Workshop on Automatic Face and Gesture Recognition*, Zurich, pp. 65–70. University of Zurich Press.

Plumbey, M. (1991). On information theory and unsupervised neural networks. Technical Report CUED/F-INFENG/TR.78, Cambridge University Engineering Department, UK.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992). *Numerical Recipes in C*, pp. 73–78, 102–104. Cambridge University Press.

Redlich, N. A. (1993a). Redundancy reduction as a strategy for unsupervised learning. *Neural Computation 5*, 289–304.

Redlich, N. A. (1993b). Supervised factorial learning. *Neural Computation 5*, 750–766.

Robinette, K. M. and J. J. Whitestone (1992). Methods for characterizing the human head for the design of helmets. Technical Report AL-TR-1992-0061, Crew Systems Directorate, Human Engineering Division, Armstrong Laboratory. Copies are available from the National Technical Information Service, Springfield, Virginia.

Rolls, E. T. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences 335*(1273), 11–20; discussion 20–1.

Ruderman, D. L. (1994a). Designing receptive fields for highest fidelity. *Network : computation in neural systems 5*(2), 147–155.

Ruderman, D. L. (1994b). The statistics of natural images. *Network : computation in neural systems 5*(4), 517.

Ruderman, D. L. (1996). Origins of scaling in natural images. In B. E. Rogowitz and J. A. Allebach (Eds.), *SPIE Proceedings*, Volume 2657. in press.

Ruderman, D. L. and W. Bialek (1994). Statistics of natural images: scaling in the woods. *Physical Review Letters 73*(6), 814–817.

Rumelhart, D. E. and J. L. McClelland (1982). An interactive activation model of context effects in letter perception: Part 2. the contextual enhancement effect and some tests and extensions of the model. *Psychological Review 89*(1), 60–94.

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer network. *Neural Networks 2*, 459–473.

Schmidhuber, J. (1992). Learning factorial codes by predictability maximization. *Neural Computation 4*, 863–879.

Sirovich, L. (1987). Turbulence and the dynamics of coherent structures. *Q. Appl. Math. XLV*, 561–590.

Sirovich, L. and M. Kirby (1987). Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America 4*, 519–524.

Tolhurst, D., Y. Tadmor, and T. Chao (1992). Amplitude spectra of natural images. *Ophthal. Physiol. Opt. 12*, 229–232.

Touretzky, D. S. (1989). Analyzing the energy landscapes of distributed winner-take-all networks. In D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems*, Volume 1, pp. 626–633. Morgan Kaufmann, San Mateo, CA.

Young, A. W. (1992). Face recognition impairments. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences 335*(1273), 47–53; discussion 54.

Zetzcshe, C. (1990). Sparse coding: The link between low level vision and associative memory. In R. Eckmiller, G. Hartmann, and G. Hauske (Eds.), *Parallel Processing in Neural Systems and Computers*. Amsterdam: North-Holland.