39

C.E. SACLAY
DSM

SERVICE D'ÉLECTRONIQUE ET D'INFORMATIQUE

FERMILAB
SEP 21 1999
LIBRARY

DAPNIA-SEI-99-02

Fermilab Library
0 1160 0070446 4

NAME | I.D. No. | M.S.

DAPNIA

DAPNIA/SEI 99-02

August 1999

MULTISCALE ENTROPY

J.L. Starck, F. Murtagh, F. Bonnarel, S. Couvidat

Submitted to the IEEE Transactions on Information Theory
Special issue on Information-Theoretic Imaging

Couverture : Sylvie Cabrit et Marc Sauvage

Le DAPNIA (**D**épartement d'**A**strophysique, de physique des **P**articules, de physique **N**ucléaire et de l'**I**nstrumentation **A**ssociée) regroupe les activités du Service d'Astrophysique (SAp), du Département de Physique des Particules Elémentaires (DPhPE) et du Département de Physique Nucléaire (DPhN).

Adresse : DAPNIA, Bâtiment 141
CEA Saclay
F - 91191 Gif-sur-Yvette Cedex

Multiscale Entropy

J.-L. Starck

DAPNIA/SEI-SAP, CEA/Saclay, 91191 Gif sur Yvette, France

F. Murtagh

School of Computer Science, The Queen's University of Belfast

Belfast BT7 1NN, Northern Ireland

F. Bonnarel and S. Couvidat

CDS, Observatoire Astronomique de Strasbourg Observatory

11, rue de l'Université, 67000 Strasbourg, France

August, 1999

Submitted to the IEEE Transactions on Information Theory

Special issue on Information-Theoretic Imaging

Abstract

Multiscale entropy is based on the wavelet transform and noise modeling. It is a means of measuring information in a data set. It has been recently developed and has been applied successfully to signal and image filtering. We describe in this paper how it can be used for deconvolution, background fluctuation analysis, and astronomical image content analysis. A range of examples illustrates the results.

Index Terms

Wavelet transform, filtering, deconvolution, image restoration, Bayesian estimation, entropy

Number of pages: 34

Number of figures: 11

List of Figures

1 Lena image (left) and the same data distributed differently (right). These two images have the same entropy, using any of the standard entropy definitions. 25

2 Corrected wavelet coefficient versus the wavelet coefficient with different α values (from the top curve to the bottom one, α is respectively equal to 0,0.1,0.5, 1, 2, 5,10). 26

3 Top, noisy blocks and filtered blocks overplotted. Bottom, filtered blocks. . 27

4 Top, real spectrum and filtered spectrum overplotted. Bottom, filtered spectrum. 27

5 Top left, simulated noisy image, top right, hard thresholding, bottom left, soft thresholding, and bottom right, MEF. 28

6 Upper left, Lena image convolved with a Gaussian (FWHM = 5) plus Gaussian noise ($\sigma = 5$). Upper right, deconvolved image by the Multiscale Entropy Deconvolution Method. Bottom left and right, zoom on a part of the previous images. 29

7 (a) Original image, (b) blurred image + Gaussian noise, (c) deconvolved image, and (d) residual image. 30

8 Mean entropy versus the scale of 5 simulated images containing undetectable sources and noise. Each curve corresponds to the multiscale transform of one image. From top to bottom, the image contains respectively 400, 200, 100, 50 and 0 sources. 31

9 Region of a simulated image containing an undetectable source at the center. 32

10 Multiscale entropy versus the number of objects: the number of objects is, respectively, obtained from (top) the USNO catalog, (middle) the Sextractor package, and (bottom) the MR/1 package. 33

11 Multiscale entropy of astronomical images versus the optimal compression ratio. Images which contain a high number of sources have a small ratio and a high multiscale entropy value. The relation is almost linear. 34

1 Introduction

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 previous entropy definitions generally used in image restoration. The main entropy definitions to date are:

- Burg [3]: $H_b(X) = -\sum_{pixels} \ln(X)$
- Frieden [11]: $H_f(X) = -\sum_{pixels} X \ln(X)$
- Gull and Skilling [13]: $H_g(X) = \sum_{pixels} X - M - X \ln(X|M)$

Each of these entropies correspond to different probability distributions that one can associate with an image [23]. The pixel distribution (correlation between pixels) is absolutely not considered and the above definitions imply that two images having the same intensity histogram have the same entropy. Fig. 1 illustates this perfectly. The second image is obtained by distributing randomly the Lena image pixel values, and the standard entropy definitions produce the same information measurement for both images. The concept of information becomes really subjective, or at least it depends on the application domain. Indeed, for someone who is not involved in image processing, 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, all 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.

A discussion was raised in [30] about what should be a good entropy measurement for signal restoration, and we proposed that the following criteria should be verified:

1. The information in a flat signal is zero.

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

3. The amount of information is dependent on the noise. A given signal Y ($Y = X+$ Noise) doesn't furnish the same information if the noise is high or small.

4. The entropy must work in the same way for a pixel which has a value $B+\epsilon$ (B being the background), and for a pixel which has a value $B-\epsilon$.

5. 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 [25] 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).

We review in the second section the principle of multiscale entropy and its application to filtering. Sections 3, 4 and 5 present, respectively, how it can be used for image deconvolution, background fluctuation analysis, and astronomical image content analysis.

2 Multiscale Entropy

2.1 Definition

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 [30], 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

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

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

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

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 non-corrupted part, and the other (H_N) to the corrupted part. We have [30]

$$H(X) = H_S(X) + H_N(X) \tag{3}$$

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

$$\begin{aligned} p_n(w_{j,k}) = \text{Prob}(W > \mid w_{j,k} \mid) &= \frac{2}{\sqrt{2\pi}\sigma_j} \int_{|w_{j,k}|}^{+\infty} \exp(-W^2/2\sigma_j^2) dW \\ &= \text{erfc}\left(\frac{\mid w_{j,k} \mid}{\sqrt{2}\sigma_j}\right) \end{aligned} \tag{4}$$

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

$$\begin{aligned} H_s(X) &= \sum_{j=1}^{l} \sum_{k=1}^{N_j} h_s(w_{j,k}) \\ H_n(X) &= \sum_{j=1}^{l} \sum_{k=1}^{N_j} h_n(w_{j,k}) \end{aligned} \tag{5}$$

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 and h_n are calculated by:

$$h_n(w_{j,k}) = \int_0^{|w_{j,k}|} p_n(\mid w_{j,k} \mid -u) \left(\frac{\partial h(x)}{\partial x}\right)_{x=u} du \tag{6}$$

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

$$h_s(w_{j,k}) = \int_0^{|w_{j,k}|} p_s(\mid w_{j,k} \mid -u) \left(\frac{\partial h(x)}{\partial x}\right)_{x=u} du \tag{7}$$

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

$$\begin{aligned} h_n(w_{j,k}) &= \frac{1}{\sigma_j^2} \int_0^{|w_{j,k}|} u \ \text{erfc}\left(\frac{\mid w_{j,k} \mid -u}{\sqrt{2}\sigma_j}\right) du \\ h_s(w_{j,k}) &= \frac{1}{\sigma_j^2} \int_0^{|w_{j,k}|} u \ \text{erf}\left(\frac{\mid w_{j,k} \mid -u}{\sqrt{2}\sigma_j}\right) \end{aligned} \tag{8}$$

2.2 Filtering

The problem of filtering or restoring data D can be expressed by the following: We search for a solution $\tilde{D}$ such that the difference between D and $\tilde{D}$ minimizes the information due

to the signal, and such that $\tilde{D}$ minimizes the information due to the noise.

$$J(\tilde{D}) = H_s(D - \tilde{D}) + H_n(\tilde{D}) \tag{9}$$

Furthermore, the smoothness of the solution can be controlled by adding a parameter:

$$J(\tilde{D}) = H_s(D - \tilde{D}) + \alpha H_n(\tilde{D}) \tag{10}$$

In practice [4], we minimize for each wavelet coefficient $w_{j,k}$:

$$j(\tilde{w}_{j,k}) = h_s(w_{j,k} - \tilde{w}_{j,k}) + \alpha h_n(\tilde{w}_{j,k}) \tag{11}$$

The solution is found by first computing the gradient $\nabla(J(\tilde{w}_{j,k}))$ (see Appendix):

$$\nabla(j(\tilde{w}_{j,k})) = \int_0^{w_{j,k}-\tilde{w}_{j,k}} p_s(u)du + \alpha(\tilde{w}_{j,k} - \int_0^{\tilde{w}_{j,k}} p_s(u)du) \tag{12}$$

which gives for the Gaussian case (see Appendix):

$$\begin{aligned}\nabla(j(\tilde{w}_{j,k})) \quad = \quad & -\frac{w_{j,k} - \tilde{w}_{j,k}}{\sigma_j^2}\mathrm{erf}\left(\frac{w_{j,k} - \tilde{w}_{j,k}}{\sqrt{2}\sigma_j}\right) + \sqrt{\frac{2}{\pi}}\frac{1}{\sigma_j}\left[1 - e^{-\frac{(w_{j,k}-\tilde{w}_{j,k})^2}{2\sigma_j^2}}\right] + \\ & \alpha\left(\frac{\tilde{w}_{j,k}}{\sigma_j^2}\mathrm{erfc}\left(\frac{\tilde{w}_{j,k}}{\sqrt{2}\sigma_j}\right) + \frac{1}{\sigma_j}\sqrt{\frac{2}{\pi}}\left[1 - e^{-\frac{\tilde{w}_{j,k}^2}{2\sigma_j^2}}\right]\right)\end{aligned} \tag{13}$$

The solution of the equation $\nabla(j(\tilde{w}_{j,k})) = 0$ can be obtained by any minimization routine. In our examples, we have used a simple dichotomy.

Fig. 2 shows the result when minimizing the functional j with different α values, and a noise standard deviation equal to 1. The corrected wavelet coefficient is plotted versus the wavelet coefficient. From the top curve to the bottom one, α is respectively equal to 0, 0.1, 0.5, 1, 2, 5, 10. The higher the value of α, the more the corrected wavelet coefficient is reduced. When α is equal to 0, there is no regularization and the data are unchanged.

Simulations have shown [28] 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). Figs. 3 and 4 show the filtering respectively on simulated noisy blocks and on a real spectrum.

Fig. 5 shows a simulation. The four sub-images present respectively the simulated noisy image, the hard and soft thresholded images, and the MEF filtered one.

A robust way to constrain α is to use the fact that we expect a residual with a given standard deviation at each scale j equal to the noise standard deviation σ_j at the same scale. Then rather than a single α we have an α_j per scale. A full description of the MEF algorithm can be found in [28].

3 Deconvolution

3.1 Introduction

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:

$$I = O * P + N \tag{14}$$

P is the point spread function (PSF) of the imaging system, and N is additive noise. In practice $O * P$ is subject to non-stationary noise which one can tackle by simultaneous object estimation and restoration [15]. The issue of more extensive statistical modeling will not be further addressed here (see [16, 17, 21]), 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 [7]).

Eqn. 14 is usually in practice an ill-posed problem. This means that there is not a unique solution.

3.2 The principle

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

$$J(O) = \sum_{k=1}^{N} \frac{(I_k - (P * O)_k)^2}{2\sigma_I^2} - \alpha H(O) \tag{15}$$

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

$$J(O) = \sum_{k=1}^{N} \frac{(I_k - (P * O)_k)^2}{2\sigma_I^2} + \alpha H(O) \tag{16}$$

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

$$J(O) = \sum_{k=1}^{N} \frac{(I_k - (P * O)_k)^2}{2\sigma_I^2} + \alpha \sum_{j=1}^{l} \sum_{k=1}^{N_j} \frac{w_{j,k}^2}{2\sigma_j^2} \tag{17}$$

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:

$$J(O) = \sum_{k=1}^{N} \frac{(I_k - (P * O)_k)^2}{2\sigma_I^2} + \alpha H_n(O) \tag{18}$$

and for Gaussian noise, H_n has been defined by

$$H_n(X) = \sum_{j=1}^{l} \sum_{k=1}^{N_j} \frac{1}{\sigma_j^2} \int_0^{|w_{j,k}|} u \text{ erf} \left(\frac{| w_{j,k} | - u}{\sqrt{2}\sigma_j} \right) \tag{19}$$

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

$$O^{n+1} = O^n - \gamma \nabla(J(O^n)) \tag{20}$$

We consider an α_j per scale, and introduce thereby a kind of adaptive regularization which depends on the signal-to-noise ratio of the input data wavelet coefficients.

3.3 The parameters

In order to introduce flexibility in the way we restore the data, we introduce two parameters $\beta_{j,k}$ and $\alpha_{j,k}$ which allow us to weight, respectively, the two terms of the equation to be minimized:

$$J(O) = \frac{1}{2\sigma_I^2} \sum_{k=1}^{N} \left(\sum_j \sum_l \beta_{j,k} w_{j,l}(R) \psi_{j,l}(k) \right)^2 + \sum_{j=1}^{l} \sum_{k=1}^{N_j} \alpha_{j,k} h(w_{j,k}(O))$$

where $R = I - P * O$, and $R = \sum_j \sum_k w_{j,k}(R)\psi_{j,k}$ ($w_{j,k}(R)$ are the wavelet coefficients of R, and $w_{j,k}(O)$ are the wavelet coefficients of O).

We consider three approaches for estimating $\beta_{j,k}$

1. No weighting: $\beta_{j,k} = 1$
2. Soft weighting: $\beta_{j,k} = p_s(w_{j,k}(I))$

   In this case, $\beta_{j,k}$ is equal to the probability that the input data wavelet coefficient is due to signal (and not to noise).
3. Hard weighting: $\beta_{j,k} = 0$ or 1 depending on $p_n(w_{j,k}(I))$ ($p_n(w_{j,k}(I)) = 1 - p_s(w_{j,k}(I))$).

   This corresponds to using only significant input data wavelet coefficients.

$\alpha_{j,k}$ is the product of two values: $\alpha_{j,k} = \alpha_u \beta'_{j,k}$.

- α_u is a user parameter (defaulted to 1) which allows us to control the smoothness of the solution. Increasing α_u produces a smoother solution.
- $\beta'_{j,k}$ depends on the input data and can take the following value:
  1. No regularization ($\beta'_{j,k} = 0$): only the first term of the functional is minimized.
  2. No protection from regularization ($\beta'_{j,k} = 1$): the regularization is applied at all positions and at all the scales.
  3. Soft protection ($\beta'_{j,k} = p_n(w_{j,k}(I))$): the regularization becomes adaptive, depending on the probability that the input wavelet coefficient is due to noise.
  4. Hard protection ($\beta'_{j,k} = 0$ or 1 depending on $p_n(w_{j,k}(I))$).

5. Soft + hard protection: ($\beta'_{j,k} = 0$ or $p_n(w_{j,k}(I))$ depending on $p_n(w_{j,k}(I))$).

We easily see that choosing a hard weighting and no regularization leads to deconvolution from the multiresolution support [27, 29].

3.4 The choice of the wavelet transform algorithm

There are many 2D WT algorithms [29]. The most well-known are perhaps the orthogonal wavelet transform proposed by Mallat [20], and its bi-orthogonal version [6]. These methods are based on the principle of reducing the redundancy of the information in the transformed data. Other WT algorithms exist, such as the Feauveau algorithm [10] (which is an orthogonal transform, but using an isotropic wavelet), the à trous algorithm which is non-orthogonal and furnishes a very redundant data set [14], or the Dyadic Wavelet Transform (DWT) [19] which is even more redundant than the à trous algorithm. All these methods have advantages and drawbacks. Non-decimated methods take more computational time and memory space, but produce better solutions. This can be explained by the fact that decimation violates the sampling theorem. Indeed, the decimation is done by first filtering the data, and then taking one pixel out of two. But the filtering step does not reduce strictly the frequency band by two. If the coefficients are modified, artifacts become clearly visible. For good quality signal or image restoration, a non-decimated method should always be preferred. The choice between the à trous algorithm and the DWT depends on the data. For images with contours, the DWT is better. For images containing mainly isotropic features such as astronomical images, or some types of medical imagery, the à trous algorithm is well adapted. The multiscale entropy method can be combined with any WT algorithm.

3.5 Examples

Fig. 6 shows a simulation. Upper left, the Lena image has been convolved by a Gaussian with a full width at half maximum (FWHM) equal to 5, and Gaussian noise has been

added. The noise standard deviation is equal to 5. Fig. 6 upper right shows the deconvolved image. The dyadic wavelet transform has been used. Fig. 6 bottom left and right shows a zoom of a part of the upper left and right figures. The PSNR of the blurred image is equal to 27.74 dB, and PSNR of the restored image is 30.70 dB.

Fig. 7 shows another simulation. The original image (panel (a)) contains stars and galaxies. Fig. 7b shows the data (blurred image + Gaussian noise), Fig. 7c shows the deconvolved image, and Fig. 7d the residual image (i.e. data minus solution reconvolved by the PSF). The blurred image SNR is 12dB, and the deconvolved image SNR is 23.11 dB.

4 Multiscale Entropy applied to Background Fluctuation Analysis

The mean entropy vector

The multiscale entropy has been defined by:

$$H(X) = \sum_{j=1}^{l} \sum_{k=1}^{N} h(w_j) \tag{21}$$

with $h(w_j = \ln(p(w_j(k))))$. In order to study the behavior of the information at a given scale, we prefer to calculate the mean entropy vector E defined by:

$$E(j) = \frac{1}{N} \sum_{k=1}^{N} h(w_j) \tag{22}$$

$E(j)$ gives the mean entropy at the scale j. From the mean entropy vector, we have statistical information on each scale separately. Having a noise model, we are able to calculate (generally from simulations) the mean entropy vector $E^{(noise)}(j)$ resulting from pure noise. Then we define the normalized mean entropy vector by

$$E_n(j) = \frac{E(j)}{E^{(noise)}(j)} \tag{23}$$

Fig. 8 shows the result of a simulation. Five simulated images were created by adding n sources to a 1024 $\times$ 1024 image containing Gaussian noise of standard deviation equal

to 1. The n sources are identical, with a maximum equal to 1, and standard deviation equal to 2. Defining the signal-to-noise ratio (SNR) as the ratio between the standard deviation in the smallest box which contains at least 90% of the flux of the source, and the noise standard deviation, we have a SNR equal to 0.25. The sources are not detectable in the simulated image, nor in its wavelet transform. Fig. 9 shows a region which contains a source at the center. It is clear there is no way to find this kind of noisy signal. The five images were created using a number of sources respectively equal to 0, 50, 100, 200 and 400, and the simulation was repeated ten times with different noise maps in order to have an error bar on each entropy measurement. For the image which contains 400 sources, the number of pixels affected by a source is less than 2.5%.

When the number of sources increases, the difference beteen the multiscale entropy curves increases. Even if the sources are very faint, the presence of signal can be clearly detected using the mean entropy vector. But it is obvious that the positions of these sources remain unknown.

5 Multiscale Entropy as a Measure of Relevant Information in an Image

Since the multiscale entropy extracts the information from the signal only, it was a challenge to see if the astronomical content of an image was related to its multiscale entropy.

For this purpose, we studied the astronomical content of 200 images of 1024 × 1024 pixels extracted from scans of 8 different plates carried out by the MAMA facility (Paris, France) [12] and stored at CDS (Strasbourg, France) in the Aladin archive [2]. We estimated the content of these images in three different ways:

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

2. By counting the number of objects estimated in the image by the Sextractor object detection package [1]. As in the case of USNO these detections are mainly point sources (stars, as opposed to spatially extended objects like galaxies).

3. By counting the number of structures detected at several scales using the MR/1 multiresolution analysis package [22].

Figs. 10 show 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 DENIS near infrared surveys. Results obtained were very similar to what was obtained above. This seems to point to multiscale entropy as being a universal measurement of image content.

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 [29]. 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. [18] and Couvidat [8] have estimated this optimal compression rate using the compression program of the MR/1 package [22].

Fig. 11 shows the relation obtained between the multiscale entropy and the optimal compression rate for all the images used in our previous tests including CCD ones. 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.

6 Conclusion

We have seen that information must be measured from the transformed data, and not from the data itself. We could have used the Shannon entropy (perhaps generalized, cf. [26]) to measure the information at a given scale, and derive the bins of the histogram from the standard deviation of the noise, but for several reasons we thought it better to directly introduce noise probability into our information measure. Firstly, we have seen that this leads, for Gaussian noise, to a very physically meaningful relation between the information and the wavelet coefficients: information is proportional to the energy of the wavelet coefficients normalized by the standard deviation of the noise. Secondly, it can be generalized to many other kinds of noise, even in the case of images with few photons/events (the histograms in this case give rise to a bias). We have seen that the equations are easy to manipulate. Finally, experiments have confirmed that this approach gives good results.

For filtering, the multiscale entropy has the following advantages:

- It provides a good trade-off between hard and soft thresholding.
- No a priori model on the signal itself is needed as with other wavelet based Bayesian methods [5, 9, 32, 31].
- It can be generalized to many kinds of noise.
- The regularization parameter α can be easily fixed automatically. Cross-validation [24] could be an alternative, but with the limitation to Gaussian noise.

Replacing the standard entropy measurements by the Multiscale Entropy avoids the main problems in the MEM deconvolution method. We have seen also that our new information measure allows us to analyze the background fluctuation and to characterize the content of an astronomical image.

References

[1] E. Bertin and S. Arnouts. Sextractor: Software for source extraction. *Astronomy and Astrophysics, Suppl. Ser.*, 117:393–404, June 1996.

[2] 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. *Astronomical Data Analysis Software and Systems VIII, ASP Conference Series, Vol. 172. Ed. David M. Mehringer, Raymond L. Plante, and Douglas A. Roberts. ISBN: 1-886733-94-5 (1999), p. 229.*, 8:229–232, 1999.

[3] J.P. Burg. Annual Meeting International Society Exploratory Geophysics, Reprinted in Modern Spectral Analysis, D.G. Childers, ed., IEEE Press, New York, 34–41, 1978.

[4] A. Chambolle, R.A. DeVore, N. Lee, and B.J. Lucier. Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Transactions on Signal Processing*, 7(3):319–335, 1998.

[5] H.A. Chipman, E.D. Kolaczyk, and R.E. McCulloch. Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association*, 92(440):1413–1421, 1997.

[6] A. Cohen, I. Daubechies, and J.C. Feauveau. Biorthogonal bases of compactly supported wavelets. *Communications in Pure and Applied Mathematics*, 45:485–560, 1992.

[7] T.J. Cornwell. Image restoration. In D.M. Alloin and J.M. Mariotti, editors, *Diffraction-Limited Imaging with Very Large Telescopes*, Dordrecht, 1989. Kluwer.

[8] S. Couvidat. DEA Dissertation, Stasbourg Observatory, 1999.

[9] M. Crouse, R. Nowak, and R. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing*, 46:886–902, 1998.

[10] J.C. Feauveau. *Analyse multirésolution par ondelettes non-orthogonales et bancs de filtres numériques.* PhD thesis, Université Paris Sud, 1990.

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

[12] J. Guibert. The MAMA facility - a survey of scientific programmes. In *Digitised Optical Sky Surveys*, pages 103+, 1992.

[13] S.F. Gull and J. Skilling. *MEMSYS5 Quantified Maximum Entropy User's Manual*, 1991.

[14] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian. *Wavelets: Time-Frequency Methods and Phase-Space*, chapter A Real-Time Algorithm for Signal Analysis with the Help of the Wavelet Transform, pages 286–297. Springer-Verlag, 1989.

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

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

[17] H. Lorenz and G.M. Richter. Adaptive filtering of HST images: preprocessing for deconvolution. In P. Benvenuti and E. Schreier, editors, *Science with the Hubble Space Telescope*, pages 203–206, Garching, 1993. European Southern Observatory.

[18] M. Louys, J.-L. Starck, S. Mei, F. Bonnarel, and F. Murtagh. Astronomical image compression. *Astronomy and Astrophysics, Suppl. Ser.*, 136:579–590, 1999.

[19] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, 1998.

[20] S. G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[21] R. Molina. On the hierarchical Bayesian approach to image restoration. Applications to astronomical images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:1122–1128, 1994.

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

[23] R. Narayan and R. Nityananda. Maximum entropy image restoration in astronomy. *Ann. Rev. Astron. Astrophys.*, 24:127–170, 1986.

[24] G.P. Nason. Wavelet shrinkage by cross-validation. *Journal of the Royal Statistical Society B*, 58:463–479, 1996.

[25] C.E. Shannon. A mathematical theory for communication. *Bell System Tech J.*, 27:379–423, 1948.

[26] J. Sporring and J. Weickert. Information measures in scale-spaces. *IEEE Transactions on Information Theory*, 45:1051–1058, 1999.

[27] J.L. Starck, A. Bijaoui, and F. Murtagh. Multiresolution support applied to image filtering and deconvolution. *CVGIP: Graphical Models and Image Processing*, 57:420–431, 1995.

[28] J.L. Starck and F. Murtagh. Multiscale entropy filtering. *Signal Processing*, 76(2):147–165, 1999.

[29] J.L. Starck, F. Murtagh, and A. Bijaoui. *Image Processing and Data Analysis: The Multiscale Approach.* Cambridge University Press, Cambridge (GB), 1998.

[30] J.L. Starck, F. Murtagh, and R. Gastaud. A new entropy measure based on the wavelet transform and noise modeling. Special Issue on Multirate Systems, Filter Banks, Wavelets, and Applications of *IEEE Transactions on CAS II*, 45(8), 1998.

[31] K. E. Timmermann and R. Nowak. Multiscale modeling and estimation of Poisson processes with applications to photon-limited imaging. *IEEE Transactions on Signal Processing*, 46:886–902, 1999.

[32] B. Vidakovic. Non-linear wavelet shrinkage with Bayes rules and Bayes factors. *Journal of the American Statistical Association*, 93:173–179, 1998.

Appendix: Derivative needed for the minimization of equations 11 and 18.

Gaussian case

We compute the contribution of the wavelet coefficient x to the noise information:

$$h_n(x) = \frac{1}{\sigma^2}\int_0^x t \operatorname{erfc}\left(\frac{x-t}{\sqrt{2}\sigma}\right) dt \tag{24}$$

$$\begin{aligned}
\frac{dh_n(x)}{dx} &= h_n(x+dx) - h_n(x) \\
&= \frac{1}{\sigma^2}\int_0^{x+dx} t \operatorname{erfc}\left(\frac{x+dx-t}{\sqrt{2}\sigma}\right) dt - \frac{1}{\sigma^2}\int_0^x t \operatorname{erfc}\left(\frac{x-t}{\sqrt{2}\sigma}\right) dt \\
&= \frac{1}{\sigma^2}\int_0^x \left[t \operatorname{erfc}\left(\frac{x+dx-t}{\sqrt{2}\sigma}\right) - t \operatorname{erfc}\left(\frac{x-t}{\sqrt{2}\sigma}\right)\right] dt + \frac{1}{\sigma^2}\int_x^{x+dx} t \operatorname{erfc}\left(\frac{x+dx-t}{\sqrt{2}\sigma}\right) dt \\
&= \frac{1}{\sigma^2}\int_0^x t \frac{\partial \operatorname{erfc}\left(\frac{x-t}{\sqrt{2}\sigma}\right)}{\partial x} + \frac{1}{\sigma^2}\int_x^{x+dx} t \operatorname{erfc}\left(\frac{x+dx-t}{\sqrt{2}\sigma}\right) dt \\
&= \frac{1}{\sigma^2}\int_0^x t \frac{\partial \operatorname{erfc}(\frac{x-t}{\sqrt{2}\sigma})}{\partial x} dt + \frac{x}{\sigma^2} \operatorname{erfc}(0)
\end{aligned} \tag{25}$$

Now, because $\operatorname{erfc}(0) = 1$ we have:

$$\frac{dh_n(x)}{dx} = \frac{x}{\sigma^2} + \frac{1}{\sigma^2}\int_0^x t \frac{\partial \operatorname{erfc}(\frac{x-t}{\sqrt{2}\sigma})}{\partial x} dt \tag{26}$$

We derive the function erfc:

$$\frac{\partial \operatorname{erfc}(x)}{\partial x} = -\frac{2}{\sqrt{\pi}} e^{-x^2} \tag{27}$$

$$\frac{\partial \operatorname{erfc}(\frac{(x-t)}{\sqrt{2}\sigma})}{\partial x} = -\frac{2}{\sqrt{\pi}}\frac{1}{\sqrt{2}\sigma} e^{-\frac{(x-t)^2}{2\sigma^2}} = -\sqrt{\frac{2}{\pi}}\frac{1}{\sigma} e^{-\frac{(x-t)^2}{2\sigma^2}} \tag{28}$$

Now we deduce for the derivative of h_n:

$$\frac{dh_n(x)}{dx} = \frac{x}{\sigma^2} + \frac{1}{\sigma^2}\int_0^x -\sqrt{\frac{2}{\pi}}\frac{1}{\sigma} t\, e^{-\frac{(x-t)^2}{2\sigma^2}} dt \tag{29}$$

$$\frac{dh_n(x)}{dx} = \frac{x}{\sigma^2} + \frac{1}{\sigma^3}\sqrt{\frac{2}{\pi}}\int_0^x t\, e^{-\frac{(x-t)^2}{2\sigma^2}} dt \tag{30}$$

We create the variable J

$$J = \int_0^x t\, e^{-\frac{(x-t)^2}{2\sigma^2}} dt \tag{31}$$

We create the variable u

$$\begin{aligned} u = \tfrac{t-x}{\sqrt{2}\sigma} \qquad & t = x + u\sqrt{2}\sigma \\ & dt = \sqrt{2}\sigma du \\ t = 0 \Rightarrow u = \tfrac{-x}{\sqrt{2}\sigma} \quad & t = x \Rightarrow u = 0 \end{aligned} \tag{32}$$

The variable J can be written with u

$$J = \int_{\frac{-x}{\sqrt{2}\sigma}}^0 (x + u\sqrt{2}\sigma)e^{-u^2}\sqrt{2}\sigma du \tag{33}$$

$$J = \sqrt{2}\sigma x \int_{\frac{-x}{\sqrt{2}\sigma}}^0 e^{-u^2} du + 2\sigma^2 \int_{\frac{-x}{\sqrt{2}\sigma}}^0 u\, e^{-u^2} du \tag{34}$$

The first part of J can be rewritten as:

$$J_0 = \sqrt{2}\sigma x \int_0^{\frac{x}{\sqrt{2}\sigma}} e^{-u^2} du \tag{35}$$

J_0 can be expressed with the error function.

$$J_0 = \sqrt{2}\sigma\frac{\sqrt{\pi}}{2} x \mathrm{erf}\left(\frac{x}{\sqrt{2}\sigma}\right) = \sigma\sqrt{\frac{\pi}{2}}\, x\, \mathrm{erf}\left(\frac{x}{\sqrt{2}\sigma}\right) \tag{36}$$

Now the second part of J is obvious

$$J_1 = 2\sigma^2 \int_{\frac{-x}{\sqrt{2}\sigma}}^0 u\, e^{-u^2} du \tag{37}$$

or

$$\frac{de^{-u^2}}{du} = -2ue^{-u^2} \tag{38}$$

We replace

$$J_1 = -\sigma^2 \int_{\frac{-x}{\sqrt{2}\sigma}}^0 d(e^{-u^2}) \tag{39}$$

$$J_1 = \sigma^2[e^{-\frac{x^2}{2\sigma^2}} - 1] \tag{40}$$

Now we can write J

$$J = J_0 + J_1 = \sigma\sqrt{\frac{\pi}{2}}\, x \operatorname{erf}\left(\frac{x}{\sqrt{2}\sigma}\right) + \sigma^2\left[e^{-\frac{x^2}{2\sigma^2}} - 1\right] \tag{41}$$

We can write the derivative of h_n

$$\begin{aligned}
\frac{dh_n(x)}{dx} &= \frac{x}{\sigma^2} - \frac{1}{\sigma^3}\sqrt{\frac{2}{\pi}}\, J \\
&= \frac{x}{\sigma^2} - \frac{x}{\sigma^2}\, x \operatorname{erf}\left(\frac{x}{\sqrt{2}\sigma}\right) + \frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\left[1 - e^{-\frac{x^2}{2\sigma^2}}\right] \\
&= \frac{x}{\sigma^2}\operatorname{erfc}\left(\frac{x}{\sqrt{2}\sigma}\right) + \frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\left[1 - e^{-\frac{x^2}{2\sigma^2}}\right]
\end{aligned} \tag{42}$$

In order to minimize the functional (11), we may want to calculate the derivative of $h_s(y-x)$, where $h_s(y-x)$ measures the amount of information contained in the residual (y being the data).

$$h_s(y-x) = \frac{1}{\sigma^2}\int_0^{y-x} t \operatorname{erf}\left(\frac{y-x-t}{\sqrt{2}\sigma}\right) dt \tag{43}$$

Denoting $z = y - x$, we have

$$\begin{aligned}
h_s(z) &= \frac{1}{\sigma^2}\int_0^{z} t \operatorname{erf}\left(\frac{z-t}{\sqrt{2}\sigma}\right) dt \\
&= \frac{1}{\sigma^2}\int_0^{z} t dt - \frac{1}{\sigma^2}\int_0^{z} t \operatorname{erfc}\left(\frac{z-t}{\sqrt{2}\sigma}\right) dt
\end{aligned} \tag{44}$$

and

$$\frac{dh_s(x)}{dx} = \frac{dh_s(z)}{dz}\frac{dz}{dx} \tag{45}$$

$$\frac{dh_s(z)}{dz} = \frac{z}{\sigma^2} - \frac{dh_n(z)}{dz} \tag{46}$$

then

$$\begin{aligned}
\frac{dh_s(y-x)}{dx} &= -\frac{y-x}{\sigma^2} + \frac{y-x}{\sigma^2}\operatorname{erfc}\left(\frac{y-x}{\sqrt{2}\sigma}\right) + \sqrt{\frac{2}{\pi}}\frac{1}{\sigma}\left[1 - e^{-\frac{(y-x)^2}{2\sigma^2}}\right] \\
&= -\frac{y-x}{\sigma^2}\operatorname{erf}\left(\frac{y-x}{\sqrt{2}\sigma}\right) + \sqrt{\frac{2}{\pi}}\frac{1}{\sigma}\left[1 - e^{-\frac{(y-x)^2}{2\sigma^2}}\right]
\end{aligned} \tag{47}$$

General Case

The contribution of the wavelet coefficient x to the noise and signal information in the general case is

$$\begin{aligned} h_n(x) &= \int_0^{|x|} p_n(x-u)\left(\frac{\partial h(x)}{\partial x}\right)_{x=u} du \\ h_s(x) &= \int_0^{|x|} p_s(x-u)\left(\frac{\partial h(x)}{\partial x}\right)_{x=u} du \end{aligned} \tag{48}$$

Assuming $h(x) = \frac{1}{2}x^2$, we have

$$\begin{aligned} h_n(x) &= \int_0^{|x|} p_n(x-u)udu \\ h_s(x) &= \int_0^{|x|} p_s(x-u)udu \end{aligned} \tag{49}$$

$$\frac{dh_s(x)}{dx} = \int_0^x \left(\frac{\partial p_s(x-u)}{\partial x}\right)_{x=u} udu + \frac{1}{dx}\int_x^{x+dx} p_s(x-u)udu \tag{50}$$

Since $p_s(0) = 0$, the second term tends to zero.

Denoting $\frac{\partial p_s(x-u)}{\partial x} = -\frac{\partial p_s(x-u)}{\partial u}$, we have

$$\begin{aligned} \frac{dh_s(x)}{dx} &= -\int_0^x \frac{\partial p_s(x-u)}{\partial u} udu \\ &= -([up_s(x-u)]_0^x - \int_0^x p_s(x-u)du) \\ &= \int_0^x p_s(x-u)du \\ &= \int_0^x p_s(u)du \end{aligned} \tag{51}$$

and from $h_n = h - h_s$ we get

$$\frac{dh_n(x)}{dx} = x - \int_0^x p_s(u)du \tag{52}$$

and

$$\frac{dh_s(y-x)}{dx} = -\int_0^{y-x} p_s(u)du \tag{53}$$

It is easy to verify that replacing $p_s(x) = \mathrm{erf}(x)$, and $p_n(x) = \mathrm{erfc}(x)$ (case of Gaussian noise) we find the same equation as in the Gaussian case.

Figure 1: Lena image (left) and the same data distributed differently (right). These two images have the same entropy, using any of the standard entropy definitions.

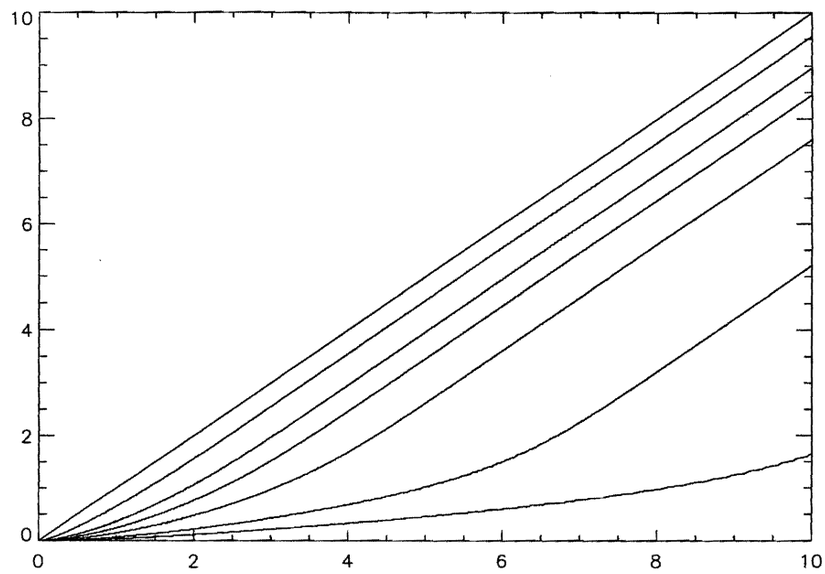

Figure 2: Corrected wavelet coefficient versus the wavelet coefficient with different α values (from the top curve to the bottom one, α is respectively equal to 0,0.1,0.5, 1, 2, 5,10).

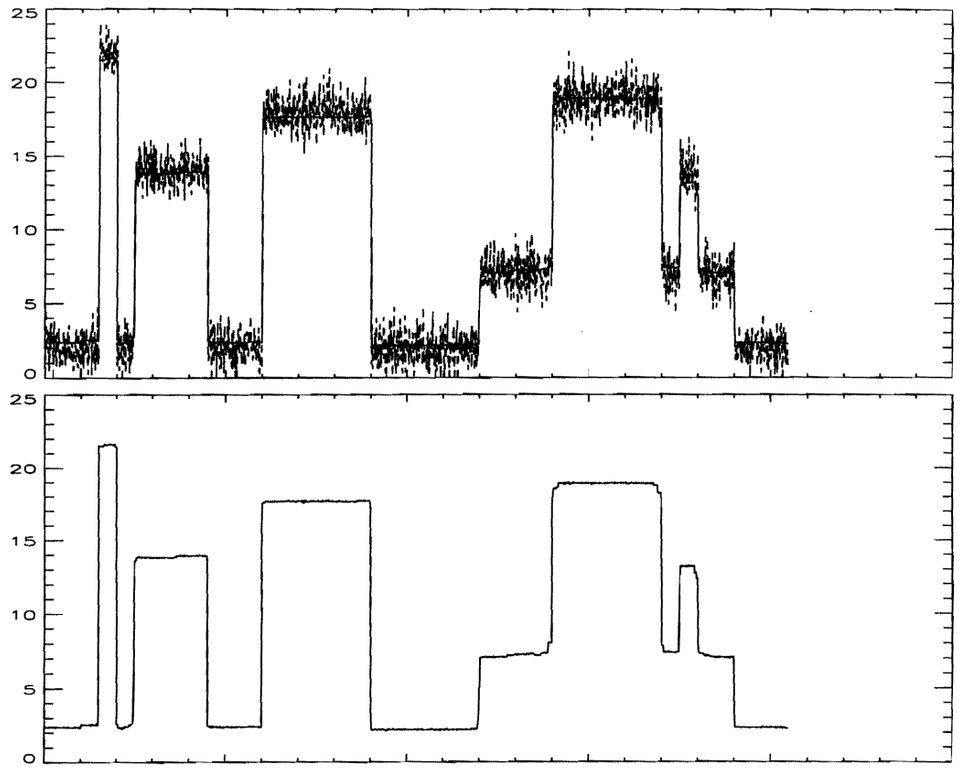

Figure 3: Top, noisy blocks and filtered blocks overplotted. Bottom, filtered blocks.

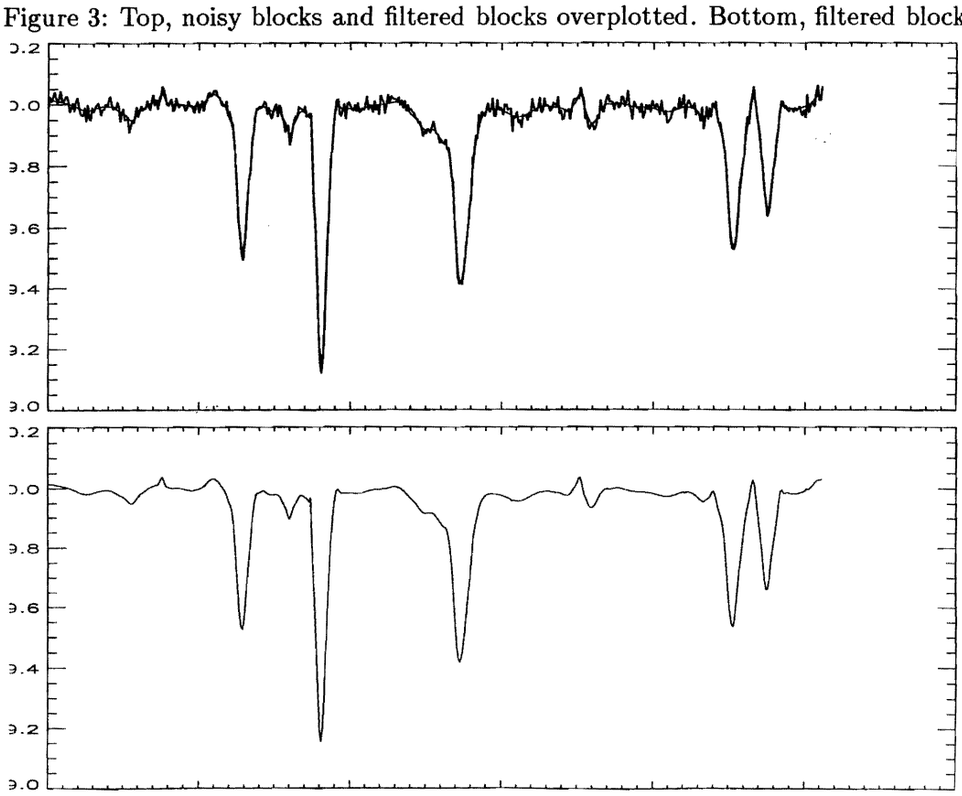

Figure 4: Top, real spectrum and filtered spectrum overplotted. Bottom, filtered spectrum.

Figure 5: Top left, simulated noisy image, top right, hard thresholding, bottom left, soft thresholding, and bottom right, MEF.

Figure 6: Upper left, Lena image convolved with a Gaussian (FWHM = 5) plus Gaussian noise (σ = 5). Upper right, deconvolved image by the Multiscale Entropy Deconvolution Method. Bottom left and right, zoom on a part of the previous images.

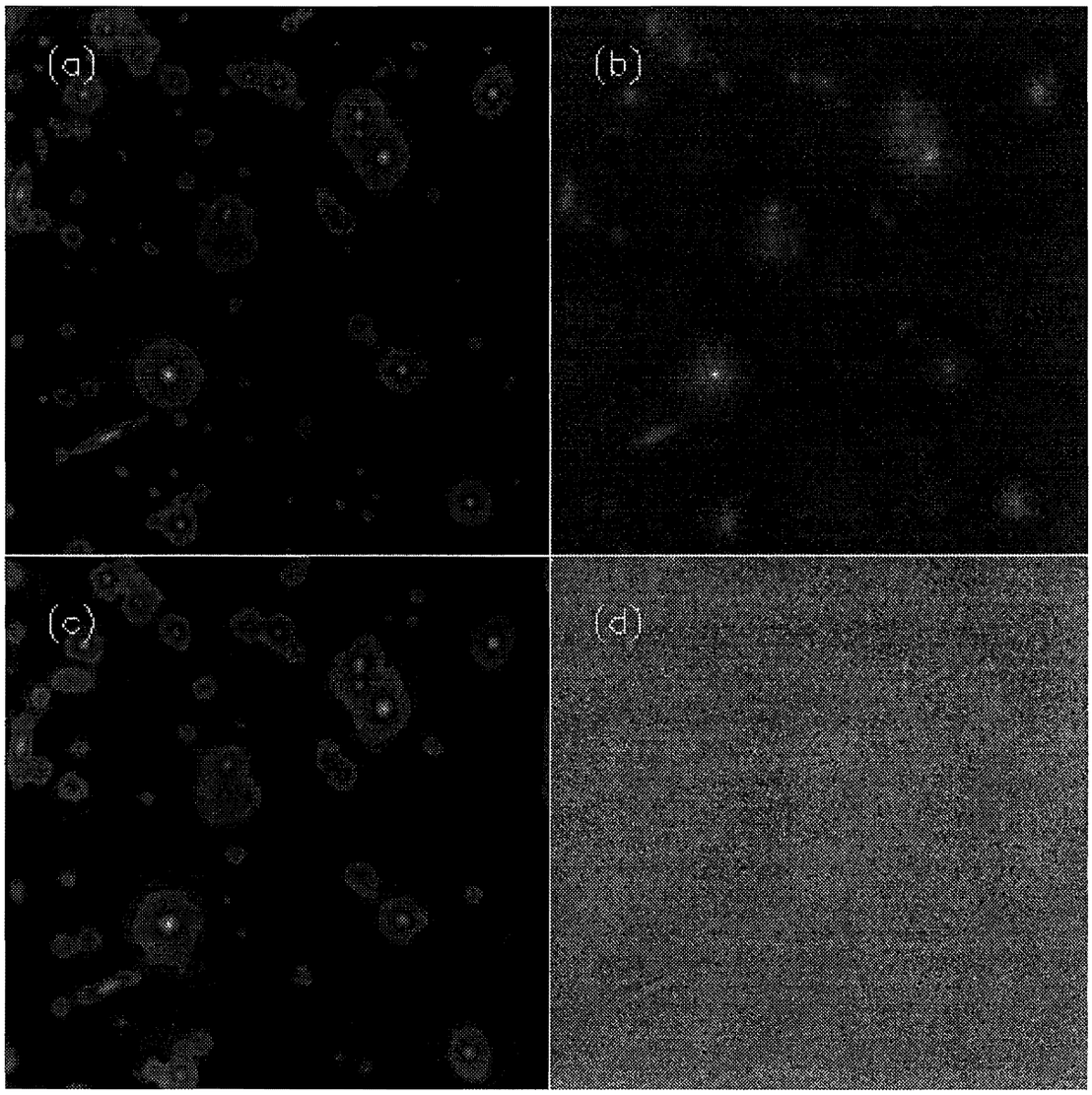

Figure 7: (a) Original image, (b) blurred image + Gaussian noise, (c) deconvolved image, and (d) residual image.

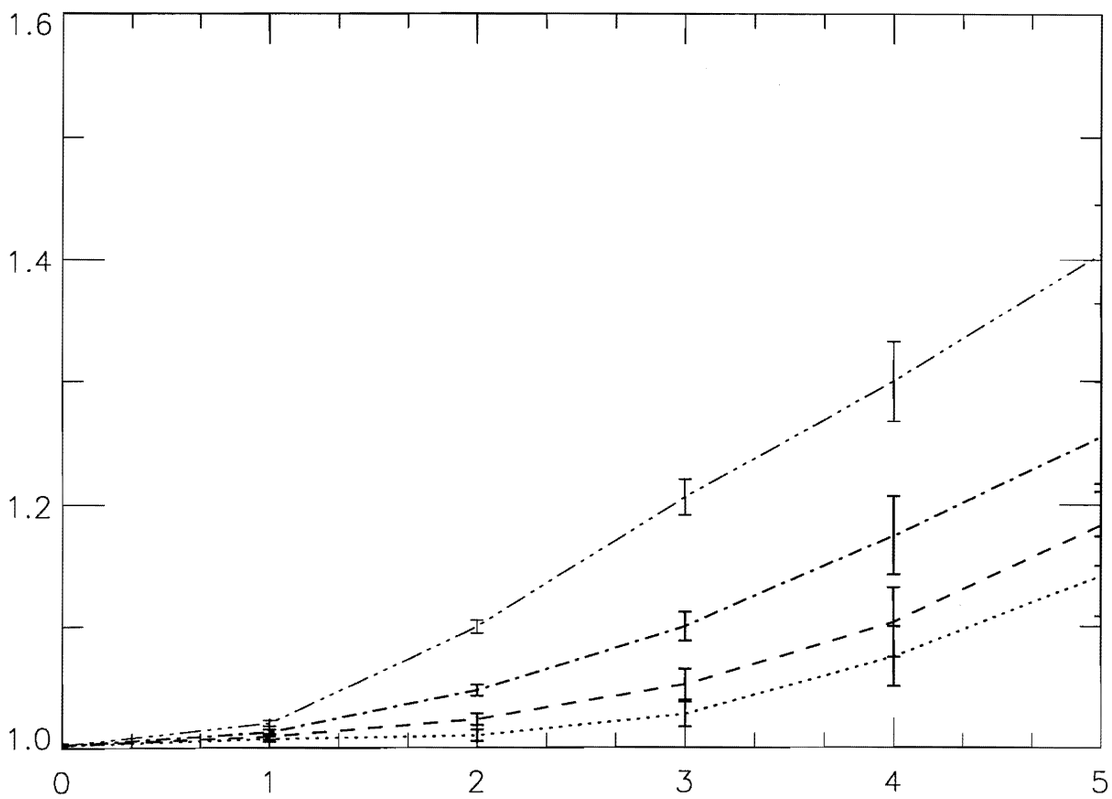

Figure 8: Mean entropy versus the scale of 5 simulated images containing undetectable sources and noise. Each curve corresponds to the multiscale transform of one image. From top to bottom, the image contains respectively 400, 200, 100, 50 and 0 sources.

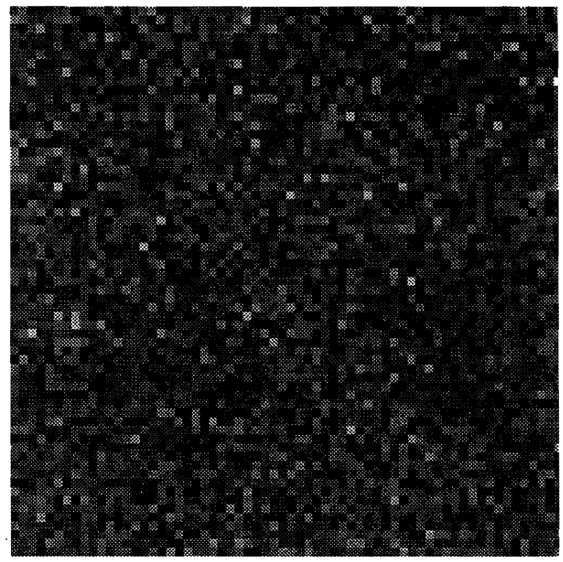

Figure 9: Region of a simulated image containing an undetectable source at the center.

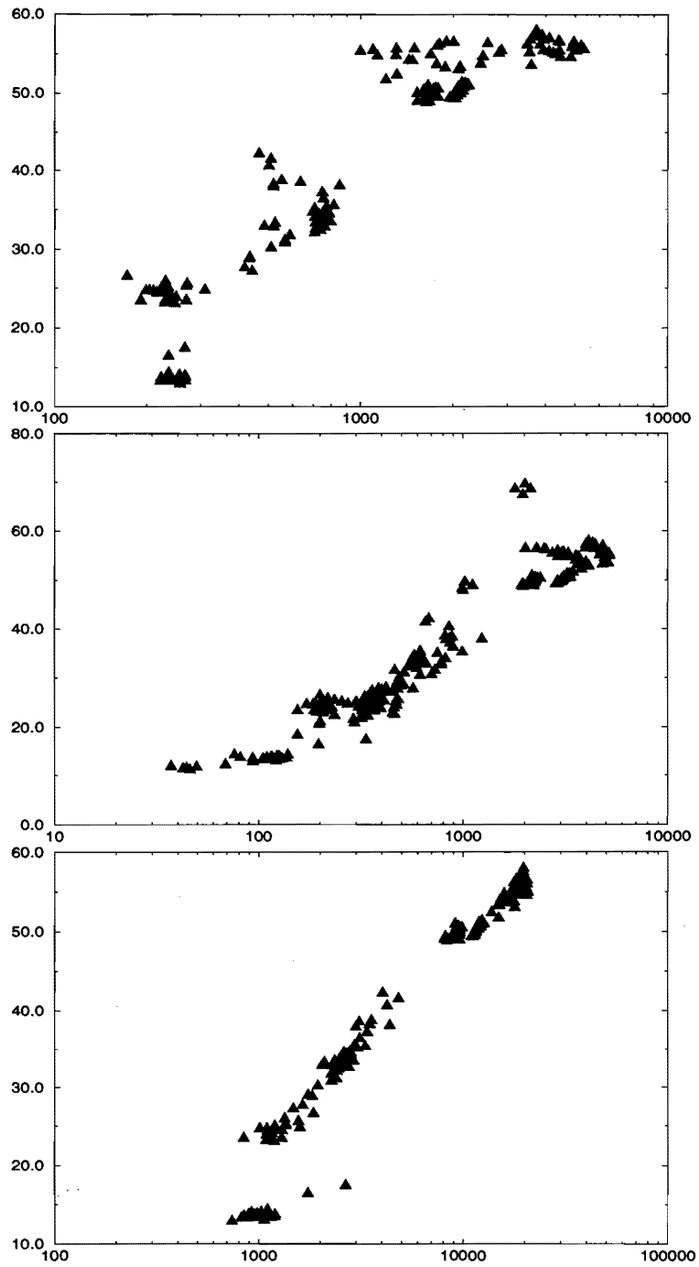

Figure 10: Multiscale entropy versus the number of objects: the number of objects is, respectively, obtained from (top) the USNO catalog, (middle) the Sextractor package, and (bottom) the MR/1 package.

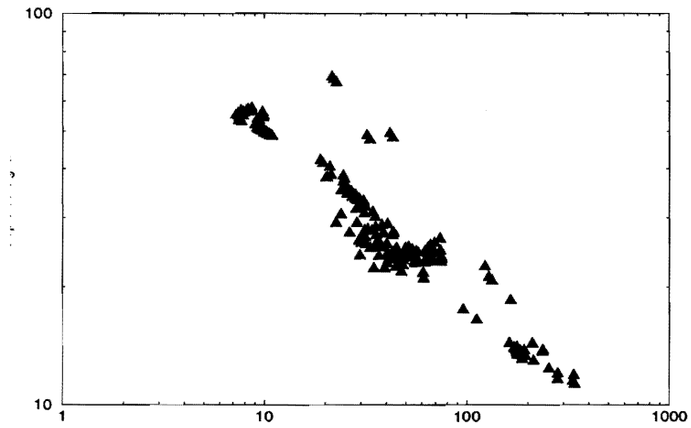

Figure 11: Multiscale entropy of astronomical images versus the optimal compression ratio. Images which contain a high number of sources have a small ratio and a high multiscale entropy value. The relation is almost linear.