Astronomical Image and Signal Processing

Looking at Noise, Information, and Scale

Jean-Luc Starck and Fionn Murtagh

The information content of an image is a key aspect of any processing task, and there are two ways to define information: we can recognize the different objects in the image, which is fundamental for most image analysis applications, or we can derive a quantity, which is representative of the amount of information in the image.

Shannon [38], in the framework of communication theory, suggested that the information related to a given measure was connected to its frequency of occurrence and proposed the entropy function as a means to calculate the amount of information in a data set. An important property of the entropy is that it gives exactly the number of bits needed to compress or transfer the data without any loss of information.

Shannon entropy does not take into account the correlation between pixels, and it is very well known that decorrelating the data decreases the entropy. Using a reversible transform such as the discrete cosine transform (DCT) leads to a new data set, which has lower entropy and can therefore be compressed with fewer bits. This result has been employed in the widely used JPEG compression method. But this result also means that we have introduced *a priori* information, namely that the data are correlated. An hypothesis of how the data are correlated is given by the choice of the transform. The more accurate this assumption is, then the more faithfully the chosen transform will represent the data and decrease the entropy.

The wavelet transform (WT) is considered one of the best tools to do this job. It allows us to separate the components of an image according to their size. It is a perfect tool for astronomical images, typically composed of objects of different size, such as stars, galaxies, interstellar clouds, clusters of galaxies, etc.

It is clear that if a multiscale representation is capable of representing information which serves to decrease the entropy, it is also useful for detection of the different components contained in the image. For instance, it has been shown that an image can be completely (or almost completely) reconstructed from its multiscale edges [31], i.e., the edges detected at each scale. So the WT is useful for both kinds of information measurement, the deterministic kind, which indicates what is in the image and where, and the statistical one, which gives a quantity relative to the amount of information.

Another aspect of astronomical images, but also of many other classes of image, is the presence of noise. The observed signal can be decomposed into two components: the useful signal (or signal of interest) and the noise. It is very important to understand what the behavior of the noise is like in wavelet space, if we want subsequently to analyze correctly the data in such a space.

We present in this article how to measure the information in an astronomical image, in both a statistical and a deterministic way. In the following sections, we discuss the wavelet transform and the noise modeling, and we describe how to measure the information and the implications for objet detection, filtering, and deconvolution.

The Wavelet and Other Multiresolution Transforms

There are many two-dimensional WT algorithms [44]. The most well known are:

▲ *The (bi-) orthogonal wavelet transform.* This wavelet transform [32], often referred to as the discrete wavelet transform (DWT), is the most widely used among available DWT algorithms. It is a nonredundant representation of the information. An introduction to this type of transform can be found in [9]. The Haar wavelet transform belongs to this class.

▲ *The Feauveau wavelet transform.* Feauveau [15] introduced **AU: spelling ok -- quincunx** analysis. This analysis is not dyadic and allows an image decomposition with a resolution factor equal to $\sqrt{2}$.

▲ *The à trous algorithm [39], [44]*. The wavelet transform of an image by this algorithm produces, at each scale j, a set $\{w_j\}$. This has the same number of pixels as the image. The original image c_0 can be expressed as the sum of all the wavelet scales and the smoothed array c_p by $c_0 = c_p + \Sigma_{j=1}^{p} w_j$ and a pixel at position x, y can be expressed also as the sum of all the wavelet coefficients at this position, plus the smoothed array $c_0(x,y) = c_p(x,y) + \sum_{j=1}^{p} w_j(x,y)$.

The multiresolution median transform [46] is one alternative to the wavelet transform for multiscale decomposition. The median transform is nonlinear (so it is not a wavelet transform, but represents an image in a similar way to the à trous algorithm) and offers advantages for robust smoothing (i.e., the effects of outlier pixel values are mitigated). The multiresolution median transform consists of a series of smoothings of the input image, with successively broader kernels. Each successive smoothing provides a new resolution scale. The multiresolution coefficient values constructed from differencing images at successive resolution scales are not necessarily of zero mean, and so the potential artifact-creation difficulties related to this aspect of wavelet transforms do not arise. For integer input image values, this transform can be carried out in integer arithmetic only which may lead to computational savings. As in the case of the à trous algorithm, the original image can be expressed by a sum of the scales and the smoothed array.

All of these methods have advantages and drawbacks. Based on the content of the image, and the nature of the noise, each of these transforms may well be considered as optimal. In astronomical images, there are generally no edges, and objects are relatively diffuse. For this reason, an isotropic or symmetric analysis produces better results. Furthermore, for most usual applications (detection, filtering, deconvolution, etc.), undersampling leads to severe artifacts which can be easily avoided by nonorthogonal transforms such as the à trous algorithm. For these reasons, the DWT is widely used for compression, but is less attractive for pattern recognition applications in the astronomical domain. The main WT used is the à trous algorithm, because it is symmetric, which is appropriate for a large range of astronomical images.

Noise and Significant Wavelet Coefficients

Here we describe how to model the noise in wavelet space. In general, data are contaminated by Gaussian noise, Poisson noise, or a combination of both. But sometimes, data are already processed (through calibration processing for instance), and the noise does not follow a known distribution. However, we can often make some assumptions, such as to suppose that it is locally Gaussian. Another relatively usual case is that of nonhomogeneous or nonstationary Gaussian noise. In such a case we can derive, for each pixel, the noise standard deviation. This means that we have a second image (the root mean square image) indicating the noise level for each pixel. It can be updated every time we coadd or otherwise process our data.

This section discusses how we handle different situations. Some of our noise modeling results are, of necessity, derived for specific wavelet transforms.

Gaussian Noise

Noise is often taken as Poisson (related to the arrival of photons) and/or Gaussian. For Gaussian noise, it is easy to derive an estimation of the noise standard deviation σ_j at scale j from the noise standard deviation, which can be evaluated with good accuracy in an automated way [42]. To detect the significant wavelet coefficients, it suffices to compare the wavelet coefficients $w_j(x,y)$ to a threshold level t_j. t_j is generally taken equal to $k\sigma_j$, and k is chosen between three and five. The value of three corresponds to a probability of false detection of 0.27%. If $w_j(x,y)$ is small, then it is not significant and could be due to noise. If $w_j(x,y)$ is large, it is significant:

$$\begin{aligned} &\text{if } |w_j| \geq t_j \quad \text{then } w_j \text{ is significant} \\ &\text{if } |w_j| < t_j \quad \text{then } w_j \text{ is not significant.} \end{aligned} \tag{1}$$

Other threshold methods have been proposed, like the *universal threshold* [12], [10], or the SURE method [11], but they generally do not produce as good results as the hard thresholding method based on the significant coefficients. An alternative to the hard thresholding is the soft thresholding, which consists of replacing each wavelet coefficient $w_{j,k}$ (j being the scale index, and k the position index) by the value $\tilde{w}_{j,k}$ where

$$\tilde{w}_{j,k} = \text{sgn}(w_{j,k})\left(|w_{j,k}| - T_j\right) \text{ if } |w_{j,k}| \geq T_j \tag{2}$$

$$= 0 \text{ otherwise.} \tag{3}$$

But for astronomical images, soft thresholding should never be used because it leads to photometry loss in regard to all objects, which can easily be verified by examining the residual map (i.e., data – filtered data).

Poisson and Combined Gaussian and Poisson noise

Poisson Noise: Case of More Than 20 Photons Per Pixel

If the noise in the data I is Poisson, the transformation [1] $t(I) = 2\sqrt{I + 3/8}$ acts as if the data arose from a Gaussian white noise model, with $\sigma = 1$, under the assumption that the mean value of I is sufficiently large. However, this transform has some limits, and it has been shown that it cannot be applied to data with less than about 20 photons per pixel. So for X-ray or gamma ray data, other solutions have to be found, which manage the case of a reduced number of events or photons under assumptions of Poisson statistics.

Combined Gaussian and Poisson Noise

The generalization of variance stabilization [35] is

$$t(I(x,y)) = \frac{2}{\alpha}\sqrt{\alpha I(x,y) + \frac{3}{8}\alpha^2 + \sigma^2 - \alpha g}$$

where α is the gain of the detector, and g and σ are the mean and the standard deviation of the read-out noise.

Poisson Noise with Few Events Using the À Trous Transform

A wavelet coefficient at a given position l and at a given scale j is

$$w_j(x,y) = \sum_{k \in K} n_k \psi\left(\frac{x_k - x, y_k - y}{2^j}\right) \tag{4}$$

where K is the support of the wavelet function ψ and n_k is the number of events which contribute to the calculation of $w_j(x,y)$ [i.e., the number of photons included in the support of the dilated wavelet centered at (x,y)].

If a wavelet coefficient $w_j(x,y)$ is due to noise, it can be considered as a realization of the sum $\Sigma_{k\in K} n_k$ of independent random variables with the same distribution as that of the wavelet function [n_k being the number of events used for the calculation of $w_j(x,y)$]. This allows the comparison of the wavelet coefficients of the data with the values which can be taken by the sum of n independent variables.

The distribution of one event in wavelet space is then directly given by the histogram H_1 of the wavelet ψ. Since we consider independent events, the distribution of a coefficient w_n (note the changed subscripting for w, for convenience) related to n events is given by n autoconvolutions of H_1:

$$H_n = H_1 \otimes H_1 \otimes \mathrm{L} \otimes H_1. \tag{5}$$

For a large number of events, H_n converges to a Gaussian. We then establish detection thresholds with 95% or 99% confidence. This constructive detection approach clearly depends on the wavelet function used.

Poisson Noise with Few Events Using the Haar Transform

The Haar transform presents the advantage of simplicity for modeling Poisson noise. A wavelet coefficient is only the difference of two values which both follow a Poisson distribution. For one-dimensional data, and using the normalized Haar transform (L2-normalization), Kolaczyk [27] proposed to use a detection level for the scale j equal to: **AU: Is there an extra parenthesis in (6)?**

$$t_j = 2^{-(j+2/2)}(2\log\left(n_l + \sqrt{(4\log n_l)^2 + 8\log n_l \lambda_l}\right) \tag{6}$$

where $n_l = 2^{\log_2 n - 1}$, n being the number of samples, and λ_l the background rate per bin in n_l.

In two dimensions, Kolaczyk and Dixon [26] proposed to use the following threshold, corresponding to a false detection rate of α:

$$t_j = 2^{-(j+1)}\left[z_{\alpha/2}^2 + \sqrt{z_{\alpha/2}^4 + 4\lambda_j z_{\alpha/2}^2}\right] \tag{7}$$

where j is the scale level ($j = 1 \ldots J$, J being the number of scales), $\lambda_j = 2^{2j}\lambda$ is the background rate over $n_j = 2^{2j}$ pixels, and $z_{\alpha/2}$ is the point under the Gaussian density function for which there falls a mass of $\alpha/2$ in the tails beyond each $z_{\alpha/2}$ point.

Jammal and Bijaoui [24] found that the probability density function (PDF) of a Haar wavelet coefficient (using the filters $h = [1,1]$ and $g = [1,-1]$) is given by

$$p\left(w_{j+1} = \nu\right) = \exp - 2^{2(j+1)} \lambda_p I_\nu\left(2^{2j+1}\lambda_p\right) \tag{8}$$

where $I_\nu(x)$ is the modified Bessel function of integer order ν, and λ_p is the background rate.

However, some experiments [40] have shown that

▲ The à trous algorithm is significantly better than any of the Haar-based methods, and

▲ The Jammal-Bijaoui threshold is a little better than the Kolaczyk one for compact source detection and is equivalent for more extended sources. But the Kolaczyk threshold requires less computation time and is easier to implement.

Hence, the Haar transform is less efficient for restoring astronomical images than the à trous algorithm. But its simplicity, and the computation time, may be attractive in practice.

Model and Simulation

Simulations can be used for deriving the probability that a wavelet coefficient is not due to the noise [14]. Modeling a sky image (i.e., uniform distribution and Poisson noise) allows determination of the wavelet coefficient distribution and derivation of a detection threshold. For substructure detection in an astronomical cluster, the large structure of the cluster must be first modeled, or otherwise noise photons related to the large scale structure will introduce false detections at lower scales. If we have a physical model, Monte Carlo simulations can also be used [13], [21], but this requires lengthy computation time, and the detections will always be model dependent. Damiani et al. [8], and also Freeman et al. [16], propose to calculate the background from the data to derive the fluctuations due to the noise in the wavelet scales. From our point of view, it is certainly better to use the histogram autoconvolution method, which is robust and fast, but the simulations may still be useful in the case where the noise does not follow exactly a Poisson distribution, due, for example, to instrumental effects. If these effects can be simulated, the detection level will also take them into account.

Root Mean Square Map

The root mean square (RMS) map $R_\sigma(x,y)$ is often associated with the data, when it is captured. For each wavelet coefficient $w_j(x,y)$ of R, the exact standard deviation $\sigma_j(x,y)$ has to be calculated from the root mean square map $R_\sigma(x,y)$ [41].

Assuming the noise is not correlated, a wavelet coefficient $w_j(x,y)$ is obtained by the correlation product between the image R and a function g_j, where x and y are discrete variables

$$w_j(x,y)=\sum_k \sum_l R(x,y) g_j(x+k,y+l). \quad (9)$$

Then we have

$$\sigma_j^2(x,y)=\sum_k \sum_l R_\sigma^2(x,y) g_j^2(x+k,y+l). \quad (10)$$

In the case of the à trous algorithm, the coefficients $g_j(x,y)$ are more easily computed by taking the wavelet transform of a discrete Dirac function (w^δ, in our notation):

$$g_{j(x,y)}=w_j^\delta(x,y). \quad (11)$$

Then the map σ_j^2 is calculated by correlating the square of the wavelet scale j of w^δ with $R_\alpha^2(x,y)$.

Other Kinds of Noise

Nonstationary Additive Noise

The noise is assumed to be locally Gaussian. So we must consider a noise standard deviation corresponding to an individual pixel and treat the problem as previously. The $R_\sigma(x,y)$ map can be obtained by taking the standard deviation in a box around each pixel.

Nonstationary Multiplicative Noise

The image is first log transformed. Then the transformed image is treated as an image with nonstationary additive noise.

Stationary Correlated Noise

The noise is stationary, but correlated. This noise modeling requires a noise map, containing a realization of the noise. The threshold at a scale j, S_j, is found by computing the wavelet transform of the noise map, and using the histogram of S_j to derive the PDF of S_j.

Undefined Noise

The standard deviation is estimated for each wavelet coefficient, by considering a box around it, and the calculation of σ is carried out in the same way as for nonstationary additive noise. The latter determines a map of variances for the image, which is used to derive the variances for the wavelet coefficients. In the case of "undefined noise" we do not assume additivity of the noise, and so we calculate the noise from local variance in the resolution scales.

Deterministic Information: Object Detection

New methods based on wavelet transforms have recently been developed for source extraction in an image [3], [47]. In the multiscale vision model [3], [44], an object in a signal is defined as a set of structures detected in the wavelet space. The wavelet transform algorithm used for such a decomposition is the à trous algorithm. The algorithm produces N images of the same size, each one containing only information in a given frequency band. From such images, we derive the multiresolution support which is then segmented to form structures. We describe the different steps leading towards structure being associated with an object.

Multiresolution Support and Its Segmentation

A multiresolution support of an image describes in a logical or Boolean way if an image I contains information at a given scale j and at a given position (x,y). If $M^{(I)}(j,x,y)=1$ (or $=$ *true*), then I contains information at scale j and at the position (x,y). M depends on several parameters:

▲ The input image.
▲ The algorithm used for the multiresolution decomposition.
▲ The noise.
▲ All additional constraints we want the support to satisfy.

Such a support results from the data, the treatment (noise estimation, etc.), and from knowledge on our part of the objects contained in the data (size of objects, linearity, etc.). In the most general case, *a priori* information is not available to us.

The multiresolution support of an image is computed in several steps:

▲ Step one is to compute the wavelet transform of the image.
▲ Binarization of each scale leads to the multiresolution support (the binarization of an image consists of assigning to each pixel a value only equal to zero or one).
▲ *A priori* knowledge can be introduced by modifying the support.

This last step depends on the knowledge we have of our images. For instance, if we know there is no interesting object smaller or larger than a given size in our image, we can suppress, in the support, anything which is due to that kind of object. This can often be done conveniently by the use of mathematical morphology [37], [20]. In the most general setting, we naturally have no information to add to the multiresolution support.

The multiresolution support will be obtained by detecting at each scale the significant coefficients. The multiresolution support is defined by

$$M(j,x,y)=\begin{cases}1, & \text{if } w_j(x,y) \text{ is significant}\\ 0, & \text{if } w_j(x,y) \text{ is not significant.}\end{cases} \tag{12}$$

Multiresolution Support Segmentation

The segmentation consists of labeling a Boolean image (0 or 1 values). Each group of connected pixels having a "1" value get a label value between 1 and $L_{\max}$, with $L_{\max}$ being the number of groups (components). This process is repeated separately at each scale of the multiresolution support. We define a "structure," S_j^i, as the group of components (the vertices comprising them are by definition significant) which has label i at a given scale j.

Multiscale Vision Model

An object is described as a hierarchical set of structures. The rule which allows us to connect two structures into a single object is called "interscale relation." Figure 1 shows how several structures at different scales are linked together and form objects. We have now to define the interscale relation: let us consider two structures at two successive scales, S_j^k and S_{j+1}^l. Each structure is located in one of the individual images of the decomposition and corresponds to a region in this image where the signal is significant. Denoting p_m the pixel position of the maximum wavelet coefficient value of S_j^k, S_j^k is said to be connected to S_{j+1}^l if S_{j+1}^l contains the pixel position p_m (i.e., the maximum position of the structure S_j^k must also be contained in the structure S_{j+1}^l). Several structures appearing in successive wavelet coefficient images can be connected in such a way, which we call an object in the interscale connectivity graph.

Reconstruction

Once an object is detected in wavelet space, it can be isolated by searching for the simplest function which presents the same signal in wavelet space. The problem of reconstruction [3] consists then of searching for a signal V such that its wavelet coefficients are the same as those of the detected structure. By denoting $\mathcal{T}$ the wavelet transform operator and P_b the projection operator in the subspace of the detected coefficients (i.e., setting to zero all coefficients at scales and positions where nothing was detected), the solution can be found by minimizing the following expression:

$$J(V)=\left\| W-(P_b \circ \mathcal{T})V \right\| \tag{13}$$

where W represents the detected wavelet coefficients of the signal. A complete description of algorithms for minimization of such a functional can be found in [3].

Statistical Information: Multiscale Entropy

The Concept of Multiscale Image Entropy

The idea behind multiscale entropy is that the information contained in an image or, more generally in a data set, is the addition of the information contained at different scales. One consequence is that the correlation between pixels is now taken into account when measuring the information. This was not the case with entropy definitions generally used in the past in image restoration. Major, and widely used, definitions of entropy, are:

▲ Burg [5]:

$$H_b(X)=-\sum_{k=1}^{N} \ln(X_k) \tag{14}$$

▲ Frieden [17]:

$$H_f(X)=-\sum_{k=1}^{N} X_k \ln(X_k) \tag{15}$$

▲ Gull and Skilling [23]:

$$H_g(X)=\sum_{k=1}^{N} X_k - M_k - X_k \ln\left(\frac{X_k}{M_k}\right) \tag{16}$$

where M is a given model, usually taken as a flat image. N is the number of pixels, and k represents an index pixel.

Each of these entropies correspond to different probability distributions that one can associate with an image [36]. The pixel distribution (correlation between pixels) is not considered in these definitions and implied by them is that two images having the same intensity histogram have the same entropy. Figure 2 illustrates this. The second image is obtained by distributing randomly the Saturn image pixel values, and the standard entropy definitions produce the same information measurement for both images. The concept of information becomes exclusively statistically. From the deterministic information viewpoint—the viewpoint of interest to the domain expert—the second image contains *less* information than the first one. For someone working on image transmission, it is clear that the second image will require more bits for lossless transmission, and from this point of view, he/she will consider that the second image contains *more* information. Finally, for data restoration—requiring use of both deterministic and statistical viewpoints—fluctuations due to noise are not of interest and do not contain relevant information. From this physical point of view, the standard definition of entropy seems badly adapted to information measurement in signal restoration.

It was discussed in [45] as to what a good entropy measurement for signal restoration or analysis should be, and we proposed that the following criteria should be verified:

▲ The information in a flat signal is zero.

▲ The amount of information in a signal is independent of the background.

▲ The amount of information is dependent on the noise. A given signal Y ($Y = X$ + Noise) does not furnish the same information if the noise is high or small.

▲ The entropy should treat a pixel with value $B + \varepsilon$ (B being the background) as equally informative as a pixel with value $B - \varepsilon$.

▲ The amount of information is dependent on the correlation in the signal. If a signal S presents large features above the noise, it contains a lot of information. By generating a new set of data from S, by randomly taking the pixel values in S, the large features will evidently disappear, and this new signal will contain less information. But the pixel values will be the same as in S.

It is clear that among all entropy functions proposed in the past, it is the Shannon one [38] which best respects these criteria. Indeed, if we assume that the histogram bin is defined as a function of the standard deviation of the noise, the first four points are verified, while none of these criteria are verified with other entropy functions (and only one point is verified for the Gull and Skilling entropy by taking the model equal to the background).

Following on from these criteria, a possibility is to consider that the entropy of a signal is the sum of the information at each scale of its wavelet transform [45], and the information of a wavelet coefficient is related to the probability of it being due to noise. Denoting h the information relative to a single wavelet coefficient, we have

$$H(X) = \sum_{j=1}^{l} \sum_{k=1}^{N_j} h(w_{j,k}) \tag{17}$$

with $h(w_{j,k}) = -\ln p(w_{j,k})$. l is the number of scales, and N_j is the number of samples in the band j ($N_j = N$ for the à trous algorithm). For Gaussian noise, we get

$$h(w_{j,k}) = \frac{w_{j,k}^2}{2\sigma_j^2} \tag{18}$$

where σ_j is the noise at scale j. We see that the information is proportional to the energy of the wavelet coefficients. The higher a wavelet coefficient, then the lower will be the probability, and the higher will be the information furnished by this wavelet coefficient. We can see easily that this entropy fulfills all our requirements. As for the Shannon entropy, the information increases with the entropy, and using such an entropy leads, for optimization purposes, to a minimum entropy method.

Since the data is composed of an original signal and noise, our information measure is corrupted by noise, and we decompose our information measure into two components, one (H_S) corresponding to the noncorrupted part, and the other (H_N) to the corrupted part. We have [45]

$$H(X) = H_S(X) + H_N(X). \tag{19}$$

We will define in the following H_S as the signal information, and H_N as the noise information. It must be clear that noise does not contain any information, and what we call "noise information" is a quantity which is measured as information by the multiscale entropy, and which is probably not informative to us.

If a wavelet coefficient is small, its value can be due to noise, and the information h relative to this single wavelet coefficient should be assigned to H_N. If the wavelet coefficient is high, compared to the noise standard deviation, its value cannot be due to the noise, and h should be assigned to H_S. h can be distributed as H_N or H_S, based on the probability $p_n(w_{j,k})$ that the wavelet coefficient is due to noise, or the probability $p_s(w_{j,k})$ that it is due to signal. We have $p_s(w_{j,k}) = 1 - p_n(w_{j,k})$. For the Gaussian noise case, we estimate $p_n(w_{j,k})$ that a wavelet coefficient is due to the noise by

$$\begin{aligned} p_n(w_{j,k}) &= \text{Prob}\left(W > \left|w_{j,k}\right|\right) \\ &= \frac{2}{\sqrt{2\pi}\sigma_j} \int_{|w_{j,k}|}^{+\infty} \exp\left(-W^2 / 2\sigma_j^2\right) dW \\ &= \text{erfc}\left(\frac{\left|w_{j,k}\right|}{\sqrt{2}\sigma_j}\right). \end{aligned} \tag{20}$$

For each wavelet coefficient $w_{j,k}$, we have to estimate now the fractions h_n and h_s of h which should be assigned to H_n and H_s. Hence signal information and noise information are defined by

$$\begin{aligned} H_s(X) &= \sum_{j=1}^{l} \sum_{k=1}^{N_j} h_s(w_{j,k}) \\ H_n(X) &= \sum_{j=1}^{l} \sum_{k=1}^{N_j} h_n(w_{j,k}). \end{aligned} \tag{21}$$

The idea for deriving h_s and h_n is the following: we imagine that the information h relative to a wavelet coefficient is a sum of small information components dh, each of them having a probability to be noise information, or signal information. Hence

$$h_n(w_{j,k}) = \int_0^{|w_{j,k}|} p_n\left(\left|w_{j,k}\right| - u\right) \left(\frac{\partial h(x)}{\partial x}\right)_{x=u} du \tag{22}$$

is the noise information relative to a single wavelet coefficient, and

$$h_s(w_{j,k}) = \int_0^{|w_{j,k}|} p_s\left(\left|w_{j,k}\right| - u\right) \left(\frac{\partial h(x)}{\partial x}\right)_{x=u} du \tag{23}$$

h_n is the signal information relative to a single wavelet coefficient. For Gaussian noise, we have

$$h_n(w_{j,k}) = \frac{1}{\sigma_j^2}\int_0^{|w_{j,k}|} u \operatorname{erfc}\left(\frac{|w_{j,k}| - u}{\sqrt{2}\sigma_j}\right) du$$
$$h_s(w_{j,k}) = \frac{1}{\sigma_j^2}\int_0^{|w_{j,k}|} u \operatorname{erf}\left(\frac{|w_{j,k}| - u}{\sqrt{2}\sigma_j}\right). \tag{24}$$

We verify easily that $h_n + h_s = h$.

Multiscale Entropy as a Measure of Relevant Information

Since the multiscale entropy extracts the information from the signal only, we ought to verify whether the astronomical content of an image is related to its multiscale entropy.

For this purpose, we studied the astronomical content of 200 images of 1024×1024 pixels extracted from scans of eight different plates carried out by the MAMA facility (Paris, France) [22] and stored at CDS (Strasbourg, France) in the Aladin archive [4]. We estimated the content of these images in three different ways:

▲ By counting the number of objects in an astronomical catalog (USNO A2.0 catalog) within the image. The United States Naval Observatory (USNO) catalog was obtained by source extraction from the same survey plates as we used in our study.

▲ By counting the number of objects estimated in the image by the Sextractor object detection package [2]. As in the case of USNO these detections are mainly point sources (stars, as opposed to spatially extended objects like galaxies).

▲ By counting the number of structures detected at several scales using the MR/1 multiresolution analysis package [34].

Figure 3 shows the results of plotting these numbers for each image against the multiscale signal entropy of the image. The best results are obtained using the MR/1 package, followed by Sextractor and then by the number of sources extracted from USNO. Of course the latter two basically miss the content at large scales, which is taken into account by MR/1.

Sextractor and multiresolution methods were also applied to a set of CCD images from CFH UH8K, 2MASS, and the DENIS near-infrared surveys. Results obtained were very similar to what was obtained above. This seems to point to multiscale entropy as being a universally applicable measurement of image content and certainly for the type of scientific image used here.

Subsequently we looked for the relation between the multiscale entropy and the optimal compression rate of an image which we can obtain by multiresolution techniques [44]. By optimal compression rate we mean a compression rate which allows all the sources to be preserved and which does not degrade the astrometry and photometry. Louys et al. [30] and Couvidat [7] have estimated this optimal compression rate using the compression program of the MR/1 package [34].

Figure 4 shows the relation obtained between the multiscale entropy and the optimal compression rate for all the images used in our previous tests including CCD images. The power law relation is obvious, thus allowing us to conclude that:

▲ The compression rate depends strongly on the astronomical content of the image. We can then say that compressibility is also an estimator of the content of the image.

▲ The multiscale entropy allows us to predict the optimal compression rate of the image.

Filtering from the Multiscale Entropy

The multiscale entropy filtering method (MEF) [45], [43] consists of measuring the information h relative to wavelet coefficients and of separating this into two parts h_s, and h_n. The expression h_s is called the signal information and represents the part of h which is definitely not contaminated by the noise. The expression h_n is called the noise information and represents the part of h which may be contaminated by the noise. We have $h = h_s + h_n$. Following this notation, the corrected coefficient $\tilde{w}$ should minimize:

$$J(\tilde{w}_j) = h_s(w_j - \tilde{w}_j) + \alpha h_n(\tilde{w}_j) \tag{25}$$

i.e., there is a minimum of information in the residual $(w - \tilde{w})$ which can be due to the significant signal, and a minimum of information which could be due to the noise in the solution $\tilde{w}_j$.

Simulations have shown [43] that the MEF method produces a better result than the standard soft or hard thresholding, from both the visual aspect and PSNR (peak signal-to-noise ratio). Figures 5 and 6 show the filtering respectively on simulated noisy blocks and on a real spectrum.

Deconvolution from the Multiscale Entropy

Consider an image characterized by its intensity distribution (the "data") I, corresponding to the observation of a "real image" O through an optical system. If the imaging system is linear and shift-invariant, the relation between the data and the image in the same coordinate frame is a convolution:

$$I = O * P + N. \tag{26}$$

P is the point spread function (PSF) of the imaging system, and N is additive noise. In practice $O * P$ is subject to nonstationary noise which one can tackle by simultaneous object estimation and restoration [25]. The issue of more extensive statistical modeling will not be further addressed here (see [28], [29], and [33]), beyond noting that multiresolution frequently represents a useful framework, allowing the user to introduce *a priori* knowledge of objects of interest.

We want to determine $O(x, y)$ knowing I and P. This inverse problem has led to a large amount of work, the main difficulties being the existence of: i) a cut-off frequency of the point spread function and ii) the additive noise (see, for example, [6], [18], and [19]).

Equation (26) is usually in practice an ill-posed problem. This means that there is not a unique solution.

The most realistic solution is that which minimizes the amount of information, but remains compatible with the data. By the MEM method, minimizing the information is equivalent to maximizing the entropy and the functional to minimize is

$$J(O) = \sum_{k=1}^{N} \frac{(I_k - (P * O)_k)^2}{2\sigma_I^2} + \alpha H(O) \tag{27}$$

where H is either the Frieden or the Gull and Skilling entropy.

Similarly, using the multiscale entropy, minimizing the information is equivalent to minimizing the entropy and the functional to minimize is

$$J(O) = \sum_{k=1}^{N} \frac{\left(I_k - (P * O)_k\right)^2}{2\sigma_I^2} + \alpha H(O). \tag{28}$$

We have seen that in the case of Gaussian noise, H is given by the energy of the wavelet coefficients. We have

$$J(O) = \sum_{k=1}^{N} \frac{\left(I_k - (P * O)_k\right)^2}{2\sigma_I^2} + \alpha \sum_{j=1}^{l} \sum_{k=1}^{N_j} \frac{w_{j,k}^2}{2\sigma_j^2} \tag{29}$$

where σ_j is the noise at scale j, N_j the number of pixels at the scale j, σ_I the noise standard deviation in the data, and l the number of scales.

Rather than minimizing the amount of information in the solution, we may prefer to minimize the amount of information which can be due to the noise. The function is now

$$J(O) = \sum_{k=1}^{N} \frac{\left(I_k - (P * O)_k\right)^2}{2\sigma_I^2} + \alpha H_n(O). \tag{30}$$

The solution is found by computing the gradient $\nabla(J(O))$ and performing the following iterative schema:

$$O^{n+1} = O^n - \gamma \nabla(J(O^n)) \tag{31}$$

Conclusion

The perspectives opened up by:

▲ The range of noise models, catering for a wide range of eventualities in physical science imagery and signals, and

▲ The new two-pronged but tightly coupled understanding of the concept of information

have given rise to better quality results in applications such as noise filtering, deconvolution, compression, and object (feature) detection. We have illustrated some of these new results in this article, and others can be found in the references. The theoretical foundations of our perspectives have been sketched out. The practical implications, too, are evident from the range of important signal processing problems which we can better address with this armory of methods.

Even broader and more ambitious objectives follow. As a launch pad for content-based image retrieval, for instance, the results and perspectives described in this article are of importance. The results described in this work are targeted at information and at relevance.

While we have focused on experimental results in astronomical image and signal processing, the possibilities are apparent in many other application domains.

Acknowledgement

The authors are grateful for the comments of two anonymous referees.

Fionn Murtagh has engineering and mathematics degrees and an M.Sc. in computer science from Trinity College Dublin, a Ph.D. in mathematical statistics from Universite P. & M. Curie, Paris 6, and an habilitation from Universite L. Pasteur, Strasbourg 1. He is a Full Professor in computer science at Queen's University, Belfast, and an Adjunct Professor at the Astronomical Observatory, Strasbourg. He worked for 12 years for the Space Science Department, European Space Agency, on the Hubble Space Telescope. He has published ten books and 150 journal papers. He is Editor-in-Chief of *The Computer Journal* and a member of the editorial boards of other journals.

Jean-Luc Starck has a Ph.D. from University Nice-Sophia Antipolis and an Habilitation from Universite Paris XI. He was visitor at the Europoean Southern Observatory (ESO) in 1993 and at Stanford statistics department in 2000. He has been a Researcher at the CEA since 1994 and is on the Infrared Space Observatory (ISO) project. He has published one book and about 50 journal papers on wavelet and multiresolution methods in astrophysics and other fields.

References

[1] F.J. Anscombe, "The transformation of Poisson, binomial and negative-binomial data." *Biometrika*, vol. 15, pp. 246-254, 1948.

[2] E. Bertin and S. Arnouts, "Sextractor: Software for source extraction," *Astron. Astrophysi*, vol. 117, pp. 393-404, June 1996.

[3] A. Bijaoui and F. Rué, "A multiscale vision model adapted to astronomical images," *Signal Processing*, vol. 46, pp. 229-243, 1995.

[4] F. Bonnarel, P. Fernique, F. Genova, J.G. Bartlett, O. Bienaymé, D. Egret, J. Florsch, H. Ziaeepour, and M. Louys, "ALADIN: A reference tool for identification of astronomical sources, in *Astronomical Data Analysis Soft-*

ware and Systems VIII, ASP Conf. Series, D.M. Mehringer, R.L. Plante, and D.A. Roberts, Ed. 1999, pp. 229-232.

[5] J.P. Burg, "Annual meeting international society exploratory geophysics," in *Modern Spectral Analysis*, D.G. Childers, Ed. New York: IEEE, 1978, pp. 34-41.

[6] T.J. Cornwell, "Image restoration," in *Diffraction-Limited Imaging with Very Large Telescopes*, D.M. Alloin and J.M. Mariotti, Eds. Dordrecht, Germany: Kluwer, 1989. **AU: please supply page numbers**

[7] S. Couvidat, DEA dissertation, Strasbourg Observatory, 1999 **AU: please supply full name of dissertation and location of Strasbourg Observatory**.

[8] F. Damiani, A. Maggio, G. Micela, and S. Sciortino, "A method based on wavelet transforms for source detection in photon-counting detector images. I. Theory and general properties," *Astrophys. J.*, vol. 483, pp. 350-369, July 1997.

[9] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 1992.

[10] D.L. Donoho, "Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data," in *Proc. Sympo. Applied Mathematics*, 47, 1993.**AU: please supply page numbers. Is 57 teh volume?**

[11] D.L. Donoho, "Translation-invariant de-noising." in *Wavelets and Statistics*, A. Antoniadis and G. Oppenheim, Eds. New York: Springer-Verlag, 1995. **AU: please supply page numbers**

[12] D.L. Donoho and I.M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," Tech. Rep. 400, Stanford Univ, 1993.

[13] E. Escalera and A. Mazure, "Wavelet analysis of subclustering—An illustration, Abell 754," *Astrophys. J.*, vol, 388, pp. 23-32, 1992.

[14] E. Escalera, E. Slezak, and A. Mazure, "New evidence for subclustering in the coma cluster using the wavelet analysis," *Astron. Astrophys.*, vol. 264, pp. 379-384, Oct. 1992.

[15] J.C. Feauveau, "Analyse multirésolution par ondelettes non-orthogonales et bancs de filtres numériques," Ph.D. dissertation, Université Paris Sud, 1990.

[16] P.E. Freeman, V. Kashyap, R. Rosner, R. Nichol, B. Holden, and D.Q. Lamb, "X-ray source detection using the wavelet transform," in *ASP Conf. Ser. 101: Astronomical Data Analysis Software and Systems V*, vol. 5. 1996, pp. 163.

[17] B.R. Frieden, *Image Enhancement and Restoration*. Berlin, Germany: Springer-Verlag, 1978.

[18] V. Di Gèsu and M.C. Maccarone, "A direct deconvolution method for two dimensional image analysis." in *Proc. Symp. COMPSTAT 1982*. Physica-Verlag, 1982.

[19] V. Di Gèsu and M.C. Maccarone, "The Bayesian direct deconvolution method: Properties and applications," *Signal Processing*, vol. 6, pp. 201-211, 1984.

[20] V. Di Gèsu, M.C. Maccarone, and M. Tripiciano, "Mathematical morphology based on fuzzy operators," in *Fuzzy Logic: State of the Art*. Norwell, MA: Kluwer, 1993. **AU: Please provide page numbers and editor of book**

[21] S.A. Grebenev, W. Forman, C. Jones, and S. Murray, "Wavelet transform analysis of the small-scale X-ray structure of the cluster Abell 1367," *Astrophys. J.*, vol. 445, pp. 607-623, 1995.

[22] J. Guibert, "The MAMA facility—A survey of scientific programmes," in *Digitised Optical Sky Surveys*. 1992, pages 103+, 1992. **AU: please provide publisher (and its location) and complete page numbers**

[23] S.F. Gull and J. Skilling, *MEMSYS5 Quantified Maximum Entropy User's Manual*. 1991.**AU: please provide publisher (and its location)**

[24] G. Jammal and A. Bijaoui, "Multiscale image restoration for photon imaging systems," in *Proc. SPIE Conf. Signal and Image Processing: Wavelet Applications in Signal and Image Processing VII*, 1999. **AU: please provide page numbers**

[25] A.K. Katsaggelos, *Digital Image Restoration*. Berlin, Germany: Springer-Verlag, 1991.

[26] E. Kolaczyk and D. Dixon, "Nonparametric estimation of intensity maps using Haar wavelets and Poisson noise characteristics," *Astrophys. J.*, vol. 534, no. 1, pp. 490-505, 2000.

[27] E.D. Kolaczyk, "Nonparametric estimation of gamma-ray burst intensities using Haar wavelets," *Astrophys. J.*, vol. 483, pp. 340-349, July 1997.

[28] J. Llacer and J.Núñez, "Iterative maximum likelihood estimator and Bayesian algorithms for image reconstruction in astronomy," in *The Restoration of HST Images and Spectra*, R.L. White and R.J. Allen, Eds. Baltimore, MD: Space Telescope Science Institute, 1990.

[29] H. Lorenz and G.M. Richter, "Adaptive filtering of HST images: preprocessing for deconvolution," in *Science with the Hubble Space Telescope*, P. Benvenuti and E. Schreier, Eds. Garching: European Southern Observatory, 1993, pp. 203-206.

[30] M. Louys, J.-L. Starck, S. Mei, F. Bonnarel, and F. Murtagh, "Astronomical image compression," *Astron. Astrophys.*, vol. 136, pp. 579-590, 1999.

[31] S.Mallat, *A Wavelet Tour of Signal Processing*. New York: Academic, 1998.

[32] S.G. Mallat, "A theory for multiresolution signal decomposition: the waveletrepresentation," *IEEE Trans. Pattern Anal. Machine Intelll.*, vol. 11, no. 7, pp. 674-693, 1989.

[33] R. Molina, "On the hierarchical Bayesian approach to image restoration. Applications to astronomical images," *IEEE Trans. Pattern Anal. Machine Intelll.*, vol. 16, pp. 1122-1128, 1994.

[34] MR/1. Multiresolution Image and Data Analysis Software Package, Version 2.0, Multi Resolutions Ltd., 1999 [Online]. Available: http://www.multiresolution.com.

[35] F. Murtagh, J.L. Starck, and A. Bijaoui, "Image restoration with noise suppression using a multiresolution support," *Astron. Astrophys.*, vol. 112, pp. 179-189, 1995.

[36] R. Narayan and R. Nityananda, "Maximum entropy image restoration in astronomy," *Ann. Rev. Astron. Astrophys.*, vol. 24, pp. 127-170, 1986.

[37] J. Serra, *Image Analysis and Mathematical Morphology*. London, U.K.: Academic, 1982.

[38] C.E. Shannon, "A mathematical theory for communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379-423, 1948.

[39] M.J. Shensa, "Discrete wavelet transforms: Wedding the à trous and Mallat algorithms," *IEEE Trans. Signal Processing*, vol. 40, pp. 2464-2482, 1992.

[40] J.L. Starck, "Wavelet transform and X-ray images," Tech. Rep., CEA, 1999.

[41] J.L. Starck, H. Aussel, D. Elbaz, D. Fadda, and C. Cesarsky, "Faint source detection in ISOCAM images," *Astron. Astrophys.*, vol. 138, pp. 365-379, 1999.

[42] J.L. Starck and F. Murtagh, "Automatic noise estimation from the multiresolution support," *Pub. Astronom. Soc. Pacific*, vol. 110, no. 744, pp. 193-199, 1998.

[43] J.L. Starck and F. Murtagh, "Multiscale entropy filtering," *Signal Processing*, vol. 76, no. 2, pp. 147-165, 1999.

[44] J.L. Starck, F. Murtagh, and A. Bijaoui, *Image Processing and Data Analysis: The Multiscale Approach*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[45] J.L. Starck, F. Murtagh, and R. Gastaud, "A new entropy measure based on the wavelet transform and noise modeling," *IEEE Trans. Circ. Syst. II*, vol. 45, no. 8, 1998. **AU: please provide page numbers**

[46] J.L. Starck, F. Murtagh, B. Pirenne, and M. Albrecht, "Astronomical image compression based on noise suppression, *Pub. Astronom. Soc. Pacific*, vol. 108, pp. 446-455, 1996.

[47] J.L. Starck, R. Siebenmorgen, and R. Gredel. "Spectral analysis by the wavelet transform," *Astrophys. J.*, vol. 482, pp. 1011-1020, 1997.

▲ *1. Example of connectivity in the wavelet space: contiguous significant wavelet coefficients form a structure, and following an interscale relation, a set of structures form an object. Two structures S_j, S_{j+1} at two successive scales belong to the same object if the position pixel of the maximum wavelet coefficient value of S_j is included in S_{j+1}.*

▲ *2. Saturn image (left) and the same data distributed differently (right). These two images have the same entropy, using any of the standard entropy definitions.*

▲ *3. Multiscale entropy (vertical) versus the number of objects (horizontal): the number of objects is, respectively, obtained from (top) the USNO catalog, (middle) the Sextractor package, and (bottom) the MR/1 package.*

▲ *4. Multiscale entropy (vertical) of astronomical images versus the optimal compression ratio (horizontal). Images which contain a high number of sources have a small ratio and a high multiscale entropy value. The relation is almost linear.*

▲ *5. Top, noisy blocks and filtered blocks overplotted. Bottom, filtered blocks.*

▲ *6. Top, real spectrum and filtered spectrum overplotted. Bottom, filtered spectrum.*

The Haar transform is less efficient for restoring astronomical images but its simplicity, and the computation time, may be attractive in practice.