SPARSE CODES FOR NATURAL IMAGES

[Davide Scaramuzza](#)
(davide.scaramuzza@epfl.ch)
Autonomous Systems Lab (EPFL)

Final rapport of Wavelet Course MINI-PROJECT (Prof. Martin Vetterli)

ABSTRACT

The human visual system, at the primary cortex, has receptive fields that are spatially localized, oriented and bandpass. It has been shown that a certain learning algorithm to produce sparse codes for natural images leads to basis functions with similar properties. This learning algorithm optimizes a cost function that trades off representation quality for sparseness, and searches for sets of natural images, which basis functions lead to good sparse approximations. The result of the learning algorithm is a dictionary of basis functions with localization in space, direction and scale.
In this paper, dictionaries for different set of images are showed and their own properties are described and verified. It will be showed that the learning algorithm leads to overcomplete bases functions that "capture" the intrinsic structure of the images. This allows efficient coding of the images with good representation quality. The results are applied to image approximation and denoising.

1. INTRODUCTION

In 1962 and '68, the Nobel prize winning discoveries of Hubel and Weisel showed that the mammals primary visual cortex (called V1) consists of cells responsive to simple and complex features in the input. For example, most cells in the visual cortex respond best to edges at some particular angle or the other (thus they are spatially oriented, see Fig.1). More generally, it has been proven that V1 has receptive fields that are characterized as being spatially *localized*, *oriented* and *bandpass*; that is, they are selective to structure of the visual input at different spatial scales [4]. One approach to understanding such response properties of visual neurons has been to consider their relationship to the statistical structure of natural images in terms of efficient coding. In [1],[2] *Olshausen et al.* showed that by designing an algorithm that attempts to find *sparse* linear codes for natural scenes, develops a complete family of localized, oriented and bandpass receptive fields, similar to those found in the primary visual cortex.

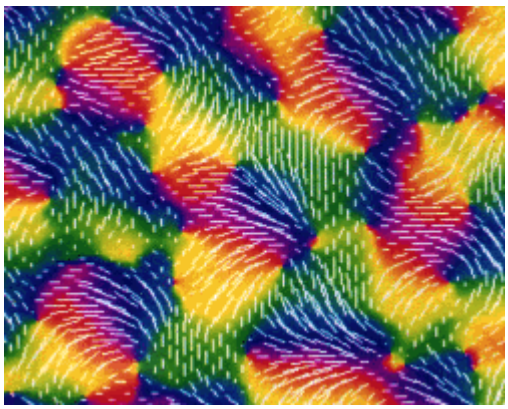


Fig.1 *Orientation columns in a patch of the monkey visual cortex, visualized with modern imaging techniques. Colors varying from red to violet indicate orientation preference of cells varying from zero to 180 degrees. The overlaid white lines also show the orientation preference in each area.*

2. GENERATIVE IMAGE MODEL

In [1] the authors propose that the neurons in V1 model the structure of images $I(x,y)$ in terms of a linear superposition of basis functions ϕ_i plus noise $\varepsilon(x,y)$:

$$I(x,y) = \sum_i s_i \phi(x,y) + \varepsilon(x,y) \qquad (1)$$

One can think to these basis functions as a simple "features vocabulary" for describing images in terms of additive functions. Note also that these bases are not necessarily orthogonal.
The goal of efficient coding is to find a set of ϕ_i that forms a complete code (that is, spans the image space) and results in the coefficient values being as *sparse* and statistically *independent* as possible over an ensemble of

natural images. *Sparseness* of coefficients allows for a small number of bases to represent an image. This resemble what found in V1, where neurons represent data using a small number of active units. The statistical independence, in terms of coding, obviously reduces the redundancy of the code, while, in neurobiological words, translates the fact that neurons are thought as independent of the activity of their neighbours.

One line of approach to this problem, accounting for sparseness and independence, is based on principal component analysis (PCA), in which the goal is to find a set of mutually orthogonal bases functions that capture the directions of maximum variance in the data and for which the coefficients s_i are pairwise decorrelated. However, the receptive fields (bases) that result from this process are not localized as it would be desirable. Moreover, it assumes that the original data are well described by a Gaussian distribution but natural scenes contain many higher order forms of statistical structure. In this case the independent component analysis (ICA) would be suitable because accounts for non-Gaussian distributions. Two limitations common to both of these techniques are that they do not allow for noise to be modelled separately from the signal structure and that they do not allow for overcomplete codes in which there are more basis functions than input dimensions.

Overcompleteness in the representation is important because it allows for the joint space of position, orientation, and spatial-frequency to be tiled smoothly without artefacts. More generally, it allows for a greater degree of flexibility in the representation, as there is no reason to believe a priori that the number of "causes" for images is less than or equal to the number of pixels.

3. CODEBOOK LEARNING ALGORITHM

The strategy for inferring the codebook from a set of natural scenes is explained in [1],[2]. It applies a Bayesian method for deriving an optimal basis that trades off reconstruction quality for sparseness.

Rewriting the (1) in matrix form we get:

$$x = A \cdot s + \varepsilon \tag{2}$$

where **A** is a LxM matrix whose columns are the basis functions, **s** is an M-element vector of basis coefficients and ε is an additive Gaussian noise.

To derive the learning algorithm for finding the basis functions it is possible to proceed in two steps: a) finding the expansion coefficients s_i, given the image and the basis set; b) learning the bases ϕ_i given a set of training images and correspondent coefficients. The above stages can be iteratively repeated until A converges towards a stable solution. Following is the explanation of coefficients extraction and the learning rule to update the bases functions.

a) Finding expansion coefficients

To solve this, we assume we are given an image **x** and a general overcomplete basis **A**. The goal is to infer the vector **s**. This can be done by exploiting Maximum Likelihood Principle, that states the most probable value **s'** for **s** is that satisfying the following:

$$s' = \max_s P(s \mid x, A) \tag{3}$$

By means of Bayes' rule:

$$P(s \mid x, A) \propto P(x \mid A, s) P(s \mid A) \tag{4}$$

Because the model is assumed to be Gaussian:

$$P(x \mid A, s) \propto e^{\frac{(x - As)^2}{2\sigma^2}} \tag{5}$$

where σ is the standard deviation of the additive noise.

The last term in (4), $P(s \mid A)$, is the prior probability distribution over the basis coefficients. If the prior is assumed to be independent of A, then $P(s \mid A) = P(s)$.

At this point observe that, in order to impose sparseness of coefficients, the prior probability P(s) has to peak at zero. The prior P(s) chosen by the authors is the Laplacian distribution (Fig 2):

$$P(s_i) \propto e^{-\vartheta_i |s|} \tag{6}$$

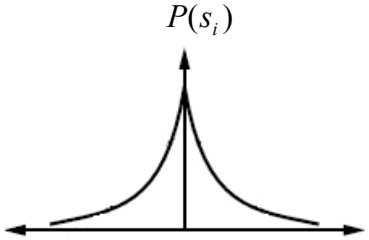


Fig.2 The probability distribution of the coefficients is peaked at zero. Such a distribution would result from a sparse activity distribution over the coefficients.

And by imposing statistical independence among coefficients:

$$P(s) = \prod_i P(s_i) \propto e^{-\vartheta^T |s|} \tag{7}$$

After substituting (4,5,6,7) in (3) we have:

$$s' = \max_s P(s \mid x, A) = \min_s \left(\frac{1}{2\sigma^2} |x - As|^2 + \vartheta^T |s| \right) \quad (8)$$

Thus, in order to find the coefficients s_i for each image presentation, (8) has to be minimized with respect to **s**.

b) Learning basis functions

The goal in learning basis vectors is to obtain a good model of the distribution of natural images. That means, finding a basis matrix which can optimally fit all image presentations. In order to do this, what we want is to maximize the average probability distribution aver as many images as possible. The cost function to be maximized is defined with respect to A is:

$$\Gamma = \langle \log P(x \mid A) \rangle \quad (9)$$

where the distribution $P(x \mid A)$ is obtained by marginalization:

$$P(x \mid A) = \int P(x \mid A, s) P(s) ds \quad (10)$$

The maximization problem can be solved by gradient ascent. This leads to the following updating term for the basis matrix A:

$$\Delta A \propto \langle e \cdot s^T \rangle \quad (11)$$

Where the error $e = |x - As|$. Details about the learning algorithm can be found in [1],[2] and [5].

4. LEARNING CODES FOR NATURAL SCENES

I applied the learning algorithm proposed by B. A. Olshausen to different sets of natural scenes (pictures of nature, pictures of buildings in Amsterdam and Lausanne and paintings of Van Gogh). I also tested it by varying the number of bases (from complete to overcomplete representations) and their dimensions.

Before starting the algorithm, the initial values for basis functions were generated to random values evenly distributed between [-1,1].

Then, the training set was preprocessed, following the suggestions of Olshausen [1], by filtering all the grey-scale images with a zero-phase whitening/lowpass filter having the following 2D Fourier transform:

$$R(f) = f \cdot e^{-\left(\frac{f}{200}\right)^4} \quad (12)$$

where $f = \sqrt{fx^2 + fy^2}$. See also Fig.3. This is because whitening filter is able to counteract the fact that the error computed at (11) preferentially weights low frequencies, while the attenuation at high spatial-frequencies eliminates the artifacts of rectangular sampling.

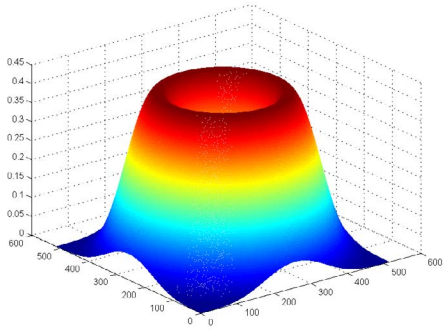


Fig.3 The 2D Fourier transform of the zero-phase whitening lowpass filter used to preprocess the natural pictures.

After these two necessary steps (random initialization of A and preprocessing), I run the algorithm separately on different set of pictures.

Let us suppose to have chosen to learn a complete codebook of 64 bases, each one of dimension 8x8 pixel. Then, in order to compute the updating rule in (11), the algorithm randomly selects a large number (actually 100) of 8x8 image patches from the training set.

A stable solution arrived after about 10,000 – 40,000 updates, that is, 1000,000 – 4000,000 image presentations (between 1 and 4 hours of computation for each set of natural scenes).

5. RESULTS OF LEARNING

The learning algorithm was applied to three sets of ten 512x512 pixel images taken from natural surroundings (Fig.4), Van Gogh paintings (Fig.5) and Amsterdam and Lausanne buildings (Fig.6). The results of the learned codebooks are also shown there.

I derived different codebooks distinguishable for the number of basis functions used (between 64 and 192) and for basis dimensions (between 8x8 and 16x16 pixels).

The results show, as wanted, the basis patches to be well oriented and spatially localized. Moreover, they seem to "capture" the intrinsic structure of the pictures. For instance, in the case of Van Gogh paintings, the extracted bases resemble the brushstrokes of the painter, while, in the pictures of cities, they capture the basic elements of the buildings, mainly composed of vertical and horizontal edges and corners.

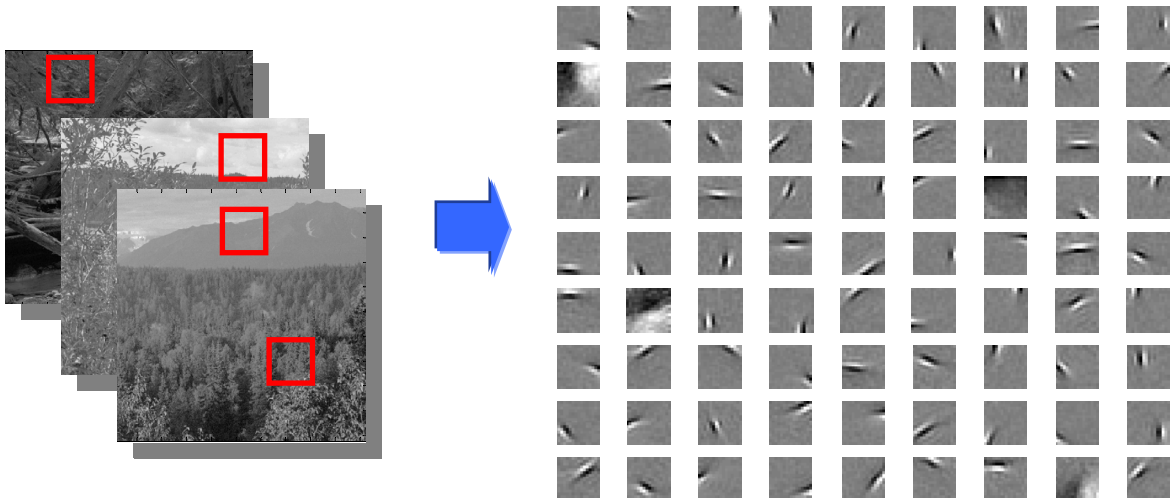
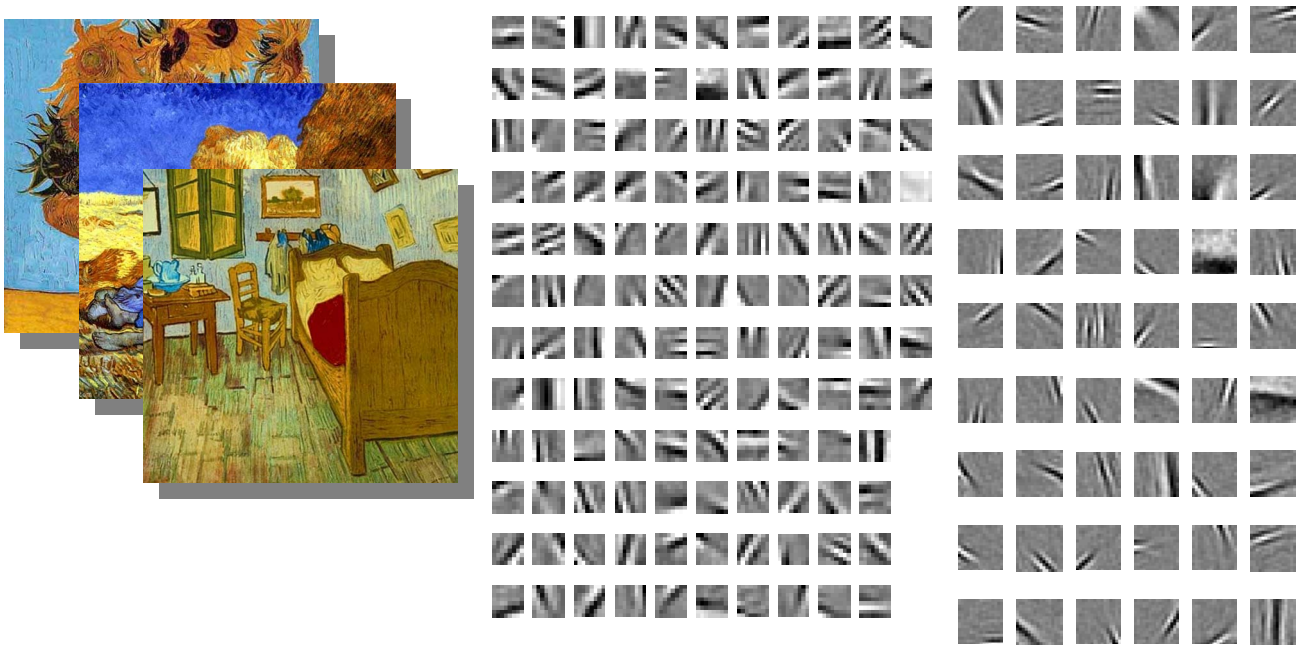


Fig.4 Results from training a system of 192 bases functions on 16x16 image patches extracted from scenes of nature (here only several bases are displayed for reasons of space). The red squares are displayed just for explaining the learning procedure, by randomly selecting image patches from the training set. The scenes were ten 512x512 images of natural surroundings in the American northwest, preprocessed by the mentioned zero-phase whitening/lowpass filter. The results shown were obtained after 40,000 iteration steps (4 hours of computation). Note well that, as desiderated, the learned bases result to be oriented along specific directions and spatially well localized.



a) **b)** **c)**

Fig.5 Results from training a 2x-overcomplete system of 128 bases functions of 8x8 pixels (b) and a system of 192 bases of 16x16 pixels (c) (in the latter only several bases are shown). The codebooks were extracted from ten 512x512 pixel images of Van Gogh paintings (a), preprocessed, as usual, by the zero-phase whitening/lowpass filter. The results shown were obtained after 20,000 (b) and 40,000 (c) iteration steps (2-4 hours of computation respectively). Note also here the learned bases result to be oriented along specific directions and spatially well localized. Moreover the bases seem to capture the intrinsic structure of Van Gogh brushstrokes (this is well visible in (c)).

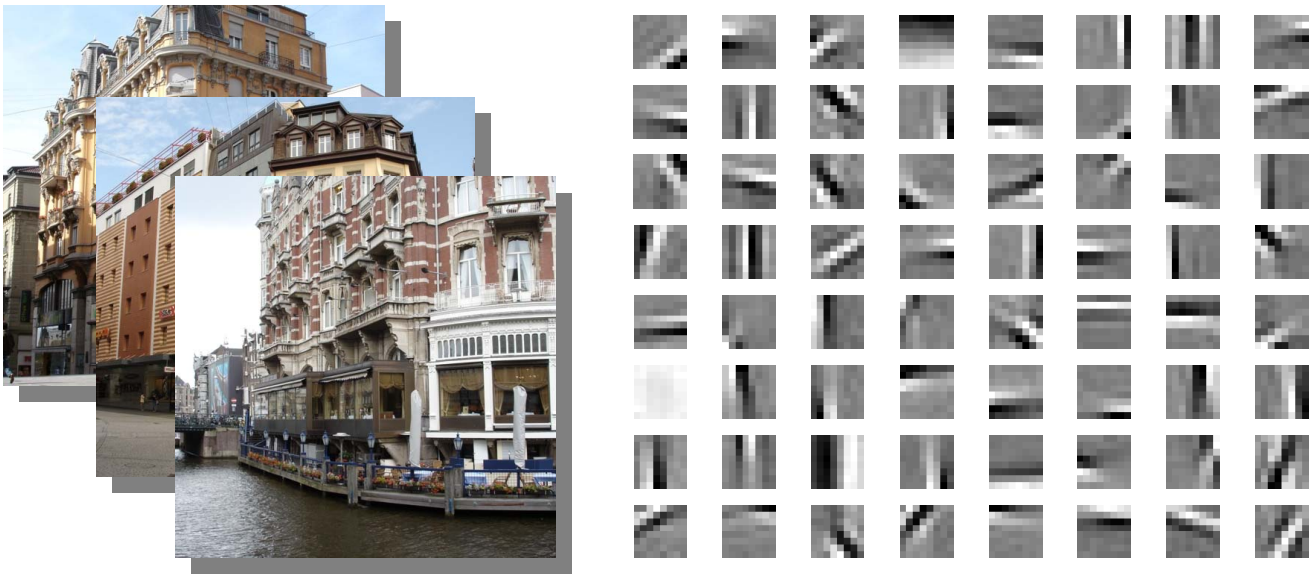


Fig.6 Results from training a complete system of 64 bases functions of 8x8 pixels. The codebook was extracted from ten 512x512 pixel images Lausanne and Amsterdam buildings. The results shown were obtained after 20,000 iteration steps (2 hours of computation). Note also here the learned bases result to be oriented along specific directions and spatially well localized. Moreover the bases seem to capture the intrinsic structure of the building elements, that result to be mainly composed of vertical and horizontal edges and corners.

As mentioned at the beginning, we are also interested in verifying the spatial-frequency localization of basis functions. This can be easily seen by analyzing the bases in the 2D Fourier Transform domain (see Fig.7).

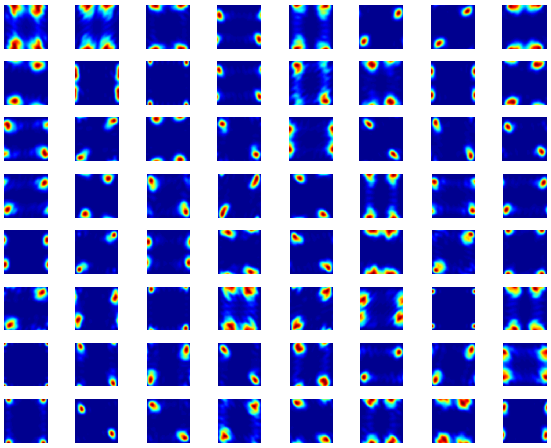


Fig.7 The Power spectrum of basis functions correspondent to Fig.5.

Fig.7 shows that bases functions are well localized in the frequency domain. Another point of interest, not visible in Fig.7, is that the basis functions largest in magnitude also have the lowest peak spatial-frequency tuning. The latter is just what would be found by using PCA technique, which assumes that the data have Gaussian structure. This could reflect an attempt of the model to capture small-amplitude noise in the images. To see better the frequency localization characteristics, the next figures show the polar plots of the peak spatial frequency tuning of all bases.

Figures 8 and 9 show the frequency tiling of basis functions for different degrees of completeness. You can note that increasing the number of bases results in a denser tiling of frequency space (and also in a denser tiling of orientations and positions). Moreover, this frequency tiling is higher at mid-high frequencies (to see this, look at frequency histograms in Fig. 11).

One trend that appears immediately evident is that the preferred orientation tends to align vertically and horizontally, but Olshausen adverts that this is an artefact due to having used a rectangular sampling grid to digitize the images, rather than a reflection of an intrinsic property of the images themselves.

Conversely, for the bases learned from pictures of buildings, basis functions result to be disposed along a few preferential directions of frequency space (Fig.10). This strongly support the mentioned conjecture that they try to capture the intrinsic structure of original images they come from, that are mainly composed of vertical, horizontal and slanting edges.

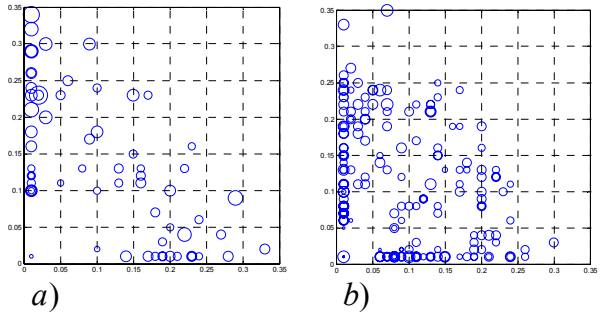


Fig.8 Spatial-frequency localization of bases functions in natural surroundings of Fig.4. The circle diameter is proportional to the bandwidth. a) Complete system of 64 bases of 8x8 pixels, b) System of 192 bases of 16x16 pixels. Note the denser tiling of frequency space by increasing the number of bases (b).

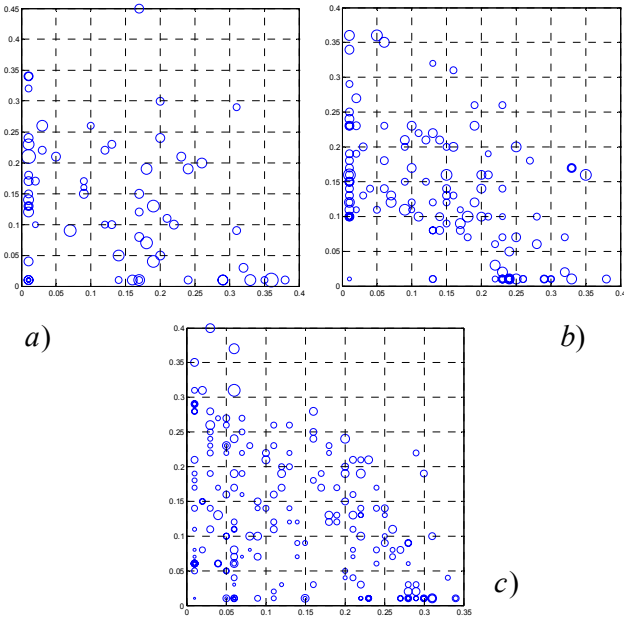


Fig.9 Spatial-frequency localization of bases functions in Van Gogh paintings (Fig.5). a) Complete system of 64 bases of 8x8 pixels, b) 2x overcomplete system of 128 bases of 16x16 pixels, c) 192 bases of 16x16 pixels.

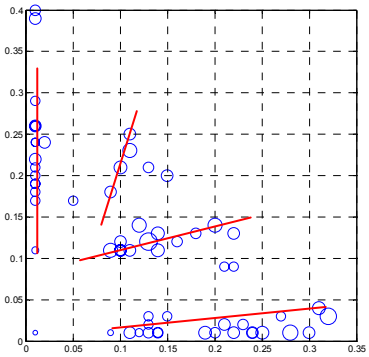


Fig.10 Spatial-frequency localization of bases functions in pictures of buildings (Fig.6). Here it is learnt a complete system of 64 bases of 8x8 pixels. Note that bases are distributed along preferential directions in the frequency domain (red lines). These preferential directions are due to the localized orientation of the correspondent bases in the spatial domain. As seen in Fig.6, they undergo the direction of horizontal, vertical and slanting edges that frequently occur in the pictures.

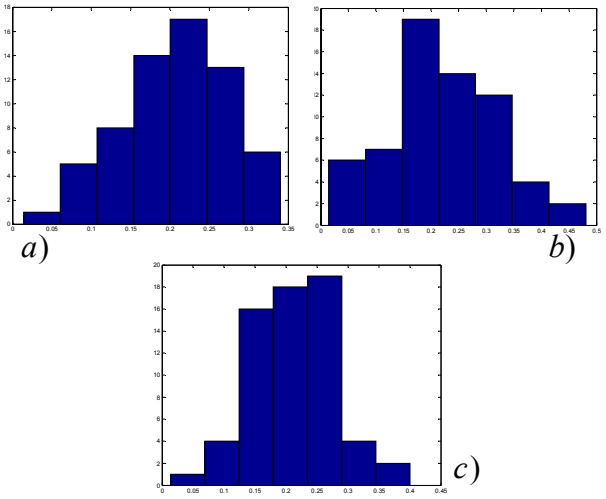
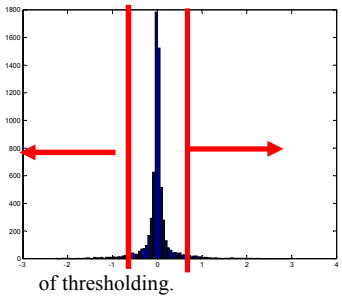


Fig.11 Histograms of peak spatial-frequency bandwidths for the complete 64 learned basis function of natural surroundings (a), Van Gogh paintings (b) and pictures of buildings (c). Note the higher density of bases at mid-high frequencies.

6. RECONSTRUCTION FROM BASES FUNCTIONS

By using the generative model defined in (1), each single image may be simply reconstructed as a linear combination of learned basis vectors. Given the probabilistic nature of the approach, this is not a perfect reconstruction of the original picture but, conversely, is its best approximation by means of the learned bases. Remember that the learning algorithm searches for those bases that better trade off representation quality for sparseness of coefficients, where the latter is essential for efficient coding. To achieve sparseness, we imposed the probability distribution of expansion coefficients to peak at zero (see Fig.12). Thus, to get the best non-linear approximation, we can get the coefficients with higher absolute values, which we interested in, and discard the others (Fig.12). In this way, each image will be represented by the smallest number of bases corresponding to the coefficients with higher activity.



of thresholding.

Then, to have a quantitative measure of the approximation-reconstruction quality, I compared the

Peak-Signal to Noise Ratio (PSNR) among the reconstructed images from the learned codebooks (Fig.13, 14).

a) Original

b) Preprocessed by whitening/lowpass

c) Approximation using 40 bases

d) Approximation using 10 bases

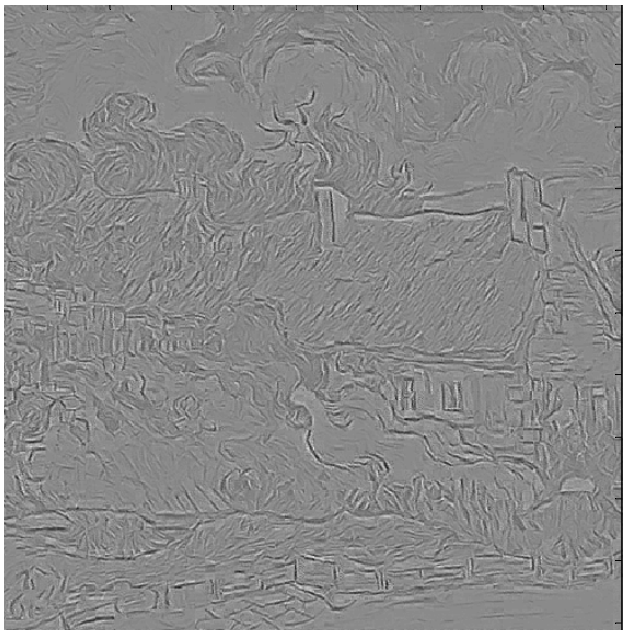
e) Approximation using 5 bases

PSNR	64 bases	40 bases	30 bases	20 bases	10 bases	5 bases	2 bases
Complete	37.06 dB	36.62 dB	35.64 dB	33.49 dB	29.84 dB	27.24 dB	25.00 dB

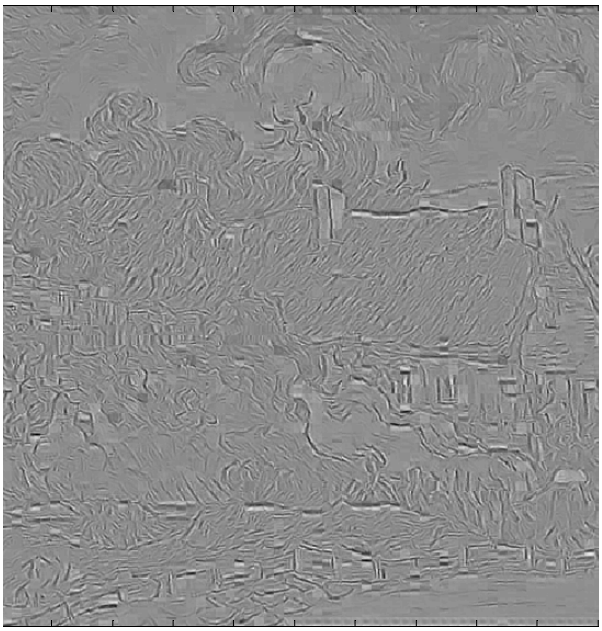
Fig.13 Approximation of the pictures of buildings by using a decreasing number of basis functions of the complete codebook. The bases are selected by taking the correspondent number of higher value coefficients. Note that e) reveals the 8x8 blocks of basis functions.

a) Original

b) Preprocessed image



c) Approximation by 5 bases of the complete codebook



d) Approximation by 5 bases of the overcomplete codebook

PSNR	128 bases	64 bases	40 bases	30 bases	20 bases	10 bases	5 bases	2 bases
Complete	-	31.92 dB	31.43 dB	30.42 dB	28.51 dB	25.60 dB	23.62 dB	22.00 dB
Overcomplete	32.14 dB	31.75 dB	30.61 dB	29.37 dB	27.45 dB	24.77 dB	23.00 dB	21.62 dB

Fig.14 PSNR values computed for different approximations of the original image a). The values are displayed in the case of a complete and overcomplete codebook. Note that d) reveals the 8x8 blocks of basis functions.

7. IMAGE DENOISING

To demonstrate the ability of the adapted bases to capture typical structures in the data, I applied the algorithm to the problem of noise removal in images. This task is well suited to the algorithm because Gaussian additive noise is incorporated into the specification of the image model. A set of bases that characterizes the probability distribution of the data well should improve noise removal properties, because they are better at inferring the most probable image in the face of uncertainty. Indeed, there results a good improvement in PSNR value between the denoised and noisy image (see Fig. 15 and 16).

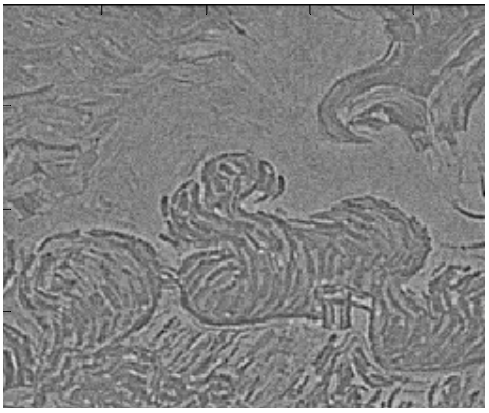


Fig.15 Noisy image. PSNR=28.56 dB

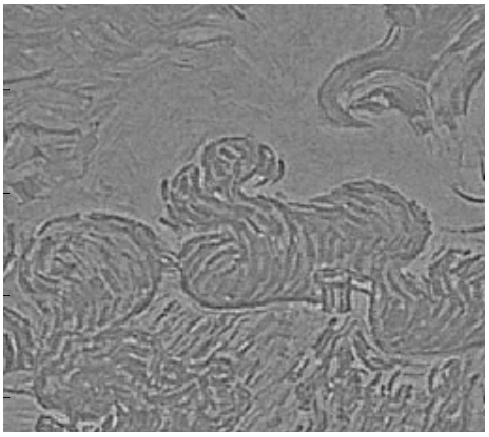


Fig.16 Denoised images. PSNR=30.06 dB

8. CONCLUSIONS

Results demonstrate that localized, oriented and bandpass receptive fields (as the ones found in the simple cells of primary visual cortex) emerge only when two global objectives are placed on a linear coding of natural images:

- that information be preserved
- and that the representation be sparse

That means, an attempt of understanding the response of primary visual neurons has been to consider their relationship to the statistical structure of natural images in terms of efficient coding.
By using Bayes' rule and imposing some constraints of the probability distributions of data, *Olshausen et al.* [1] derived a learning algorithm that provides a set of basis functions minimizing a certain cost function.
Applying the learning algorithm I verified that the learned bases behave as feature detectors able to capture the intrinsic structure of natural images (as seen in Van Gogh paintings and pictures of buildings).
Moreover, increasing the degree of completeness results in a higher density tiling of frequency space.
Because of sparseness and statistical independence among coefficients it was possible to achieve efficient and good quality representations of natural images. Actually, decreasing the number of coefficients up to 10 elements of higher value, still gives a good quality of reconstruction.
Besides, I applied the algorithm to the problem of noise removal in images. Because Gaussian noise is already incorporated into the specification of the image model, the reconstruction stage infers the most probable image in the face of uncertainty, carrying to good results.

Now, the last question, in neurobiological terms, would be: "How close are we to understanding the human visual cortex?"
What we learnt is that:

- Sparseness of coefficients resemble the sparse activity of neuronal receptive fields.
- Learned bases from natural scenes reveal the intrinsic structure of the training pictures: they behave as feature detectors (edges, corners) like V1 neurons.

But:

- The learned bases show higher density in tiling the frequency space only at mid-high frequencies, while the majority of recorded receptive fields appear to reside in the mid to low frequency range.
- Receptive field reveal bandwidths of 1 - 1.5 octaves, while learned bases have bandwidths of 1.7 – 1.8 octaves.
- Finally, neurons are not always statistically independent of their neighbours, as it is instead assumed in our probabilistic model.

Moreover, there still remain several challenges that have to be won by computational algorithms, that are:

- accounting for non-linearity, as shown by neurons at later stages of visual system.
- accounting for forms of statistical dependence.

5. REFERENCES

[1] B. A. Olshausen and D. J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, in Nature, nr. 381, pp. 607-609, 1996.

[2] M. S. Lewicki and B. A. Olshausen Probabilistic framework for the adaptation and comparison of image codes, Journal of the Optical Society of America, 16(7), pp. 1587-1601, 1999.

[3] R. M. Gray, Vector quantization, IEEE ASSP Magazine, pp. 4-29, April 1984.

[4] B. A. Olshausen and D. J. Field, How close are we to understanding V1?, submitted to Neural Computation, 2005.

[5] B. A. Olshausen, and M. S. Lewicki, Sparse Codes and Spikes, Probabilistic Models of the Brain: Perception and Neural Function. R. P. N. Rao, , Eds. MIT Press. pp. 257-272, 2002.