

# Entropy and Gibbs distribution in Image Processing: an historical perspective

Sebastiano Battiato

*Università di Catania, Dipartimento di Matematica ed Informatica,  
Viale A. Doria 6, I-95125 Catania, Italy*

E-mail: {battiato@dmi.unict.it}

November 19, 2001

## Abstract

This paper presents an historical overview about the *entropy* and its applications for the solution of inferential statistical problems in image processing. This survey covers some of the more important *entropy*-based research approaches. A brief introduction to the mathematical details and foundations about the basic concepts of Markov Random Fields (MRF) and related Gibbs sampling is also given.

The information *entropy* is a mathematical measure of information or uncertainty derived from a probabilistic model. The paper starting from the seminal works of C. Shannon and of E.T. Jaynes and of S. Geman and D. Geman discusses results obtained using different related techniques in image restoration, analysis and synthesis of textures and saliency maps construction. The paper moreover gives useful suggestions about the trend of development in future research.

## 1 Introduction

After five decades from the first appearance in the literature the *entropy* concept is now universally known and widely used: new applications are increasingly coming from the core *old* idea of information measurement. The aim of this paper is to provide an overview of the present status and future perspective with particular emphasis to image processing. The presentation is organized according to the evolution of the ideas in the field: from the birth of Information Theory [ 23 ], to the contribution of E. T. Jaynes, the founding father of *entropy*-based techniques, whose research have greatly influenced the modern approach to the inferential statistical problems [ 19 ], to the more recent milestone provided by the work

of S. Geman and D. Geman [ 14 ]. At the same time , it is aim to provide an elementary introduction to the field in order to gain insights and perspective on future research. Roughly speaking, the information *entropy* has been introduced as a measure of uncertainty of a probability distribution and the experimental evidences have shown its fundamental usefulness. Despite its simplicity, it provides a reliable statistical approach where the resolution of *ill-conditioned* problems, imposes the assumption of maximum generality about the unknown parameters, a very common situation in the image processing. In this field *entropy* has found several important applications. Nowadays, the use of statistical assumptions to construct probability models incorporating the principal features of the images under study is a standard methodology to process digital images, representing real or artificial scenes.

Assembling together concepts of Physical Statistics and/or in Equilibrium Theory of chemical process, S. Geman and D. Geman published in 1984 a fundamental seminal paper [ 14 ] that has allowed the introduction in Computer Vision of techniques and methodology of works, strictly related with Gibbs distribution and Markov Random Field. Thanks to the research spurred from this work is now possible to obtain *true* samples of probability distribution, underlying digital images, whose probabilistic model is too complex to manipulate analitically. This is realized using a sampling techniques called *Gibbs Sampling* (see below for more details).

The paper is structured as follows. Section 2 introduces the *entropy* concept and reports some of its important properties. It also provides some of useful links about Information and Coding Theory. Section 3 introduces the *Maxent* or *Maximum entropy* principle due to E.T. Jaynes, with discussions and critiques about its application when only partial information about a given inferential statistical problem is available. Gibbs distributions and MRF modeling are introduced in next Section. In Section 5 some interesting applications in image processing using *entropy*-based approaches are reported. Finally new directions and future works are discussed.

## 2 Shannon's entropy: the measure of randomness

The concept of *entropy* in science, was introduced for the first time in 1948 [ 23 ], but it wasn't a very novel idea: it was already known in thermodynamics and in statistical mechanics. In particular Clausius and, later, Boltzmann gave the first functional expression for *entropy* as a measure of the degree of disorder existing in a thermodynamical system [ 31 ]. Motivated by the problem of efficient transmission of information over a noisy communication channel, Claude Shannon introduced a revolutionary new probabilistic way of thinking about communication and simultaneously created the first truly mathematical theory of *entropy*. His ideas were rapidly developed along two main lines of research:

- *Information Theory*: a probabilistic theory to study the statistical characteristics of data and communication systems;

- *Coding Theory*: a set of algebraic and geometric tools to obtain efficient codes for various communication problems. A well known algorithm in this field, for example, is the *Huffman encoding* where the value of the *entropy* provides a close estimate to the average length of the optimal injective encoding [ 15 ], [ 28 ].

A large set of *entropy*-based applications, other than those mentioned above, have been introduced since 1948. It is possible to apply *entropy* to the study of cryptosystems: the conditional *entropy* is used in order to measure the information about the key revealed by the ciphertext [ 5 ], [ 25 ]. Moreover, it is possible to find applications and related works in Economics, Biology, Social Sciences, Dynamical System, Logic and Theory of Algorithms, other than in Information-Coding Theory and statistical inference and prediction. See for more details the exhaustive Web site [ 16 ].

In this Section we will introduce only the basic ideas behind the *entropy* concept. The development and the application to image processing will be covered in the successive Sections.

## 2.1 Basic Concepts

After fifty years the original paper of Shannon is still, in many ways, the best introduction to the modern concept of *entropy*. This presentation is hence largely derived from it.

Suppose we have a random variable  $X$  which takes on a finite set of values according to the probability distribution  $p(X)$  with mutually exclusive events  $X = \{x_1, x_2, \dots, x_n\}$  and such that  $\sum_{i=1}^n p(x_i) = 1$ .

We are looking for a consistent measure of the uncertainty of such information-source; in other words we want to find the information gained by an event which takes place according to the distribution  $p(X)$ . The only partial knowledge that we have is the prior probability distribution. We can also state that in order to estimate the degree of uncertainty with a function  $H(X)$  the following assumptions must be granted:

1. The function  $H(X)$  exists: it is possible to set up some kind of associations between uncertainty and real numbers;
2.  $H(X)$  has a continuous functional dependence on the probability distribution  $p(X)$ : arbitrarily small changes in  $p(X)$  should lead to small changes in the amount of uncertainty;
3. If  $p(x_i) = \frac{1}{n}$   $i = 1, 2, \dots, n$  then  $H(X)$  should be a monotonic increasing function of  $n$ . In other words, when there are more possibilities, we are in a condition of greater uncertainty;
4.  $H(X)$  must be obeys to the *composition* or *additivity* law: if there is more than one way to work out its value, we must get the same answer for every possible way; e.g. if

a choice can be broken in more successive choices, the global  $H$  should be the weighted sum of the values of  $H$  computed for each choice.

It is possible to prove the following result, due to Shannon ([ 23 ]):

**Theorem 2.1** *The only function  $H$  that satisfies the properties 1-4, above, is of the form:*

$$H(X) = -k \sum_{i=1}^n p(x_i) \log p(x_i) \quad (1)$$

where  $k$  is a positive constant. It is common to set  $k$  to 1.

The function  $H$  is called *entropy*, *Shannon's entropy* or, better, *information entropy*. This name is due to the similar functional expression of the thermodynamical entropy but it is important to note that this is an unfortunate terminology. The *information entropy* defined above is a property of any probability distribution and the *experimental entropy* of thermodynamics is instead a property that measures different observed quantities as volume, pressure, temperatures of some physical system. In the following with the word *entropy* we will refer to the *information entropy*.

When  $p(x_i) = 0$  for some  $i$  the relative term in the expression above is not defined. Since  $\lim_{x \rightarrow 0} x \log x = 0$  there is no real difficulty allowing  $p(x_i) = 0$  for some  $i$ . However, in the sum we also implicitly consider only the terms not equal to zero.

It is possible to generalize the function *entropy* for a continuous information source. In this case the distribution of the random variable  $X$  is expressed in terms of probability density function (p.d.f.) which is assumed, for sake of simplicity, to be continuous. The functional expression for the *entropy* function  $H$  is in this case:

$$H(X) = -k \int_{-\infty}^{+\infty} p(x) \log p(x) dx \quad (2)$$

## 2.2 Some properties of information entropy

Some important *entropy* properties are listed as follows:

1.  $H(X) \geq 0$  and  $H(X) = 0$  if and only if  $p(x_i) = 1$  for some  $i$  and  $p(x_j) = 0$  for all  $i \neq j$ . It is a concrete measure of the uncertainty of the available data;
2.  $H(X)$  is maximal and equal to  $\log n$  when  $p(x_i) = \frac{1}{n}$  for all  $i = 1, 2, \dots, n$ . This property states that the maximum uncertainty is present when all events have the same probability to occur;
3. If  $X$  and  $Y$  are two random variables, then  $H(X, Y) \leq H(X) + H(Y)$  where  $H(X, Y)$  denote the *entropy* of the joint probability distribution. Equality holds if and only if  $X$  and  $Y$  are independent events;

Figure 1: a) Chinese character for the Entropy b) Entropy of two events with probability  $p$  and  $(1 - p)$  as a function of  $p$ .

4. If  $X$  and  $Y$  are two random variables and we are interested to consider the conditional probability distribution  $(X | Y)$  it is possible to define the *conditional entropy*  $H(X | Y)$  as follows:

$$H(X | Y) = - \sum_y p(y) \sum_x p(x | y) \log p(x | y). \quad (3)$$

The conditional *entropy* is the weighted average, with respect to the probability  $p(y)$  of the entropies  $H(X | y)$  over all possible values  $y$ ;

5.  $H(X, Y) = H(X | Y) + H(Y)$ . Note that when  $X$  and  $Y$  are independent we have  $H(X | Y) = H(X)$ .

A proof of the results illustrated above can be found in [ 23 ] or in [ 25 ].

### 3 The MaxEnt principle

The next important step in the application of *entropy* to digital signal processing is the, so called, **MaxEnt** principle. The **MaxEnt** principle in statistical inference and prediction was introduced in 1957 by E.T.Jaynes [ 19 ]. Jaynes reformulated statistical mechanics results, due to J.Willard Gibbs, in terms of probability distributions using the principle of maximum *entropy*. This reformulation of the theory simplifies the mathematics, allowing for fundamental extensions of the theory. In this framework even statistical mechanics could

be reinterpreted as inference based on incomplete information.

The **MaxEnt Principle** simply states:

**Definition 3.1** *Given a collection of facts and/or experimental data, in order to estimate the underlying probability distribution it is sufficient to choose a model which is consistent with all the facts, but otherwise has maximum entropy.*

This principle assumes a great relevance solving problems based on measured and very often incomplete or noisy data. The **MaxEnt** principle wasn't quickly accepted in the scientific community: it took several years before it was universally understood and used in different research areas [ 20 ]. The power of this idea, however, is today testified by the fact that every year there is an international **MaxEnt** meeting where it is possible to find the more recent applications and results of the method with regard to different areas: mathematics, physics, engineering, natural language, image processing and so on.

In practice if a certain probability distribution maximizes *entropy* an automatic justification for using that distribution for the inference is obtained. In this manner nothing more of what is explicitly known is assumed. It is a transcription into mathematics of an ancient principle of wisdom: *Nunquam ponenda est pluralitas sine necessitate*, (Occam's Razor). Strictly related with this idea is the *Laplacian principle of indifference*; it states that two or more events are to be assigned equal probabilities unless there is a reason to think otherwise.

To see better the relation existing between the *Laplacian principle of indifference* and the **MaxEnt** principle consider the following optimization problem:

$$\text{Max}(H(p_1, p_2, \dots, p_n) = - \sum_{i=1}^n p_i \log p_i) \quad (4)$$

with

$$\sum_{i=1}^n p_i = 1 \quad (5)$$

Applying the Lagrange multiplier method the solution one finds:  $p_i = \frac{1}{n}$   $i = 1, 2, \dots, n$ . In other words the maximum *entropy* approach does not make any unwarranted assumptions. It learns exactly what the data says: this is the main difference with other traditional predictive models (e.g. neural networks) that inadvertently make spurious assumptions about their data. The availability of powerful computers has allowed the application of **MaxEnt** principle to high dimension data problem, with significative improvements over the previous techniques.

### 3.1 The mathematical details

Suppose a random variable  $X$  can take  $n$  different values  $\{x_1, x_2, \dots, x_n\}$  with an unknown probability distribution. Suppose moreover that there are  $m$  different functions of  $x$ , called **features**:  $f_k(X)$ ,  $k = 1, \dots, m$ . Usually **features** are strictly related with the inference problem and represent the marginal distributions of the model under detection. These functions can be used in constraining the expected value that the model assigns to the corresponding feature function:

$$\tilde{F}_k = \sum_{i=1}^n p_i f_k(x_i) \quad k = 1, \dots, m. \quad (6)$$

where  $\tilde{F}_k$  is the empirical distribution of  $x_i$  directly measured on the available training data. The goal is to construct a statistical model  $p$  of the random variable  $X$  which is generated from a given training sample set. Applying the **MaxEnt** principle (3.1), it is enough to find the solution to the following problem:

$$\text{Max}(H(p_1, p_2, \dots, p_n) = - \sum_{i=1}^n p_i \log p_i) \quad (7)$$

subject to the constraints (5) and (6).

It is possible to solve the problem applying the method of Lagrange multipliers from the theory of constrained optimization, see also [ 19 ] for further details. For each feature  $f_i$  there is a parameter  $\lambda_i$ , a Lagrange multiplier such that the problem is reduced to compute the unconstrained maximum of the Lagrangian:

$$\Lambda(p, \lambda) = H(p_1, \dots, p_n) + \sum_{k=1}^m \lambda_k \left( \sum_{i=1}^n p_i f_k(x_i) - \tilde{F}_k \right) \quad (8)$$

whose solution is:

$$p_\lambda(x_i) = \frac{1}{Z(\lambda_1, \dots, \lambda_m)} e^{-\sum_{k=1}^m \lambda_k f_k(x_i)} \quad i = 1, \dots, n. \quad (9)$$

where  $Z(\Lambda)$  is a normalizing constant function determined by the constraint (5):

$$Z(\lambda_1, \dots, \lambda_m) = \sum_{i=1}^n e^{(-\lambda_1 f_1(x_i) - \dots - \lambda_m f_m(x_i))} \quad (10)$$

It is possible to prove that the value  $H_{max}$  is a function  $S$  of  $\tilde{F}_1, \dots, \tilde{F}_m$ :

$$H_{max} = S(\tilde{F}_1, \dots, \tilde{F}_m) = \log Z(\lambda_1, \dots, \lambda_m) + \sum_{k=1}^m \lambda_k \tilde{F}_k \quad (11)$$

From this an explicit expression for the  $\lambda_k$ 's can be easily derived:

$$\lambda_k = \frac{\partial S}{\partial \tilde{F}_k} \quad k = 1, \dots, m. \quad (12)$$

However it is not so easy to find the value  $\lambda_k$  analitically because the substantial incalculability of the function  $Z$ . For this reason in real applications it is necessary to apply numerical techniques (see next section).

A key factor of this approach is the right choice of the set of features. This problem has been studied in [ 10 ] [ 34 ]. It will be considered with some detail in Section 5; here suffices to note that this approach is powerful but high expensive in terms of computational complexity, so it is necessary to select a set of features, sufficiently expressive but at the same time not too big. Other than in Image Processing the **MaxEnt** principle is also applied directly in the frequency domain, on the Fourier Transform of the input data and finds application in 1-D Spectral Analysis area where object of study is the processing of radar, seismic and speech signals ([ 30 ]). It is possible to find several applications of the **MaxEnt** principle in the construction of statistical models in natural language processing [ 4 ] or in experiments conducted on nuclear magnetic resonance [ 11 ]. Yet, a simulated annealing algorithm based on the same principle has been proposed, together with a well-known fuzzy  $c$ -partition algorithm, in order to obtain an effective threshold selection method [ 7 ], [ 17 ].

## 4 Gibbs Random Fields

The Principle of the **Maximum Entropy** is strictly related with the **Gibbs distribution** and **Markov Random Field** (MRF) modeling theory. The final expression (9) of the Lagrangian function in Sect.3.1, naturally describes a well known function used in statistical mechanics, physics and in thermodynamic. Geman and Geman in a pioneering paper published in 1984 [ 14 ] introduced the possibility to apply similar techniques also in Computer Vision. They suggest to process pixel gray levels and image properties like molecules or atoms in physical systems applying ideas and techniques used for the study of equilibrium states of chemicals processes at different temperatures. In this Section we give a brief introduction to the basic concepts of such methodology. It should be remarked that Geman and Geman obtained their results in an independent way with respect to the existing works regarding the application of **MaxEnt** principle. As we shall see in the next Section, only recently have been published some papers where these results are re-combined and studied together [ 4 ], [ 10 ], [ 33 ].

### 4.1 Markov Random Fields and Gibbs Distribution

Let  $S = \{s_1, s_2, \dots, s_N\}$  be a set of *sites* or *places* defined on a lattice system (e.g. a given image **I**) and let  $\Psi = \{\Psi_s, s \in S\}$  be a *neighborhood* system for  $S$  that is a collection of

subsets of  $S$  such that  $\forall s \in S$  we have 1)  $s \notin \Psi_s$  2)  $s \in \Psi_r \Leftrightarrow r \in \Psi_s$ . The elements in  $\Psi_s$  are the *neighbors* of  $s$ . A subset of  $S$  is a *clique*  $C$  if every pair of distinct sites in  $C$  are neighbors. Based on different specific application areas, different neighborhoods systems can be identified and used. For example, a well-known homogeneous neighborhoods system in image processing is the following:

$$\Psi = \{\mathfrak{S}_{x,y}, (x, y) \in I\} \quad (13)$$

where

$$\mathfrak{S}_{x,y} = \{(k, l) \in I : 0 < (k - x)^2 + (l - y)^2 \leq c\} \quad (14)$$

The parameter  $c$  controls the size of the system. Obviously, the number of *cliques* types grows rapidly with  $c$ . Another neighborhood system, actually applied to the problem of automatic words classification in natural language processing, is described in detail in [ 4 ]: this approach uses the features learned with the proposed algorithm regarding natural linguistic properties (single lower-case letter, single upper-case letter, adjacent lower-case letters, etc.) in order to redefine these information as graphs features.

**Definition 4.1** Let  $X = \{X_s, s \in S\}$  be a random variable indexed by  $S$ , let  $G$  be the set of possible values (for example number of gray levels in an image) such that  $X_s \in G, \forall s \in S$  and let  $\Omega$  be the set of all possible configurations.  $X$  is a Markov Random Field (MRF) with respect to  $\Psi$  if  $P(X = \omega) \geq 0, \forall \omega \in \Omega$  and  $P(X_s = x_s | X_r = x_r, r \neq s) = P(X_s = x_s | X_r = x_r, r \in \Psi_s)$ .

Roughly speaking in a MRF is possible to evaluate the probability to have a specific value in a *state* of the system having only knowledge of the relative neighbors set.

**Definition 4.2** Given a set of sites  $S$  and a neighborhood system  $\Psi$ , a Gibbs distribution is a probability measure  $p$  on  $\Omega$  such that:

$$p(\omega) = \frac{1}{Z} e^{\frac{-U(\omega)}{T}} \quad (15)$$

where  $Z$  is the normalizing or partition constant defined as

$$Z = \sum_{\omega \in \Omega} e^{\frac{-U(\omega)}{T}} \quad (16)$$

$T$  is the constant relative to the temperature and  $U$  is the energy function of the form

$$U(\omega) = \sum_C V_C(\omega) \quad (17)$$

where the  $V_C$  functions are called potentials or potential functions and are referred to specific cliques  $C$ .

The *potential* functions are clique dependent, that is their value depends only on the values assumed by the states presents in the given clique  $C$ . The global parameter  $T$ , is used to gradually simulate an *annealing* process that converges to an equilibrium stage of the system. This technique, has been proposed in [ 14 ] in order to obtain a Maximum *a posteriori* (MAP) estimate of an observed and degraded input image. The choice of the *neighborhood* system is crucial and is strictly related with the real-world knowledge about the problem under processing.

## 4.2 The Gibbs Sampling

One of the biggest problem working with the Gibbs distribution is the constant  $Z$ ; it is often impossible to calculate its value explicitly because it involves a great number of possible configurations. To obtain some quantitative information, simulation is required and a good sampling procedure is an unavoidable step in most of the practical problem. Such sampling procedures generally involve *Monte Carlo Markov Chain* (MCMC) methods (e.g *Metropolis* algorithms) that work only on samples extracted from the theoretical distribution. This can be accomplished generating a sample sequence. The evolution of the sample along the sequence is controlled from the energy variations. The operation of sampling cannot be done deterministically in order to avoid to pick out local extrema of the energy function. This methodology permits to converge versus *true* samples of the distribution with *low* temperature that are more reliable and expressive.

A fundamental relationship between MRF and Gibbs distribution is stated in the following theorem, called the *Hammersley-Clifford's theorem*:

**Theorem 4.1** *Let  $\Psi$  be a neighborhood system. Then  $X$  is a MRF with respect to  $\Psi$  if and only if  $p(\omega) = P(X = \omega)$  is a Gibbs distribution with respect to  $\Psi$ .*

The proof can be found in [ 3 ]. Thanks to this important theoretical result is possible to combine together MRF and Gibbs distribution to properly extract sample from the population in accord with the distribution under detection. Let us observe that the conditional probability with respect to the local characteristics of a Gibbs distribution is expressed as:

$$p(x_s | x_r, r \neq s) = \frac{p(\omega)}{\sum_{x_s \in G} p(\omega)} \quad s \in S, \quad \omega \in \Omega \quad (18)$$

It is now possible to express this distribution in a more effective way using theorem (4.1). Fix  $\omega \in \Omega$  and  $s \in S$ , if  $p(\omega) = P(X = \omega)$  is Gibbsian and  $X$  is a MRF then

$$P(X_s = x_s | X_r = x_r, r \neq s) = \frac{1}{Z_s} e^{-\frac{1}{T} \sum_{C:s \in C} V_C(\omega)} \quad (19)$$

where

$$Z_s = \sum_{x \in G} e^{-\frac{1}{T} \sum_{C:s \in C} V_C(\omega^x)} \quad (20)$$

The symbol  $\omega^x$  denotes the configuration whose value is  $x$  at site  $s$  and agrees with  $\omega$  everywhere else. This is the basic idea of *Gibbs sampling*. This term denotes a sampling algorithm that with a sequence of suitable replacements of *sites* converges to a sample drawn from the theoretical distribution. The sampling generates a sequence of configurations  $X(t) = (X_{s_1}(t), X_{s_2}(t), \dots)$  for  $t = 1, 2, \dots$ . Each configuration is obtained sampling under the conditional probability above, in the following manner: all the *sites* are repeatedly visited and only one *site* could be changed at each time  $t$  realizing what is called a *sweep*. Given the site  $s_t$  and knowing the functional expression of the general probability model  $P$ , it is possible to sample a new value  $x \in G$  for it, in accordance with (18) and (19), where  $s = s_t$  and  $\omega = X(t - 1)$ . In other words we obtain a new configuration where  $X_{s_t}(t) = x$  and  $X_s = X_s(t - 1)$ ,  $s \neq s_t$ . The *Gibbs sampling* is hence an iterative process capable to evaluate in each step, the probability distribution to have a *change* in a given *site*, using the (19), simply observing the neighbors of the *site* in object. It is possible to show that this simple algorithm converges to the distribution  $P$  regardless of the initial configuration. The convergence is realized also using a decreasing sequence of temperatures. The proof of the correctness of these algorithms called *Relaxation* and *Annealing*, can be found in [ 14 ].

### 4.3 A toy example

In order to better understanding the Gibbs Sampling and its practical usefulness in statistical problems, a typical example is here presented and discussed. Starting from a random string of ASCII characters, of unspecified length, it is possible to generate strings obeying a specific probabilistic model. For example if we want to model a string of the form *bab*, it is possible to apply the Gibbs Sampling using the following exponential function:

$$p(\omega) = \frac{1}{Z} e^{\sum_I f(\omega)} \quad (21)$$

where

$$Z = \sum_{\omega} e^{\sum_I f(\omega)} \quad (22)$$

and the *feature* function  $f$  expresses the *weight* associated with the characteristic that we want to model. In this case it assumes its maximum value i.e. a constant  $K$ , when, in the given neighborhood  $I$ , i.e. 4 closest letters, there is exactly the configuration under detection *bab*, zero otherwise. In this manner, using the Gibbs Sampling described above, it is possible to apply the conditional probability (18) and (19) straightaway. Restricting for sake of simplicity the range of the possible ASCII characters to  $\mathbf{A}=\{a,b,c\}$ , and supposing to extract, at a certain time of computation, a *place* corresponding to the center  $x$  of the substring **cbcb**a, the sampling procedure computes:

$$p(x = a | \mathbf{cb\_ba}) = \frac{e^{f_{MAX}}}{e^{f_{MAX}} + e^0 + e^0} = \frac{e^K}{e^K + 2}$$

$$p(x = b|\mathbf{cb\_ba}) = \frac{e^0}{e^{f_{MAX}} + e^0 + e^0} = \frac{1}{e^K + 2}$$

$$p(x = c|\mathbf{cb\_ba}) = \frac{e^0}{e^{f_{MAX}} + e^0 + e^0} = \frac{1}{e^K + 2}$$

The choice of the replacement letter that substitutes the original values  $c$  is simply obtained sampling the right value/letter in accord with the distribution above. Obviously in the given example, it will be more probable to choose the letter  $a$  that reproduces exactly the distribution in object than the other two.

## 5 Applications to Image Processing

In the previous sections we have presented the historical evolution that starting from the ordinary concept of *entropy*, has allowed to obtain new and powerful techniques. In this Section we will describe as these results can be applied in Image Processing in order to find significative and effective responses to different and specific topics of this area.

### 5.1 The MiniMax entropy principle and texture modeling

Recently, Zhu and Mumford [ 33 ], [ 34 ] have proposed a new theory for building statistical models for images (or signals) in a variety of applications. In their work filtering theory, information theoretical arguments and Markov Random Field [ 9 ], [ 14 ] come together in a general purpose automated learning approach. Let  $F_\alpha(\mathbf{I})$ ,  $\alpha = 1, 2, \dots, k$  be a series of well specified features of a given image  $\mathbf{I}$ . The approach uses as a fundamental statistical indicator, for each feature  $F_\alpha(\mathbf{I})$ , the value  $E_f(F_\alpha(\mathbf{I}))$  i.e. the expectation of the underlying probability distribution  $f(\mathbf{I})$ . Since the distribution  $f(\mathbf{I})$  is generally unknown the values  $E_f(F_\alpha(\mathbf{I}))$  can be estimated by the sample mean of the feature computed from the training images. What is wanted is to construct a model  $p(\mathbf{I})$  such that:

$$E_p(F_\alpha(\mathbf{I})) = E_f(F_\alpha(\mathbf{I})), \quad \alpha = 1, 2, \dots, k \quad (23)$$

Zhu and mumford suggest to obtain one such model balancing between model generality and model simplicity by two seemingly contrary criteria:

1. **The maximum entropy principle** (3.1). Among all models  $p(\mathbf{I})$  satisfying (23) this principle choose the simplest one, maximizing the *entropy* over all distribution that reproduces the particular feature statistics.
2. **The minimum entropy principle.** It applies on feature selection/extraction phase: among all plausible sets of feature statistics choose the set whose maximum *entropy* distribution has the minimum Kullback-Leibler distance between  $f(\mathbf{I})$  and  $p(\mathbf{I})$ . It is possible to prove that this value is equal to the *entropy* of the model  $p(\mathbf{I})$  (up to a constant).

The estimate of the joint probability distribution  $f(\mathbf{I})$  plays a significant role in several areas of computer vision like visual coding, pattern recognition, neural networks, statistical decision theory, computational vision. Of particular interest is the application of this approach to analysis and synthesis of texture images where a model called FRAME (Filters, Random fields, And Maximum Entropy) has been introduced with promising results [ 34 ]. In this case the choice of the features, given a training texture image, concerns the histograms of the filtered image with a *suitable* set of filters. Filters can have kernel of any size, and can be linear or nonlinear. These histograms represents an estimate of the marginal distribution of the underlying  $f(\mathbf{I})$ . Experimental results show that it is sufficient to choose a limited set of features/filters. The parameters of the estimated model  $p(\mathbf{I})$  are obtained by an algorithm that does an extensive use of the Gibbs sampling discussed above, both in the feature selection phase and in the model construction phase. The derived FRAME model has the following functional expression:

$$p_{\Lambda_k}(\mathbf{I}) = \frac{1}{Z(\Lambda_k)} e^{-\sum_{\alpha=1}^k \lambda_{\alpha}(\mathbf{I}_{\alpha})} \quad (24)$$

where  $k$  is the number of filters,  $\Lambda_k = (\lambda_1, \lambda_2, \dots, \lambda_k)$  and  $\mathbf{I}_{\alpha}$  is the image  $\mathbf{I}$  filtered by the filter  $\alpha$ . The  $\lambda_{\alpha}$ ,  $\alpha = 1, 2, \dots, k$  are optimal values obtained using an iterative algorithm based on the following equations:

$$\frac{\partial \lambda_{\alpha}}{\partial t} = E_{p_{\Lambda_k}}(H_{\alpha}) - H_{\alpha}^{obs} \quad (25)$$

where  $H_{\alpha}$  and  $H_{\alpha}^{obs}$  are respectively the histograms of the estimated model image and of the observed input image either filtered by  $\alpha$ . The expressive power of the existing inference model is limited by the exponential number of parameters that must be considered in order to obtain significative results. In Mumford's approach the computational cost is high, although it is able to reproduce also non-Gaussian and/or highly complex visual structures of input texture images. The previous MRF models are strongly conditioned by the neighborhood size: the parameter size is too large also for a relative modest neighborhood (e.g. 13 x 13). The FRAME model increases the expressiveness characterizing the local interactions by non-linear function of filter responses. An example of textures synthesized with this method is shown in Fig.2 where the two input textures are placed on the left hand side of the page.

## 5.2 Image Restoration

The quality of digital images representing real or artificial scenes is often reduced by different reasons (e.g. physical acquisition factors, transmission, ecc.) that introduce noise in the images. In order to extract useful information from a degraded image it is necessary to

Figure 2: Some results obtained applying the FRAME model

*invert*, if possible, this frequently unknown process of degradation. In general, if we suppose to have *additive* noise, all that we have is a degraded image  $I$  obtained as:

$$I = O * H + N \quad (26)$$

where  $N$  is the noise,  $H$  is a linear or non linear filter applied to the initial image  $O$  and  $*$  is the classical convolution operator. REstoration algorithm try to invert the relation above and to obtain an image as close to the true image as possible in its perceptive characteristics. Different lines of research are present in literature about the image restoration problem; the same work of Geman and Geman [ 14 ] mentioned above was devoted in this direction but here we are interested to survey only the *entropy* related approaches. The first applications involving the **MaxEnt** principle were due to Frieden [ 12 ], [ 13 ]. In these works he shows a series of restoration methods *entropy*-based that naturally give a standard maximization problem. In fact, if we consider the formation process of an image as a series of silver grains that falls in a *random* position obeying a specific probability law it is possible to express the total number of possible ways  $W$ , referred to a given set of pixels  $(o_1, o_2, \dots, o_N)$  in such a way:

$$W(o_1, o_2, \dots, o_N) = \frac{(\sum_{i=1}^N o_i)!}{o_1! o_2! \dots o_N!} \quad (27)$$

With simple re-formulation (that can be seen in [ 19 ] and in [ 30 ]) it is possible to prove that maximize the expression (27) is the same that maximize the *entropy* associated with the observed  $\{o_i, i = 1, \dots, N\}$  subject to the constraints due to the noise  $N$ . The final expression for the matrix  $O$  is:

$$\tilde{o}_i = e^{(-1 - \sum_m \lambda_m h_{i,m})} \quad i = 1, 2, \dots, N \quad (28)$$

where the index  $m$  indicates the number of observed available data,  $\lambda_i$  are the free parameters defining the object and the  $h_{i,m}$  are the corresponding values in the matrix  $H$  above. Other related works and algorithms can be found in [ 30 ]. Most of this works are used in astronomy, medical imaging, X-Ray diagnosis, SAR images and so on.

The FRAME model of Zhu and Mumford [ 32 ] has several applications in this field. In particular the methods classifies the potential functions learned by the *Minimax entropy* approach in two different categories: *diffusion terms* and *reaction terms*. Experimental studies show that *diffusion terms* work like denoising components while *reaction terms* form patterns and enhance preferred image features. Following these ideas, Zhu and Mumford have proposed a model called *Gibbs Reaction and Diffusion Equations (GRADE)*. Using this model it is possible to obtain a suitable energy function  $U$  for texture pattern, rendering, denoising and image enhancement. See Fig.3 for an image processed according to this methodology.

Figure 3: GRADE clutter removal

### 5.3 Saliency Maps Construction

We have recently proposed [ 1 ] adopted a simple but promising algorithm to construct **Saliency Maps** i.e. a function  $S$ , useful to detect object in digital scenes especially in early vision tasks. This new image-space approach learns the most relevant, in an information-theoretic sense, mixture of measured characteristic of a pixel to determine a saliency value. Given an image  $\mathbf{I}$ , a saliency map is a function:

$$S : I \rightarrow [0, 1] \tag{29}$$

$S(x, y)$  is a measurements of the likelihood that a pixel in a digital scene of intensity  $\mathbf{I}(x, y)$  bears the trace of the object under detection. Saliency maps are a useful tool in the preliminary detection of subsets of the picture where to focus the attention for more refined and precise higher level of processing. The method tries to bring together *entropy*-based image processing methods and statistic smoothing of irregular and sparse data.

There are two general approaches to the construction of saliency maps: scene-based and image-based. In the first approach 3D information about the object that have generated the images under examination is available and is integrated in the analysis. The second approach, on the other hand, tries to classify the pixels of an image according to properties that can be measured and detected in the image itself. The method tries to bring together *entropy*-based image processing methods [ 13 ] and statistics smoothing of irregular and sparse data [ 27 ] to obtain a saliency map algorithm. The algorithm resembles those based on Principal Components Analysis [ 21 ]. The main idea here is to learn from a training set of positive example the *right mixture* of features in order to obtain a Saliency map as a *weighted* linear combination of simpler features this approach is known in literature

as feature extraction [ 18 ], [ 26 ]. The weights in the linear combination are chosen as function of the *entropy* of the observed distribution of each feature. Finiteness of the training set of positive examples could lead to irregular histograms. Since *entropy* is sensitive to this situation, a more robust estimation of the weights can be obtained using the *entropy* of histograms that are previously regularized by a smoothing Gaussian kernel. Given in input a sequence of  $n$  positive examples of the object to recognize  $E_1, E_2, \dots, E_n$ , a set of  $h$  features  $F_1, F_2, \dots, F_h$  and one picture to analyze  $\mathbf{I}$  the algorithm works in two distinct phases. First, for each feature  $F_j$ ,  $j = 1, 2, \dots, h$  it derives the relative histogram  $Histo(F_j)$  where  $Histo(F_j)(v)$  denotes the observed frequency of the value  $v$  of the feature  $F_j$  over the universe of positive example pixels. The histogram shapes are smoothed through a convolution with a Gaussian Kernel of moderate size. The regularized histogram is denoted as  $HistoR(F_j)$ .

Finally for each characteristic feature  $F_j$  a weight factor  $w_j$  can be assigned having the following functional expression:

$$w_j = (K_1 f(j) + K_2 H_j)^{-1} \quad j = 1, 2, \dots, h \quad (30)$$

where  $f(j)$  is a user-defined function that incorporates the a priori knowledge about the recognition problem,  $H_j$  indicates the *Shannon's entropy* of the random variable whose statistical distribution is given by the regularized histogram  $HistoR(F_j)$  and  $K_1, K_2$  are suitable positive constants. The final expression for the Saliency map is:

$$S(x, y) = \sum_j w_j G(F_j, HistoR(F_j), x, y) \quad j = 1, 2, \dots, h \quad (31)$$

where:

$$G(F_j, HistoR(F_j), x, y) = \left( 1 + \frac{|F_j(x, y) - HistoR(F_j)^{MAX}|}{Z_j} \right)^{-1} \quad (32)$$

The symbol  $Z_j$  denotes a normalization constant while  $HistoR(F_j)^{MAX}$  indicates the modal value of the relative histogram. The preliminary results obtained can be seen in Fig.4. In this particular example two saliency maps able to detect verticals rows with a given gray level, are constructed by two different training set of positive examples. The saliency maps are depicted on the right hand side of the figure, where brightness indicates an higher degree of probability that in such a zone there is the vertical row under detection. The combination of a set of features to use in recognition, as the daily practice of image processing shows, is a key factor to boost the performance of a recognizer. Our techniques allows to take into account very large sets of features with a moderate burden on the computational resources required. The feature set should, moreover, considered as an input, or at least, as a crucial parameter of the algorithm. Observe that, in general, is preferable to take into account many features and there is no need for a preliminary decision on the relevance of each one of them. The final step of the learning phase, indeed, automatically assigns weights to each characteristic. Our first results are obtained choosing as features the gray levels of each pixel

Figure 4: Preliminary results

at a given position. Further improvement of the idea regards the possibility to choose the features in a more sophisticated way. Recent studies show that the marginal distribution of an unknown probability model can be entirely reconstructed using its marginal distribution approximated by the histograms of the original images filtered by an effective set of filters.

## 6 Conclusions and future works

This paper traces the historical development of the *entropy* concept and of its applications to image processing. We have presented an overview of the main theoretical results from the original idea of C. Shannon to the Principle of the Maximum Entropy of E.T. Jaynes to the use of Gibbs distribution in image processing pioneered by S. Geman and D. Geman. A brief introduction to the mathematical details and various applicative methodology where the information *entropy* find significative applicability is given. The underlying ideas and techniques are described and discussed in detail. Our historical excursion ends with a quick overview of recent important results of Mumford and Zhu. Their minimax entropy principle is a powerful tool to process textures in digital images and it is our feeling that it will lead to further advances in image processing. The final example where the use of entropy has been demonstrated, is a very simple but effective algorithm for feature extraction.

## Acknowledgments

The author wishes to thank G. Gallo for help and assistance during the redaction of the paper.

## References

- [ 1 ] S.Battiato, G.Gallo, *An Information-Theoretical Approach to Saliency Maps Construction*, In *Proceedings EUFIT'98*, Aachen, Germany, Vol.2, pp.1375-1380, 1998.

- [ 2 ] S.Battiato, G.Gallo, *Multi-resolution Clustering of Texture Images* Chapter in *Texture Analysis in Machine Vision*, Ed. M. Pietikinen - Series in Machine Perception and Artificial Intelligence - Vol. 40, pp.41-51, World Scientific, October 2000.
- [ 3 ] J.Besag, *Spatial interaction and the statistical analysis of lattice systems (with discussion)*, J. Royal Statist. Soc., series B, Vol.36, pp.192-236, 1973.
- [ 4 ] A.L.Berger, S.A.Della Pietra, V.J.Della Pietra, *A Maximum Entropy Approach to Natural Language Processing*, Computational Linguistics, Vol.22, No.1, 1996.
- [ 5 ] G.Brassard, P.Bratley, *Algorithms, Theory and Practice*, Prentice Hall, 1988.
- [ 6 ] J.P.Burg, Maximum entropy spectral analysis, PhD dissertation, Stanford University, Stanford, CA, 1975.
- [ 7 ] H.D.Cheng, J.R.Chen, J.Li, *Threshold selection based on fuzzy c-partition entropy approach*, Pattern Recognition, Vol.31, No.7, pp.857-870, 1998.
- [ 8 ] Z.Chi, H.Yan, T.Pham, *Fuzzy Algorithm: With Applications to Image Processing and Pattern Recognition*, World Scientific, 1996.
- [ 9 ] G.R.Cross, A.K.Jain, *Markov Random Field texture models*, IEEE Trans. PAMI, Vol.5, pp.25-39, 1983.
- [ 10 ] S.Della Pietra, V.Della Pietra, J.Lafferty, *Inducing features on random fields*, IEEE Trans. PAMI, Vol.19, No.4, 1997.
- [ 11 ] M.A.Delsuc, *A new maximum entropy processing algorithm, with applications to nuclear magnetic resonance experiments*, In *Maximum entropy and Bayesian Methods*, Ed. by J.Skilling, Netherlands, pp. 285-290, 1989.
- [ 12 ] B.R.Frieden, J. Optical Soc. Amer., Vol.62, pp.511-518, 1972.
- [ 13 ] B.R.Frieden, *Statistical Models for the Image Restoration Problem*, Computer Graphics and Image Processing, Vol.12, pp.40-59, 1980.
- [ 14 ] S.Geman, D.Geman, *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*, IEEE Trans. PAMI, Vol.6, No.6, pp.721-741, 1984.
- [ 15 ] C.M.Goldie, R.G.E.Pinch, *Communication theory*, Cambridge University Press, 1991.
- [ 16 ] Chris Hillman, *Entropy on the Web*,  
< <http://www.math.washington.edu/~hillman/entropy.html> > , 1998.

- [ 17 ] L.K.Huang, M.J.Wang, *Image thresholding by minimizing the measure of fuzziness*, *Pattern Recognition* , No.28, pp.41-51, 1995.
- [ 18 ] A.Jain, D.Zongker, *Feature Selection: Evaluation, Application, and Small Sample Performance*, IEEE Trans. PAMI, Vol.19, No.2, pp.153-158, Feb. 1997.
- [ 19 ] E.T.Jaynes, *Information Theory and statistical mechanics*, Physical Review 106, pp.620-630, 1957.
- [ 20 ] E.T.Jaynes, *Notes on present status and future prospects*, In W.T.Grandy and L.H.Schick *Maximum Entropy and Bayesian Methods*, Kluwer, 1-13, 1990.
- [ 21 ] I.Jolliffe, *Principal Component Analysis*, New York, Springer Verlag, 1986.
- [ 22 ] B.Moghaddam, A.Pentland, *Probabilistic Visual learning for Object representation*, IEEE Trans. On PAMI Vol.19, No.7, pp.696-710, 1997.
- [ 23 ] C.E.Shannon, *A Mathematical Theory of Communication*, The Bell System Technical J., Vol.27, 1948.
- [ 24 ] C.E.Shannon, *Communication Theory of Secrecy Systems*, The Bell System Technical J., Vol.28, 1949.
- [ 25 ] D.R.Stinson, *Cryptography: Theory and Practice*, CRC Press, 1995.
- [ 26 ] M.Turk, A.Pentland, *Eigenfaces for recognition*, J.of Cognitive Neuroscience, Vol.3, No.1, 1991.
- [ 27 ] R.R.Yager, D.P.Filev, *A generalized de-fuzzification method via BAD distributions*, Intern.Jour.of Intelligent Systems, Vol.6, pp.687-697, 1991.
- [ 28 ] D.Welsh, *Codes and Cryptography*, Oxford Science Publications, 1988.
- [ 29 ] G.Winkler, *Image Analysis, Random Fields and dynamic Monte Carlo Methods*, Springer-Verlag, 1995.
- [ 30 ] N.Wu, *The Maximum Entropy Method*, Springer Verlag Series in Information Sciences, Vol.32, Springer Verlag, 1997.
- [ 31 ] M.W.Zemansky, *Heat and Thermodynamics*, McGraw-Hill, New York, NY 1981.
- [ 32 ] S.C.Zhu, D.Mumford, *Prior Learning and Gibbs Reaction-Diffusion*, IEEE Trans. PAMI, Vol.18, No.11, pp.1236-1250, Nov. 1997.
- [ 33 ] S.C.Zhu, Y.N.Wu, D.Mumford, *Minimax Entropy Principle and Its Application to texture modeling*, Neural Computation Vol.9, No.8, Nov. 1997.

- [ 34 ] S.C.Zhu, Y.N.Wu, D.Mumford, *Filters, Random fields And Maximum Entropy (FRAME)*, Intl Journal of Computer Vision 27(2) 1-20, March/April 1998.