# Wavelets and Natural Image Statistics

Jarmo Hurri, Aapo Hyvärinen and Erkki Oja
Helsinki University of Technology
Laboratory of Computer and Information Science
Rakentajanaukio 2 C, FIN-02150 Espoo, Finland
Email: jarmo.hurri,aapo.hyvarinen,erkki.oja@hut.fi

**Abstract**

It is well-known that wavelets provide a transformation of image data that has excellent properties with respect to image compression. The reasons for the compression ability of wavelets however have not been fully understood. In this paper we show that there is an interesting connection between wavelets and statistical properties of real-world images. This might lead to new theoretical and practical results in the domain of image processing. We show how wavelet-like filters emerge automatically as a result of applying a statistical technique, called independent component analysis, on natural images.

## 1 Introduction

Recently wavelet theory [17] has proved to be one of the most promising approaches to image processing, especially image compression. Wavelets give an orthonormal linear transformation of image data that has the property that the energy of the data is concentrated in only a few coefficients. Different compression schemes have been developed to exploit this property. The basic idea is to preserve only those coefficients that are significantly non-zero; the set of such coefficients is different for every image window. The reasons for the success of such compression schemes has not been fully understood. Some researches have proposed that wavelets are efficient because they mimic the properties of neurons in the visual cortex [13].

In this paper we provide a different insight into the success of wavelets. This is based on statistical analysis of real-world images. We suggest that wavelets are efficient because they closely resemble filters that are obtained when a certain technique, recently developed in statistical signal processing, is applied on real-world images. This technique is independent component analysis [3].

In independent component analysis, or ICA, one tries to decompose a random vector linearly into components that are not only decorrelated, but also as independent as possible in the sense of higher-order statistics. Often this implies that one must find a transformation that provides a vector whose components are as 'sparse' as possible. Sparsity means that the probability of a component to be significantly different from zero is very low.

We suggest that it is this conjunction of independence and sparsity that explains why wavelets are successful. Sparsity implies that the number of simultaneously 'active' components is very small. Thus sparsity embodies exactly the property that seems to be behind the success of wavelets. In our experiments, we applied ICA on real-world images. The obtained filters resemble wavelets, and thus the experiments back up our theoretical arguments.

Similar observations have been done before by others [1, 14]. Our results show a wide variety of structure, while the results of others have been qualitatively more limited. We also try to quantify the connections between our results and wavelets instead of just visually observing that the results are 'wavelet-like.'

## 2 Wavelets

A fundamental problem in signal processing is to find a suitable representation for a signal. Usually different signal representations are based on linear transformations of the signals onto different bases. For example, one way to represent a time signal is to consider its Fourier transformation, which can be seen as projecting it onto a basis of functions each of which consists of just one frequency.

Whereas the usual signal representation is well localized in time — or in the case of images, in space — the Fourier representation of the signal is well localized in frequency. It has been observed that for signals with time varying spectra, i.e., signals in which the frequency contents do not stay constant over time, a representation of the signal localized both in time and in frequency would be very useful for many purposes. One method used to solve the problem is the short time Fourier transformation (STFT), where we restrict our attention around a certain time point of the original signal by multiplying it with a window function (e.g., a Gaussian function) and Fourier transforming this modified signal. The problem with STFT is the constant window size, which is difficult to choose and fixed for all frequencies [17].

Wavelets are one method developed to solve this problem [4]. Wavelets are families of basis functions, each family being generated by scaling and translating a 'model function' called *mother wavelet.* From a mathematical point of view this restricted form of functions is an essential property because it makes the theoretical analysis feasible. But it may not be crucial from the point of view of applications. An important property of wavelets is that they are localized both in time and frequency [17]. This is also typical of STFT, but the localization of wavelets is specialized so that wavelets that respond to low frequencies are more frequency selective, i.e., localized in the frequency domain, but also more spread in time than wavelets which respond to high frequencies. So wavelets have variable window size, and the window size is connected to the frequency response of the wavelet. These qualities will be the basis of our examination of the similarities between ICA and wavelets.

Wavelets represent a compromise between good time resolution and good frequency resolution, a trade-off being forced by the linear approach [17] (for other, nonlinear methods see [2]). Some applications in which they have proven to be useful are feature detection, compression, noise removal, computer vision and graphics and time-frequency description of signals [15].

# 3 Independent Component Analysis and Sparse Coding

## 3.1 The Basic Model

Independent Component Analysis (ICA) [3, 11] is a statistical signal processing technique whose goal is to express a set of random variables as linear combinations of statistically independent component variables. Some applications of ICA are blind source separation [11], feature extraction [1, 7], and, in a slightly modified form, blind deconvolution [6]. In the simplest form of ICA [3], we observe $m$ scalar random variables $x_1, x_2, ..., x_m$ which are assumed to be linear combinations of $n$ *unknown independent components*, or ICs, $s_1, s_2, ..., s_n$ that are mutually statistically independent, and zero-mean. Note that the assumption of zero mean is in fact no restriction, as this can always be accomplished by a preliminary centering of the observed data. To enable estimation of the ICs, we must also assume that $n \leq m$. Let us arrange the observed variables $x_i$ into a vector $\mathbf{x} = (x_1, x_2, ..., x_m)^T$ and the IC variables $s_i$ into a vector $\mathbf{s}$, respectively; then the linear relationship is given by

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

Here, $\mathbf{A}$ is an unknown $m \times n$ matrix of full rank, called the mixing matrix. The basic problem of ICA is then to estimate the realizations of the original ICs $s_i$ using only the mixtures $x_j$ or, equivalently, to estimate the mixing matrix $\mathbf{A}$. The fundamental restriction of the model is that we can only estimate non-Gaussian ICs (except if just one of the ICs is Gaussian). Moreover, neither the energies nor the signs of the ICs can be estimated because any constant multiplying an IC in eq. (1) could be canceled by dividing the corresponding column of the mixing matrix $\mathbf{A}$ by the same constant. For mathematical convenience, one usually defines that the ICs $s_i$ have unit variance. This makes the (non-Gaussian) ICs unique, up to their signs [3]. Note that no order is defined between the ICs.

If ICA is used for *feature extraction* [1, 7], the columns of $\mathbf{A}$ represent features, and $s_i$ signals the presence and the 'amplitude' of the $i$-th feature in the observed data $\mathbf{x}$.

## 3.2 Contrast functions

The basic principle of many algorithms for ICA estimation is the use of a contrast function. A very popular contrast function is kurtosis. Kurtosis, or the fourth-order cumulant [10] is defined for a zero-mean random variable $v$ as $\mathrm{kurt}(v) = E\{v^4\} - 3(E\{v^2\})^2$. Kurtosis is a contrast function for ICA in the following sense. Consider a linear combination of the observed mixtures, say $\mathbf{w}^T\mathbf{x}$, where the vector $\mathbf{w}$ is constrained so that $E\{(\mathbf{w}^T\mathbf{x})^2\} = 1$. When $\mathbf{w}^T\mathbf{x} = \pm s_i$ for some $i$, i.e., when the linear combination equals, up to the sign, one of the ICs, the kurtosis of $\mathbf{w}^T\mathbf{x}$ is minimized or maximized [5, 10]. This property is widely used in ICA algorithms.

For our purposes, it is interesting that contrast functions can in most cases be interpreted as measures of sparsity. It is well-known [12] that distributions with high positive kurtosis are usually quite sparse, i.e., their densities are peaked at zero, thus making the probability of significantly non-zero values small.

In fact, under suitable assumptions, ICA estimation simply means finding those linear combinations $\mathbf{w}^T\mathbf{x}$ in which the kurtosis, and thus sparseness, are maximal. The main assumption needed is that the ICs must have positive kurtoses, i.e., they are sparser than the Gaussian distribution. This assumption seems to be true for most components in image data. Thus we see that the ICA transformation provides us with components that are at the same time as independent and as sparse as possible.

### 3.3 A Fixed-Point Algorithm for ICA

To actually perform the ICA estimation we use a fast fixed point algorithm. A necessary prerequisite for the algorithm is that the data be uncorrelated or *white.* For this purpose the data is whitened by

$$\mathbf{v} = \mathbf{D}^{-1/2}\mathbf{E}^T\mathbf{x},$$

where $\mathbf{E}$ is the matrix of eigenvectors of the covariance matrix of $\mathbf{x}$ and $\mathbf{D}$ the diagonal matrix of corresponding eigenvalues. This is called *PCA whitening.* At this point it is also possible to use the properties of PCA to reduce the dimension of the data by selecting only the largest eigenvalues and their corresponding eigenvectors to form $\mathbf{D}$ and $\mathbf{E}$.

The computation of ICA is accomplished using the following fixed-point algorithm, proposed in [8, 9]. In this algorithm, we have a set of vectors $\mathbf{w}_i, i = 1, ..., n$, each of which is updated in the $(k+1)$-th step as follows:

$$\mathbf{w}^*(k+1) = E\{\mathbf{v}g(\mathbf{w}(k)^T\mathbf{v}) - g'(\mathbf{w}(k)^T\mathbf{v})\mathbf{w}(k)\} \tag{2}$$
$$\hat{\mathbf{w}}(k+1) = \frac{\mathbf{w}^*(k+1)}{||\mathbf{w}^*(k+1)||}$$

where $g$ is a suitable non-linearity, e.g. one of the following:

$$g_1(u) = \tanh(u),\ g_1'(u) = \frac{1}{\cosh^2(u)} \tag{3}$$
$$g_2(u) = u\exp(-u^2/2),\ g_2'(u) = (1-u^2)\exp(-u^2/2) \tag{4}$$

(In practice, the expectation in (2) is estimated using a sufficiently large sample of data.) After updating the vectors as in (2), they must also be orthogonalized. This is done using a method that changes the matrix minimally in the Frobenius norm sense

$$\mathbf{W} = \hat{\mathbf{W}}(\hat{\mathbf{W}}^T\hat{\mathbf{W}})^{-1/2},$$

where $\mathbf{W} = [\mathbf{w}_1 \ \cdots \ \mathbf{w}_n]$ and $\hat{\mathbf{W}} = [\hat{\mathbf{w}}_1 \ \cdots \ \hat{\mathbf{w}}_n]$. For details on this algorithm, see [8], where it is also shown that the algorithm searches for the right extrema of a contrast function which can be seen as a robustified version of kurtosis, thus finding the ICs as linear combinations $\mathbf{W}^T\mathbf{v}$.

After matrix $\mathbf{W}$ has been determined, matrix $\mathbf{A}$ is calculated by $\mathbf{A} = \mathbf{E}\mathbf{D}^{1/2}\mathbf{W}$. This is valid even if we have reduced the dimension of the problem — then the determination of matrix $\mathbf{A}$ of ICA basis vectors is underdetermined, and the solution above is the minimum norm or pseudoinverse solution to the problem.

## 4 Experiments

The data used in the experiments consisted of 15 different natural images describing different scenes, plants and animals. In the experiments a set of 10000 (possibly overlapping) subimages of size $12 \times 12$ pixels was extracted randomly from the image set. These subimages were vectorized into 144-dimensional vectors, which were used as the mixed data $\mathbf{x}$ of the ICA model (1).

Two preprocessing steps were used. First, low frequency components of the overall images were discarded by subtracting from each sample vector the mean of its components. Second, in order to avoid the domination of high variance areas we equalized the local variance in each sample to 1 by dividing each sample by its norm. The subtraction of the mean reduces one dimension of the data, so PCA was used to reduce the dimension of the input data by 1 when the data was whitened.

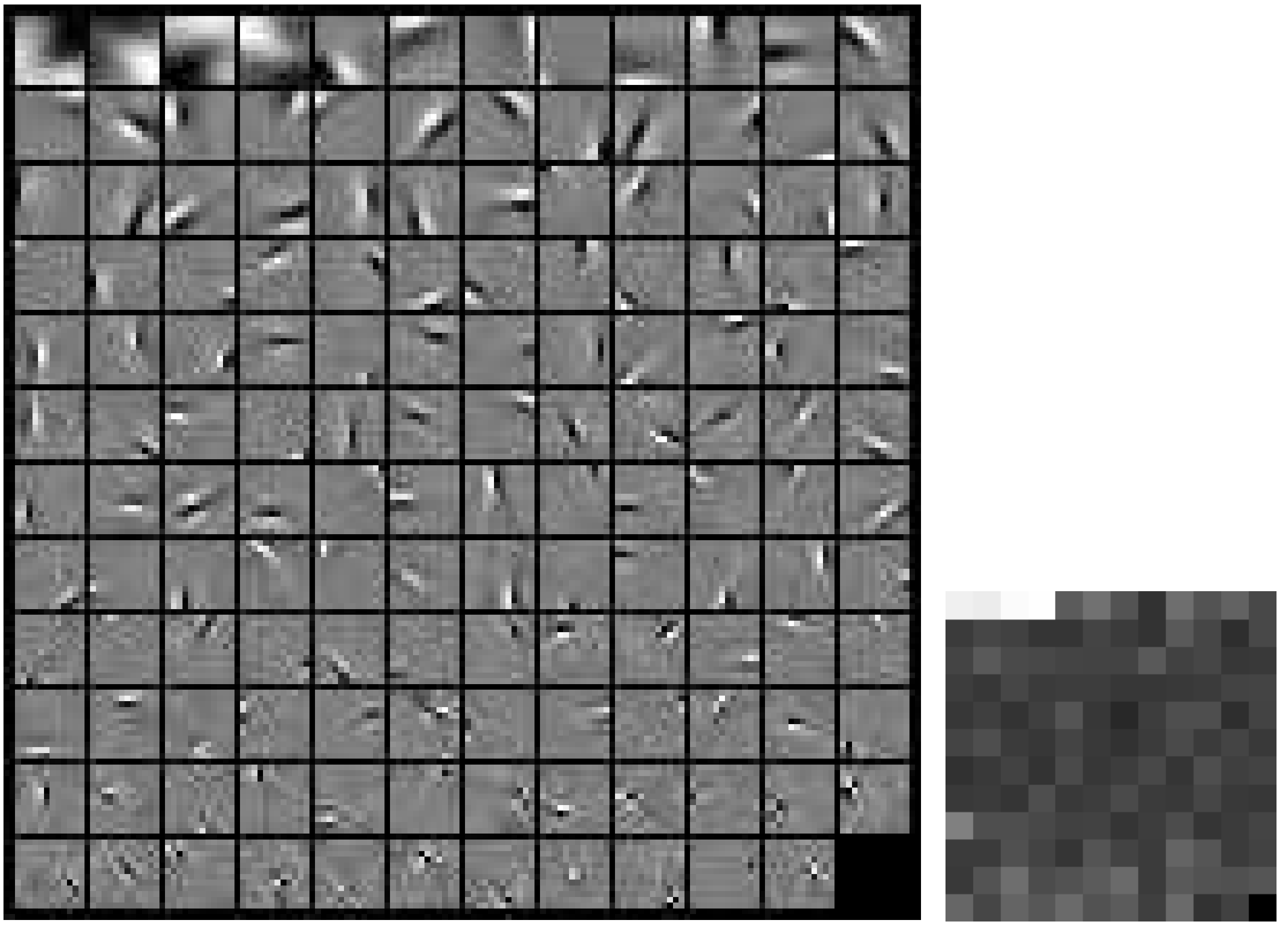

Figure 1: Basis vectors obtained using the fixed point algorithm and their spreads. The image has been scaled for visual display so that the mean grayscale value corresponds to value zero in the original vectors. A lighter color indicates a larger value.

# 5 Results

Experiments using ICA were conducted with the fixed point algorithm in (2), using $g_1$ in (3) as the nonlinearity $g$. The results can be seen in Figure 1. The subimages presented here are *ICA basis vectors,* that is, column vectors of matrix $\mathbf{A}$ in the ICA model (1). Assuming that the hypothesized ICA model (1) holds here, we would deduce that each image block in the data set is built of construction blocks of Figure 1, the coefficient of each block being given by the value of the corresponding independent component.

In order to analyze the results we introduce some useful concepts from time-frequency analysis [2] and generalize these to two dimensions. The techniques we shall use are called *time and frequency distributions.* Consider the energy of a time signal at each moment, that is, the square of the signal. Assume that we normalize the total energy of the signal to be 1. Then we can consider this energy function to be an *energy density* [2]. This density can be used in a similar manner as a probability density to calculate properties of the signal. For example for a 1-D signal $s(t)$ which has been normalized to have an energy of 1, the *mean time* of the signal is calculated as $\langle t \rangle = \int t\,|s(t)|^2\,dt$. The *duration* of the signal — or variance in time — is calculated as $\left\langle (t - \langle t \rangle)^2 \right\rangle$. Similar concepts can be defined in the frequency domain, giving us the *mean frequency* and *bandwidth* (or variance in frequency).

Now consider generalizing these definitions for image data. Then the energy densities in both spatial and frequency domain are functions of two variables — in the spatial domain we have the image function $s(x, y)$ and in the frequency domain the corresponding cosine transform $S(u, v)$ — so the densities are similar to joint probability densities of two random variables. The *mean location* and mean frequency of an image are calculated as for usual random variables. The result is a two-dimensional vector. To calculate the spatial variance, which we shall call

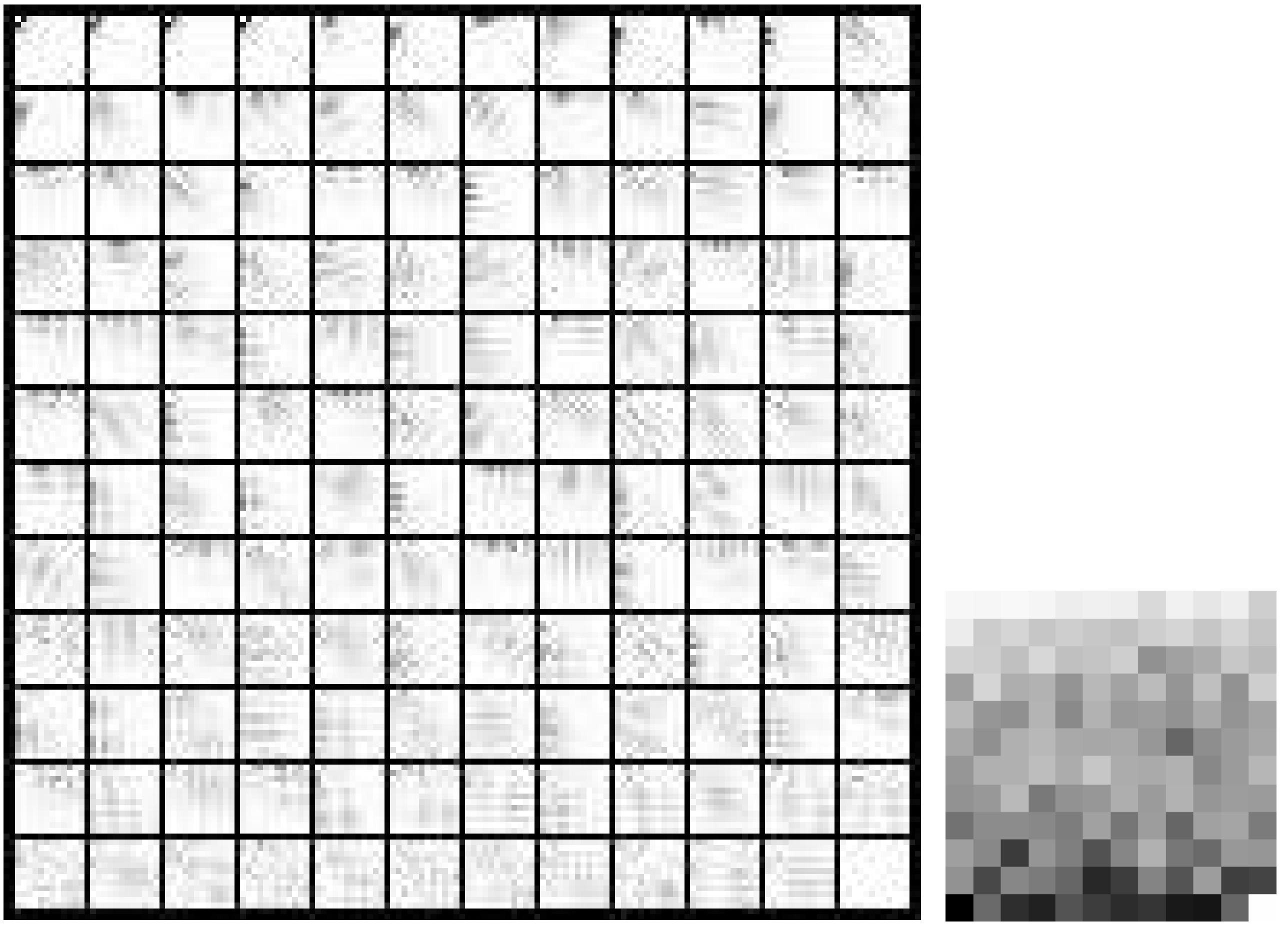

Figure 2: Magnitudes of the 2-dimensional discrete cosine transformations of basis vectors of Figure 1 and their bandwidths. The energies of the vectors have been equalized. Here a darker color indicates a larger value.

*spread*, we add the variances of $x$- and $y$-directions together. The bandwidth of the image is defined similarly.

ICA basis vectors in Figure 1 have been ordered by the magnitudes of their mean frequencies. In the smaller subimage we can also see their spreads. In Figure 2 we can see the two-dimensional discrete cosine transformations of the basis vectors of Figure 1 along with the corresponding bandwidths.

As can be seen in Figures 1 and 2, most basis vectors obtained using the fixed point algorithm are localized in both space and frequency, that is, the energies of the vectors are localized to a subset of the space. We can also see that those basis functions representing lower frequencies are spatially more spread, but also more localized in the frequency domain than those representing high frequencies. These are central properties of wavelets [17, 16], which makes this an important result for the analysis of similarities and connections between ICA and wavelets.

When different initial starting points were selected in the algorithm (2), the results were quantitatively different but qualitatively alike, that is, the basis vectors fulfilled the wavelet-like properties described above. Mostly the obtained vectors were translated and/or rotated versions of the basis vectors of other runs. This suggests that the ICA model does not hold in the sense that there may be more sources than the number of measured signals, and that the algorithm finds only an orthogonal subset of possible 'independent' directions.

# 6 Conclusions

In this paper we examined the connection between wavelets and the statistical properties of natural images. We hypothesized that natural images follow the ICA model in small scale (small subwindows), and we used an ICA algorithm to extract the parameters of this linear model. The algorithm used is a fast fixed-point algorithm, based on the optimization of a contrast function [8, 9]. We found out that ICA basis vectors obtained using this method resemble wavelets, i.e., they are localized in both space and frequency and their spatial localization is directly proportional and frequency localization inversely proportional to the frequencies to which the wavelet responds. This provides a new insight into the theory of wavelets from the viewpoint of statistical properties of natural images and signals, and may give us a way to search data or application dependent bases for their representation.

# References


[1] Anthony Bell and Terrence Sejnowski. Edges are the independent components of natural scenes. In *Advances in Neural Information Processing Systems 9*. The MIT Press, Cambridge, Massachusetts, 1997.

[2] Leon Cohen. *Time-Frequency Analysis*. Prentice Hall Signal Processing Series. Prentice Hall, Englewood Cliffs, New Jersey, 1995.

[3] Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[4] Ingrid Daubechies. Where do wavelets come from? — a personal point of view. *Proceedings of the IEEE*, 84(4):510–513, April 1996. Special Issue on Wavelets.

[5] Nathalie Delfosse and Philippe Loubaton. Adaptive blind source separation of independent sources: A deflation approach. *Signal Processing*, 45:59–83, 1995.

[6] Simon Haykin. *Adaptive Filter Theory*. Prentice Hall International, 3rd edition, 1996.

[7] Jarmo Hurri, Aapo Hyvärinen, and Erkki Oja. Image feature extraction using independent component analysis. In *Proceedings of the IEEE Nordic Signal Processing Symposium (NORSIG) '96*, Espoo, Finland, 1996.

[8] Aapo Hyvärinen. A family of fixed-point algorithms for independent component analysis. Technical Report A 40, Helsinki University of Technology, Faculty of Information Technology, Laboratory of Computer and Information Science, 1996.

[9] Aapo Hyvärinen. A family of fixed-point algorithms for independent component analysis. In *Proceedings of the International Conference on Acoustics, Speech & Signal Processing (ICASSP) '97*, Munich, Germany, 1997.

[10] Aapo Hyvärinen and Erkki Oja. One-unit learning rules for independent component analysis. In *Advances in Neural Information Processing Systems 9*. The MIT Press, Cambridge, Massachusetts, 1997.

[11] Christian Jutten and Jeanny Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[12] Maurice Kendall and Alan Stuart. *The Advanced Theory of Statistics*, volume 1 – Distribution Theory. Charles Griffin & Company Ltd, London, 1958.

[13] Stéphane Mallat. Wavelets for a vision. *Proceedings of the IEEE*, 84(4):604–614, April 1996. Special Issue on Wavelets.

[14] Bruno Olshausen and David Field. Wavelet-like receptive fields emerge from a network that learns sparse codes for natural images. *Nature*, Accepted.

[15] *Proceedings of the IEEE*, volume 84, April 1996. Special Issue on Wavelets.

[16] Gilbert Strang and Truong Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, Wellesley, MA, 1996.

[17] Martin Vetterli and Jelena Kovačević. *Wavelets and Subband Coding*. Prentice Hall Signal Processing Series. Prentice Hall, Englewood Cliffs, New Jersey, 1995.