(a)

(b)

Fig. 6. Reconstructed Lena images ($\mathrm{PSNR} = 30$ dB). (a) Using BWTH at 52.15 : 1 compression. (b) Using JPEG at 32.47 : 1 compression.

superposition operations. We add a new member, K1-transform, to the family of reversible embedded wavelet transforms that may be used in lossless compression. It has a higher degree of regularity than the two existing reversible embedded wavelet transforms. The performance measure of our lossless compression using K1-transform was shown to have a 10% improvement over the lossless JPEG. For lossy compression, we present a fast reconstruction algorithm based on multiplierless 2-D filter masks that take advantage of the characteristics of the wavelet transformed data; the Hilbert scanning is applied to gain an additional compression. In comparison to JPEG, this BWTH compression demonstrated a 60% improvement.

References

[1] J. M. Shapiro, "Embedded image coding using zero-trees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.

[2] M. Antonini, M. Barlaud, P. Marthieu, and I. Daubechies, "Image coding using wavelet transforms," *IEEE Trans. Image Processing*, vol. 1, pp. 205–220, Apr. 1992.

[3] G. Strang and T. Nguyen, *Wavelets and Filter Banks.* Cambridge, U.K.: Wellesley-Cambridge Press, 1996.

[4] A. Zandi, J. Allen, E. Schwartz, and M. Boliek, "CREW: Compression with reversible embedded wavelets," in *Proc. 5th IEEE Data Compression Conf.*, Snowbird, UT, 1995, pp. 212–221.

[5] D. LeGall and A. Tabatabai, "Sub-band coding of digital images using symmetric short Kernel filters and arithmetic coding techniques," in *IEEE Proc. Int. Conf. ASSP*, 1988, pp. 761–764.

[6] I. Daubechies, *Ten Lectures on Wavelets*, vol. CBMS-61. Philadelphia, PA: SIAM, 1992.

[7] H. Kim and C. C. Li, "A fast reversible wavelet image compressor," in *Proc. SPIE, Math. Imaging: Wavelet Appl. Signal Image Processing IV*, vol. 2825, 1996, pp. 929–940.

[8] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[9] H. Kim, "Unified image compression using reversible and fast biorthogonal wavelet transforms," Ph.D. dissertation, Dep. Elect. Eng., Univ. Pittsburgh, Pittsburgh, PA, 1996.

A New Entropy Measure Based on the Wavelet Transform and Noise Modeling

J.-L. Starck, F. Murtagh, and R. Gastaud

***Abstract*—We present in this brief a new way to measure the information in a signal, based on noise modeling. We show that the use of such an entropy-related measure leads to good results for signal restoration.**

I. Introduction

The term "entropy" is due to Clausius (1865), and the concept of entropy was introduced by Boltzmann into statistical mechanics, in order to measure the number of microscopic ways that a given macroscopic state can be realized. Shannon [11] founded the mathematical theory of communication when he suggested that the information gained in a measurement depends on the number of possible outcomes out of which one is realized. Shannon also suggested that the entropy can be used for maximization of the bits transferred under a quality constraint. Jaynes [7] proposed to use the entropy measure for radio interometric image deconvolution, in order to select in a set of possible solutions which contains the minimum of information, or following his entropy definition, that which has a maximum entropy. In principle, the solution verifying such a condition should be the most reliable. A lot of work has been carried out in the last 30 years on the use of entropy for the general problem of data filtering and deconvolution [1], [3]–[5], [8]–[10], [12], [16].

Traditionally, information and entropy are determined from events and the probability of their occurrence. Signal and noise are basic building blocks of signal and data analysis in the physical sciences. Instead of the probability of an event, in this work we are led to consider the probabilities of our data being either signal or noise.

Observed data Y in the physical sciences are generally corrupted by noise, which is often additive and which follows in many cases a Gaussian distribution, a Poisson distribution, or a combination of both. Using Bayes' theorem to evaluate the probability of the

Manuscript received November 21, 1996; revised December 15, 1997.

J.-L. Starck and R. Gastaud are with Centre D'Etudes de Saclay, Service d'Astrophysique, F-91191 Gif sur Yvette Cedex, France.

F. Murtagh is with the Faculty of Informatics, University of Ulster, Magee College, Londonderry BT48 7JL, Northern Ireland.

Publisher Item Identifier S 1057-7130(98)04689-8.

realization of the original signal X, knowing the data Y, we have

$$\mathrm{Prob}(X \mid Y) = \frac{\mathrm{Prob}(Y \mid X) \cdot \mathrm{Prob}(X)}{\mathrm{Prob}(Y)}. \tag{1}$$

$\mathrm{Prob}(Y \mid X)$ is the conditional probability of getting the data Y given an original signal X, i.e., it represents the distribution of the noise. It is given, in the case of uncorrelated Gaussian noise with variance σ^2, by

$$\mathrm{Prob}(Y \mid X) = \exp\left(-\sum_{\mathrm{pixels}} \frac{(Y-X)^2}{2\sigma^2}\right). \tag{2}$$

The denominator in (1) is independent of X and is considered as a constant (stationary noise). $\mathrm{Prob}(X)$ is the *a priori* distribution of the solution X. In the absence of any information on the solution X except its positivity, a possible course of action is to derive the probability of X from its entropy, which is defined from information theory.

The main idea of information theory [11] is to establish a relation between the received information and the probability of the observed event [2]. If we note $\mathcal{I}(E)$ the information related to the event E, and p the probability of this event happening, then we consider that

$$\mathcal{I}(E) = f(p). \tag{3}$$

Then we assume the two following principles.

- The information is a decreasing function of the probability. This implies that the more information we have, the less will be the probability associated with one event.
- Additivity of the information. If we have two independent events E_1 and E_2, the information $\mathcal{I}(E)$ associated with the happening of both is equal to the addition of the information of each of them.

$$\mathcal{I}(E) = \mathcal{I}(E_1) + \mathcal{I}(E_2). \tag{4}$$

Since E_1 (of probability p_1) and E_2 (of probability p_2) are independent, then the probability of both happening is equal to the product of p_1 and p_2. Hence,

$$f(p_1 p_2) = f(p_1) + f(p_2). \tag{5}$$

Then we can say that the information measure is

$$\mathcal{I}(E) = k \ln(p) \tag{6}$$

where k is a constant. Information must be positive, and k is generally fixed at -1.

Another interesting measure is the mean information, which is denoted

$$H = -\sum_i p_i \ln(p_i). \tag{7}$$

This quantity is called the entropy of the system, and was established by Shannon in 1948 [11].

This measure has the following several properties.

- It is maximal when all events have the same probability $p_i = 1/N_e$ (N_e being the number of events), and is equal to $\ln(N_e)$. It is in this configuration that the system is the most undefined.
- It is minimal when one event is sure. In this case, the system is perfectly known, and no information can be added.
- The entropy is a positive, continuous, and symmetric function.

Then if we know the entropy H of the solution (the next section describes different ways to calculate it), we derive its probability by

$$\mathrm{Prob}(X) = \exp(-\alpha H(X)). \tag{8}$$

Given the data, the most probable image is obtained by maximizing $\mathrm{Prob}(X \mid Y)$. Taking the logarithm of (1), we thus need to maximize

$$\begin{aligned}&\ln(\mathrm{Prob}(X \mid Y))\\ &\quad = -\alpha H(X) + \ln(\mathrm{Prob}(Y \mid X)) - \ln(\mathrm{Prob}(Y)).\end{aligned} \tag{9}$$

The last term is a constant and can be omitted. Then, in the case of Gaussian noise, the solution is found by minimizing

$$J(X) = \sum_{\mathrm{pixels}} \frac{(Y-X)^2}{2\sigma^2} + \alpha H(X) = \frac{\chi^2}{2} + \alpha H(X) \tag{10}$$

which is a linear combination of two terms: the entropy of the signal, and a quantity corresponding to χ^2 in statistics measuring the discrepancy between the data and the predictions of the model. α is a parameter that can be viewed alternatively as a Lagrangian parameter or a value fixing the relative weight between the goodness-of-fit and the entropy H.

For the deconvolution problem, the object–data relation is given by the convolution

$$Y = P * X \tag{11}$$

where P is the point spread function, and the solution is found (in the case of Gaussian noise) by minimizing

$$J(X) = \sum_{\mathrm{pixels}} \frac{(Y - P * X)^2}{2\sigma^2} + \alpha H(X). \tag{12}$$

The way the entropy is defined is fundamental, because the solution will depend on its definition. The next section discusses the different approaches which have been proposed in the past.

II. The Concept of Entropy

We wish to estimate an unknown probability density $p(x)$ of the data. A direct approach would be to build up the histogram of values $X(i)$, using a suitable interval Δx, counting up how many times m_k each interval $(x_k, x_k + \Delta x)$ occurs among the N occurrences. Then the probability that a data value belongs to an interval k is $p_k = m_k/N$, and each data value has a probability p_k. The entropy is defined by

$$H_s(X) = -\sum_{k=1}^{m} p_k \ln(p_k) \tag{13}$$

where m is the number of intervals. The entropy is minimum and equal to zero when the signal is flat, and increases when we have some fluctuations. Using this entropy in (10) for restoration leads to a minimum entropy restoration method.

The trouble with this approach is that, because the number of occurrences is finite, the estimate p_k will be in error by an amount proportional to $m_k^{-(1/2)}$ [6]. The error becomes significant when m_k is small. Furthermore, this kind of entropy definition is not easy to use for signal restoration, because the gradient of (10) is not easy to compute. For these reasons, other entropy functions are generally used. The main ones are

- Burg [4]:
$$H_b(X) = -\sum_{\text{pixels}} \ln(X) \tag{14}$$
- Frieden [5]:
$$H_f(X) = -\sum_{\text{pixels}} X \ln(X) \tag{15}$$
- Gull and Skilling [8]:
$$H_g(X) = \sum_{\text{pixels}} X - M - X \ln(X \mid M). \tag{16}$$

Each of these entropies can be used, and they correspond to different probability distributions that one can associate with an image [9] (see [5], [12], [13] for descriptions). The last definition of the entropy has the advantage of having a zero maximum when X equals the model M, usually taken as a flat image. All of these entropy measures are negative, and maximum when the image is flat. They are negative because an offset term is omitted which has no importance for the minimization of the functional. The fact that we consider that a signal has maximum information value when it is flat is evidently a curious way to measure information. The probability of X must be defined by $\mathrm{Prob}(X) = \exp(\alpha H(X))$. The sign has been inverted [see (8)], which is natural if we want the best solution to be the smoothest. These three entropies lead to the maximum entropy restoration method, for which the solution is found by minimizing (for Gaussian noise)

$$J(X) = \sum_{\text{pixels}} \frac{(Y-X)^2}{2\sigma^2} - \alpha H(X). \tag{17}$$

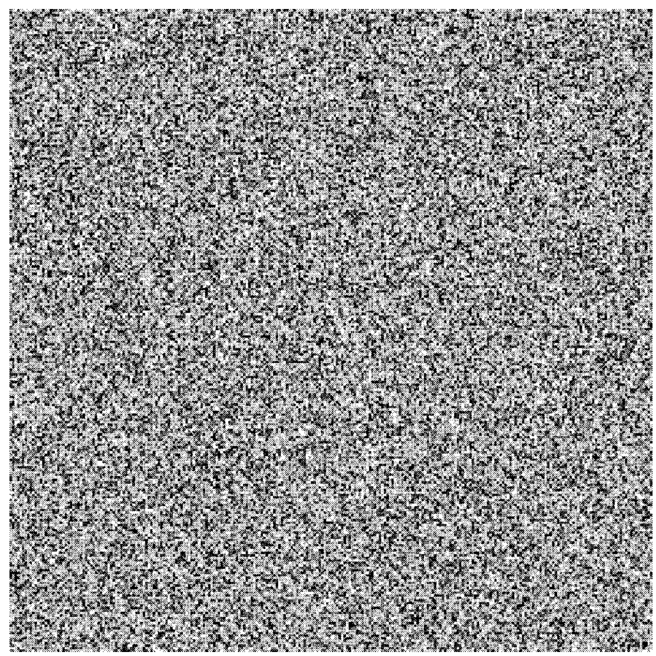
In 1986, Narayan and Nityanda [9] compared several entropy functions, and finally concluded by saying that all were comparable if they have good properties, i.e., they enforce positivity, and they have a negative second derivative which discourages ripple. They showed also that results varied strongly with the background level, and that these entropy functions produced poor results for negative structures, i.e., structures under the background level (absorption area in an image, absorption band in a spectrum, etc.), and compact structures in the signal. The Gull and Skilling entropy gives rise to the difficulty of estimating a model. Furthermore, it has been shown [3] that the solution was dependent on this choice.

Many studies [3], [10], [16] have been carried out in order to improve the functional to be minimized. But the question which should be raised is: what is a good entropy for signal restoration?

Trying to answer this corresponds to asking what is the information in the signal. The entropy should verify the following criteria.

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 ($Y = X + \textit{Noise}$) 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.

(a)



(b)

Fig. 1. (a) Lena image and (b) the same data distributed differently. These two images have the same entropy, using any of the standard entropy methods.

Fig. 1(a) and (b) shows, respectively, the Lena image and an image obtained by distributing randomly the Lena image pixel values. 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 a lossless transmission, and from this point of view, he will consider that the second one contains more information. The standard entropy methods produce exactly the same value for both images and, for such methods, both images contain the same amount of information. For data restoration, all fluctuations due to noise are not of interest, and do not contain relevant information. From this physical point of view, that is the reason why the standard definition of entropy seems badly adapted to information measurement in signal restoration.

III. Entropy from Noise Modeling

In the case of signal restoration, the noise is the main problem. This means that we should not consider the probability of appearance of a pixel value in an image, but rather its probability of being due to the signal (or to the noise). If we consider a variable x which follows a probability distribution $p(x)$, we can define the information in x by $-\ln(p(x))$, and a signal S can be considered as a set of individual variables x_k (pixels), each of which follows the same probability distribution. Then the information contained in the data can be measured by $-\sum_{pixel} \ln(p(x))$. If x follows a Gaussian distribution with zero mean, we have

$$H(X) = \sum_{\text{pixels}} \frac{x^2}{2\sigma^2}. \tag{18}$$

The energy gives a good measurement of information. But many of the required criteria are not fulfilled by using such an entropy (correlation between pixels, background-independent, etc.). It seems difficult to derive a good probability distribution from the pixel values which fulfill the entropy requirements.

This is not so for transformed data, especially when using the wavelet transform. This has already been done, in fact, for finding threshold levels in filtering methods by means of wavelet coefficient thresholding [14]. Thus we must introduce the concept of multiresolution into our entropy. We will now consider that the information contained in some dataset is the sum of the information at different resolution levels j. Choosing the "à trous" wavelet transform (see [14] for a description of this wavelet transform algorithm), a signal S can be represented by

$$S(k) = \sum_{j=1}^{l} w_j(k) + c_l(k) \tag{19}$$

where k is the pixel index, w_j are the wavelet coefficients of S, j the resolution level, and c_l the smoothed version of S. Due to the properties of the wavelet transform, the set $w_j(x)$ for all x has a zero mean. From noise modeling, we can derive the probability distribution in the wavelet space of a wavelet coefficient, assuming it is due to the noise. The entropy becomes

$$H(X) = -\sum_{j=1}^{l}\sum_{k=1}^{N} \ln(p(w_j(k))). \tag{20}$$

For Gaussian noise, we get

$$H(X) = \sum_{j=1}^{l}\sum_{k=1}^{N} \frac{w_j(k)^2}{2\sigma_j^2} \tag{21}$$

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, 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 the requirements of Section II. If we consider two signals S_1, S_2, derived from a third one S_0 by adding noise

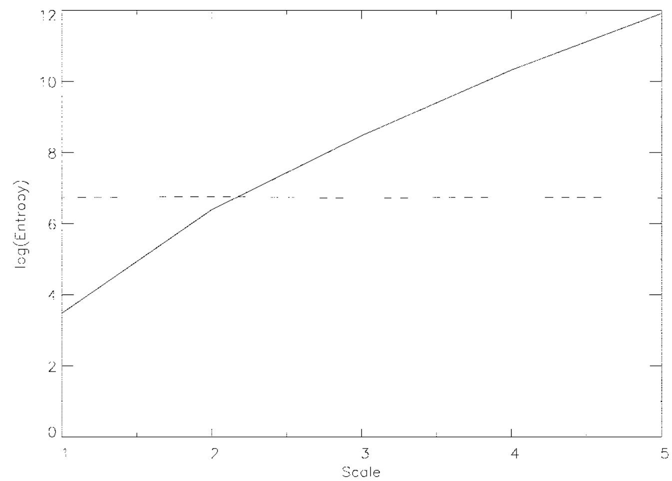


Fig. 2. Multiscale entropy of the Lena image (continuous curve), and multiscale entropy of the scrambled image (dashed curve).

$$\begin{aligned} S_1 &= S_0 + N_1(\sigma_1) \\ S_2 &= S_0 + N_2(\sigma_2) \end{aligned} \tag{22}$$

then we have

$$\text{if } \sigma_1 < \sigma_2 \qquad \text{then} \quad H(S_1) > H(S_2) \tag{23}$$

and a flat image has zero entropy.

Our entropy definition is completely dependent on the noise modeling. If we consider a signal S, and we assume that the noise is Gaussian, with a standard deviation equal to σ, we won't measure the same information compared to the case when we consider that the noise has another standard deviation value, or if the noise follows another distribution.

Fig. 2 shows the information measure at each scale for both the Lena image and its scrambled version. The global information is the addition of the information at each scale. We see that for the scrambled image (dashed curve), the information-versus-scale curve is flat, while for the unscrambled Lena image, it increases with the scale.

IV. Signal Information and Noise Information

A. Definition

In the previous section, we have seen how it was possible to measure the information related to a wavelet coefficient. Since the data are composed of an original signal and noise, our information measure is corrupted by noise. Trying to decompose our information measure into two components—one (H_S) corresponding to the noncorrupted part—and another (H_N) to the corrupted part—we have

$$H(X) = H_S(X) + H_N(X). \tag{24}$$

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.

As described in the previous section, the information h relative to a wavelet coefficient w_j is $-\ln(p(w_j))$. If the wavelet coefficient is small, its value can be due to the noise, and h should be assigned to H_N. If the wavelet coefficient is high, compared to the noise standard deviation, h 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 $p_n(w_j)$ that the wavelet coefficient is due to noise, or the probability $p_s(w_j)$ that it is due to signal. We have $p_s(w_j) = 1 - p_n(w_j)$. We consider that $h_n(w_j) = -p_n(w_j)\ln(p(w_j))$ is the noise information, and $h_s(w_j) = -p_s(w_j)\ln(p(w_j))$ is the signal information. Hence signal information and noise information are defined by

$$\begin{aligned} H_s(X) &= \sum_{j=1}^{l}\sum_{k=1}^{N} h_s(w_j(k)) \\ &= -\sum_{j=1}^{l}\sum_{k=1}^{N} p_s(w_j(k))\ln(p(w_j(k))) \\ H_n(X) &= \sum_{j=1}^{l}\sum_{k=1}^{N} h_n(w_j(k)) \\ &= -\sum_{j=1}^{l}\sum_{k=1}^{N} p_n(w_j(k))\ln(p(w_j(k))). \end{aligned} \tag{25}$$

For the Gaussian noise case, we estimate $p_n(w_j)$ that a wavelet coefficient is due to the noise by

$$\begin{aligned} p_n(w_j) &= \mathrm{Prob}(W > |w_j|) \\ &= \frac{2}{\sqrt{2\pi}\sigma_j}\int_{|w_j|}^{+\infty} \exp\left(-\frac{W^2}{2\sigma_j^2}\right) dW \\ &= \mathrm{erfc}\left(\frac{|w_j|}{\sqrt{2}\sigma_j}\right) \end{aligned} \tag{26}$$

and

$$\begin{aligned} H_s(X) &= \sum_{j=1}^{l}\sum_{k=1}^{N} \frac{w_j^2}{2\sigma_j^2}\,\mathrm{erf}\left(\frac{|w_j|}{\sqrt{2}\sigma_j}\right) \\ H_n(X) &= \sum_{j=1}^{l}\sum_{k=1}^{N} \frac{w_j^2}{2\sigma_j^2}\,\mathrm{erfc}\left(\frac{|w_j|}{\sqrt{2}\sigma_j}\right). \end{aligned} \tag{27}$$

Note that $H_s(X) + H_n(X)$ is always equal to $H(X)$. For Gaussian noise, the functional to minimize becomes

$$J(X) = \sum_{\text{pixels}} \frac{(Y - X)^2}{2\sigma^2} + \alpha(H_s(X) + H_n(X)). \tag{28}$$

If we want to preserve features with high signal-to-noise ratio from the regularization, we just omit $H_s(X)$ and we get

$$J(X) = \sum_{\text{pixels}} \frac{(Y - X)^2}{2\sigma^2} + \alpha H_n(X). \tag{29}$$

We seek a solution which minimizes the amount of information which could be due to the noise.

By this measure, information relative to high wavelet coefficients is completely assigned to the signal. This allows us also to exclude wavelet coefficients with high signal-to-noise ratio (SNR) from the regularization. It leads to perfect fit of the solution with the data at scales and positions with high SNR. If we want to consider the information due to noise, even for significant wavelet coefficients, the noise information relative to a wavelet coefficient is

$$h_n(w_j) = \int_0^{|w_j|} p_n(u \mid w_j)\left(\frac{\partial H(x)}{\partial x}\right)_{x=u} du \tag{30}$$

which gives for Gaussian noise

$$h_n(w_j) = \frac{1}{\sigma_j^2}\int_0^{|w_j|} u\,\mathrm{erfc}\left(\frac{|w_j| - u}{\sqrt{2}\sigma_j}\right) du \tag{31}$$

and the noise and signal information in a signal are

$$\begin{aligned} H_s(X) &= \sum_{j=1}^{l}\sum_{k=1}^{N} \frac{1}{\sigma_j^2}\int_0^{|w_j|} u\,\mathrm{erf}\left(\frac{|w_j| - u}{\sqrt{2}\sigma_j}\right) du \\ H_n(X) &= \sum_{j=1}^{l}\sum_{k=1}^{N} \frac{1}{\sigma_j^2}\int_0^{|w_j|} u\,\mathrm{erfc}\left(\frac{|w_j| - u}{\sqrt{2}\sigma_j}\right) du. \end{aligned} \tag{32}$$

Equations (27) and (32) lead to two different ways to regularize a signal. The first requires that we use all the information which is furnished in high wavelet coefficients, and leads to an exact preservation of the flux in a structure. If the signal presents high discontinuities, artifacts can appear in the solution due to the fact that the wavelet coefficients located at the discontinuities are not noisy, but have been modified like noise. The second equation doesn't have this drawback, but a part of the flux of a structure (compatible with noise amplitude) can be lost in the restoration process. It is, however, not as effective as in the standard maximum entropy methods.

B. A New Approach for Signal Restoration

The new definition of the information contained in noisy data can easily lead to a new approach for restoration of images.

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.

$$J(\tilde{D}) = H_s(D - \tilde{D}) + H_n(\tilde{D}). \tag{33}$$

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

$$J(\tilde{D}) = H_s(D - \tilde{D}) + \alpha H_n(\tilde{D}). \tag{34}$$

Here, α is considered as a constant value, but we can easily imagine having a regularization parameter per scale, or even per wavelet coefficient, depending on the signal-to-noise ratio of the data. This direction will be investigated in the future.

The following three points must be noted.

1) The positivity of the solution is not enforced.
2) There is no constraint on the flux.
3) The last scale of the wavelet transform is not taken into account in this entropy.

The first two points can be easily resolved by introducing strict *a priori* constraints on the solution [17]

$$J(Z) = H_s(D - \mathcal{C}(Z)) + \alpha H_s(\mathcal{C}(Z)). \tag{35}$$

And the real solution is evidently $\tilde{D} = \mathcal{C}(Z)$. Positivity and total flux conservation impose

$$\mathcal{C}(Z)(x) = \frac{\sum_x I(x)}{\sum_x Z(x)^2} Z(x)^2. \tag{36}$$

Any other constraint can evidently be introduced into the function $\mathcal{C}$.

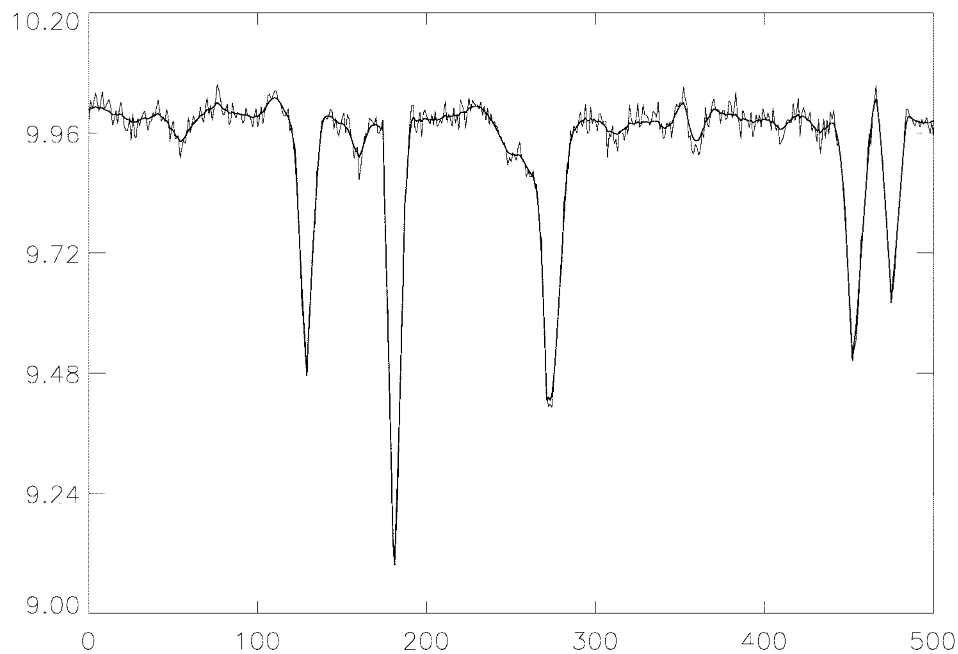


Fig. 3. Spectrum and filtered spectrum superimposed.

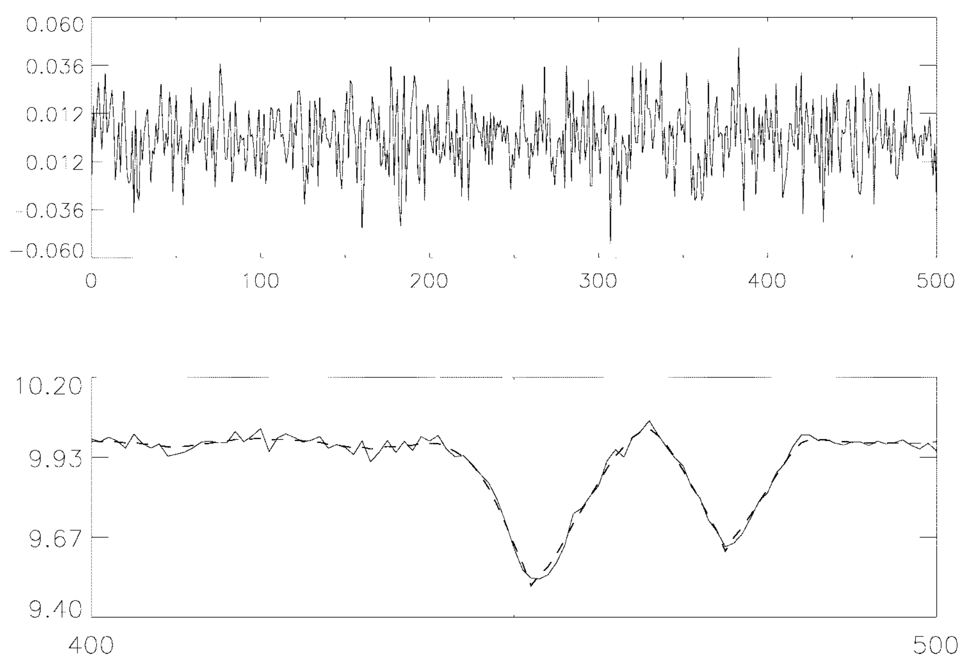


Fig. 4. Difference (upper part) between the real spectrum and its smoothed version. Part (pixels 400 to 500) of the spectrum (continuous curve), with the filtered spectrum overplotted (dashed).

There is no constraint to be introduced to cater to the third point above, but this should not be a problem if the number of scales we use for the entropy is high enough. Indeed, in this case, the last scale becomes flat, and flux normalization should correctly fix this level.

C. Example

Fig. 3 presents a spectrum and the result (overplotted) after filtering using the multiscale entropy. The difference between the spectrum and its smoothed version is plotted in Fig. 4 (upper part). As we can see, the residual contains only noise. In order to better see the quality of the smoothing, we have plotted only a part of the spectrum (see lower part of Fig. 4), and the filtered spectrum superimposed. The absorption lines are not modified using our filtering technique.

Fig. 5(a) shows the Lena image (cf. Fig. 1) to which Gaussian noise of standard deviation 10 has been added. Fig. 5(b) shows the result using (32) with a regularization parameter value of 2.

V. Conclusion

We have seen that information must be measured from the transformed data, and not from the data itself. This approach has been used in fact for several years in the domain of image compression. Indeed, modern image compression methods consist first of applying

(a)

(b)

Fig. 5. (a) Lena + Gaussian noise. (b) Filtered image.

a transformation (cosine transform for JPEG, wavelet transform, etc.) to the image, and then coding the coefficients obtained. A good transform for image compression is obviously an orthogonal transform because there is no redundancy, and the number of pixels is the same as in the original image. The exact number of bits necessary to code the coefficients is given by the Shannon entropy. For signal restoration, the problem is not to reduce the number of bits in the representation of the data, and we prefer to use a nonorthogonal wavelet transform, which avoids artifacts in reconstruction due to undersampling.

We could have used the Shannon entropy 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. First, we have seen that this leads, for Gaussian noise, to a very physical 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. Second, it works even in the case of images with few photons/events (the histograms in this case present a bias). We have seen that the equations are easy to manipulate. Finally, experiments have confirmed that this approach gives good results. We have also seen that our new information measure leads naturally to a new method for signal restoration. We are now experimenting with this method, and working on generalizations to other classes of noise.

References

[1] J. G. Ables, *Astronomy Astrophys. Suppl. Series,* vol. 15, pp. 383–393, 1974.

[2] A. Bijaoui, *Introduction au Traitement Numérique des Images.* Paris: Masson, 1984.

[3] Tj. R. Bontekoe, E. Koper, and D. J. M. Kester, "Pyramid maximum entropy images of IRAS survey data," *Astronomy Astrophysics,* vol. 294, pp. 1037–1053, 1994.

[4] J. P. Burg, presented at the *Annu. Meet. Int. Soc. Explor. Geophys.,* 1967; reprinted in *Modern Spectral Analysis,* D. G. Childers Ed. New York: IEEE Press, pp. 34–41, 1978.

[5] B. R. Frieden, "Image enhancement and restoration," in *Topics in Applied Physics,* vol. 6. Berlin: Springer-Verlag, 1975, pp. 177–249.

[6] ——, *Probability, Statistical Optics, and Data Testing: A Problem Solving Approach,* 2nd ed. Berlin: Springer-Verlag, 1991.

[7] E. T. Jaynes, *Phys. Rev.,* vol. 106, pp. 620–630, 1957.

[8] S. F. Gull and J. Skilling, *MEMSYS5 User's Manual,* 1991.

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

[10] E. Pantin and J. L. Starck, "Deconvolution of astronomical images using the multiresolution maximum entropy method," *Astronomy Astrophys. Suppl. Series,* vol. 118, pp. 575–585, 1996.

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

[12] J. Skilling and S. F. Gull, in *Proc. Amer. Math. Soc.,* vol. 14, 1984, p. 167.

[13] J. Skilling, "Classic maximum entropy," in *Maximum Entropy and Bayesian Methods,* J. Skilling Ed. Dordrecht: Kluwer, 1989, pp. 45–52.

[14] J. L. Starck, F. Murtagh, and A. Bijaoui, "Multiresolution support applied to image filtering and deconvolution," *Graph. Models Image Processing,* vol. 57, pp. 420–431, 1995.

[15] ——, *Image Processing and Data Analysis: The Multiscale Approach.* Cambridge: Cambridge University Press, 1998.

[16] N. Weir, "A multi-channel method of maximum entropy image restoration," in *Astronomical Data Analysis Software and System 1,* D. M. Worral, C. Biemesderfer, and J. Barnes Eds. San Francisco: Astronomical Society of the Pacific, 1992, pp. 186–190.

[17] E. Thiebault and J.-M. Conan, "Strict *a priori* constraints for maximum-likelihood blind deconvolution," *JOSA,* vol. 12, pp. 485–492, 1995.