ELSEVIER

June 1994

Pattern Recognition Letters 15 (1994) 567–573

Pattern Recognition Letters

Entropic feature for sequence pattern through iterated function systems *

R. Román-Roldán [a,**], P. Bernaola-Galván [a], J.L. Oliver [b]

[a] *Departamento de Física Aplicada, Universidad de Granada, Spain*
[b] *Departamento de Genética, Universidad de Granada, Spain*

Received 29 June 1992

Abstract

Entropy and relative entropy are proposed as features extracted from symbol sequences. Firstly, a proper Iterated Function System is driven by the sequence, producing a fractal-like representation (CSR) with a low computational cost. Then, two entropic measures are applied to the CSR histogram of the CSR and theoretically justified. Examples are included.

Key words: Entropy; Histogram; Iterated function system; Chaos game; Pattern recognition; Feature extraction; Randomness; Sequence

1. Introduction

Feature extraction is an important topic in pattern recognition. The subject is covered in the literature before, namely Kittler (1986, p. 60). Recognition of patterns or structures in sequences (Valiveti and Oommen, 1991) is a particular subject which the present work deals with. In this context, a sequence is understood as any string of symbols or data drawn from a finite alphabet. They can, of course, come from any other figure, such as a digital image, by a properly defined ordering method of scanning and reading.

We propose a new feature to be extracted from sequences. It is intended for applying either to any of a battery of classification procedures in use, together with any other features, or in a characterization process such as revealing non-randomness in a sequence.

The novelty of the proposed entropic feature is twofold: the procedure for scrutinizing the sequence, through the Chaos Sequence Representation (CSR) and the application of Shannon's entropy and Kullback's relative entropy to the histogram of the above CSR (see for example Cover and Thomas, 1991).

A chaotic representation for sequences is presented, which is based on the so-called *Iterated Function System* (IFS) by Barnsley (1988, p. 82). Jeffrey (1990) and Oliver et al. (1992) have applied a similar method in analysing DNA sequences. The attractor obtained, here called the *Chaos Sequence Representation* (CSR), is suitable for processing. The procedure has proved to be a saving in computational effort when determining statistics concerning large subsequences in the whole string. The subject is described in Section 2.

The histogram of the CSR as well as its entropy and relative entropy are theoretically interpreted and discussed as a suitable feature. Gray-level histograms in digital images have been reported before by Román-Roldán et al. (1991) in a similar fashion. The spatial multi-resolution analysis provided there is translated

* This work was partially supported by a grant from the DGICIT (TIC-0646) of the Spanish Government to R. Román.

** Corresponding author. Email: rroman@ugr.es.

SSDI 0167-8655(93)E0063-T

into multi-length substring analysis here. Section 3 deals with these subjects.

The entropic feature is introduced in Section 4 as a statistical one. Since an entropic measure is obtained for each substring length from, for example, 1 to M, the result is a feature vector or, graphically, an entropic profile. The underlying meaning of these features is given, particularly when a fully random sequence is used as a reference. In such a case, this results in a useful criterion for non-randomness.

Section 5 presents the algorithm used for the computations. Section 6 offers three kinds of applications as examples: (1) sequences produced by random number generators, which are intended for full randomness; (2) DNA chains, where nature has imprinted meaning; (3) digital images.

2. The chaos representation of sequences

2.1. Iterated function system

The basic subject of Chaos Theory to be applied in this work, used similarly by Jeffrey (1990) and Oliver et al. (1992), is described by Barnsley (1988, p. 82) as the *Iterated Function System* (IFS):

$$\{(X, d); w_n, n=1, 2, \dots, N\},$$

where (X, d) is a metric space and w_n is the nth contraction mapping of the form

$$w : X \to X, \text{ such that } \forall x, y \in X,$$

$$d(w(x), w(y)) \leqslant s \cdot d(x, y),$$

where s is the contractivity factor, $0 \leqslant s < 1$.

For simplicity, attention is usually restricted to IFSs of the form $\{\mathbb{R}^M; w_n, n=1, 2, \dots, N\}$, where each mapping is an affine transformation. Most often M equals 2, since it allows for a simple visual representation. The IFSs can be represented as follows:

$$w_i \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e_i \\ f_i \end{bmatrix}.$$

For an isolated contraction mapping, the attractor would be a *fixed point*, according to the *contraction mapping theorem* (Barnsley, 1988, p. 76). However, the attractor for the IFS depends upon the ordering by which the functions w_i are applied. In a random algorithm, the rule is given through a probability distribution on the set of functions w. This, together with the coefficients of $\{w_i\}$, constitute the *IFS code*; for the Sierpinsky triangle with vertices (e_i, f_i), the code is given in Table 1.

2.2. The chaos game

For our present interest, the following choices are selected.

(A) All contractivity factors (a, b, c, d in the above example) are set at 0.5.

(B) The number of functions is constrained to the number of vertices of a hypercube drawn in the space $X=\mathbb{R}^M$; $N=2^M$.

(C) The constants e_i, f_i are taken as the co-ordinates (x, y) of the ith vertex. Under these conditions, attractors are obtained without any structure due to the IFS itself. The fractal structure, if any, will come from the probabilities p_i.

(D) Lastly a sequence of symbols, drawn from a finite alphabet, is used to determine the function w_i running each time. Thus, the probability distribution is substituted by the frequency vector (called *type* by Cover (1991, p. 279)) of the sequence.

A simple procedure to go from a sequence to the corresponding attractor may be described in the following way (restricting ourselves to $N=4$ and $M=2$ without any loss of generality, thus obtaining a picture on a piece of paper).

(1) Locate four dots (vertices) anywhere on the paper (without any three of them in a line).

(2) Label them from 1 to 4.

(3) Pick a point anywhere, which will be the initial point.

(4) Randomly generate symbol-labels (from the alphabet $\{1, 2, 3, 4\}$).

(5) Mark a new point half way between the previous one and the newly indicated vertex.

The dancing point makes a fractal picture which

Table 1
IFS code for the Sierpinsky triangle

w	a	b	c	d	e	f	p
1	0.5	0	0	0.5	1	1	1/3
2	0.5	0	0	0.5	1	50	1/3
3	0.5	0	0	0.5	50	50	1/3

has been called *Chaos Game* by Barnsley (1988, p. 2), thus revealing some underlying structure in the sequence.

2.3. The chaos sequence representation (CSR)

Each and every symbol in the sequence corresponds to one point in the CSR. Four important features of the diagrams should be emphasized:

(1) When drawing the CSR, point by point, each can be allocated by knowing the corresponding symbol and the previous point only.

(2) The location of each point in the CSR allows us to retrace the whole sequence up to it; thus, the final point indicates the total sequence.

(3) By partitioning the CSR into 4^l small l-squares, these correspond one to one with the different l-lengthened subsequences. Varying l we get a multi-length subsequence statistical description.

(4) The density of points (number of points in a given l-square) is the frequency of the corresponding l-subsequence in the chain; thus, the relative densities of points correspond to probabilities of finding subsequences.

As an example, Fig. 1 shows a CSR for (a) the DNA chain of Human Betaglobin Region Chromosome 11 (HUMHBB) with 73357 bases, and (b) a sequence generated by a congruential generator with the same number of symbols.

We take advantage of these properties for both theoretical analysis and practical implementation. Some tools for image processing could be applied to the CSR. For example, a method of multi-resolution-information analysis for digital images has recently been developed by Román-Roldán et al. (1991). First, the entropy of the gray-level histogram there is translated here into the entropy of the point-density histogram in CSRs; second, different resolutions of images are translated into different square sizes of CSRs, which in turn correspond to different subsequence lengths. In this case, a particular histogram may be used as a reference, thus obtaining some measure of divergence of the present CSR with respect to it.

3. The density distribution in the CSR

3.1. The type of the sequence

We shall use the geometric terminology (points, squares, densities, ...) for an easier and intuitive description. For each l-resolution, we have 4^l squares in the CSR. Let the series $N_1, N_2, ..., N_i, ..., N_L$, $\sum_i N_i = N$, be the number of points in the different squares. It is the CSR itself, plus a rather arbitrary ordering of

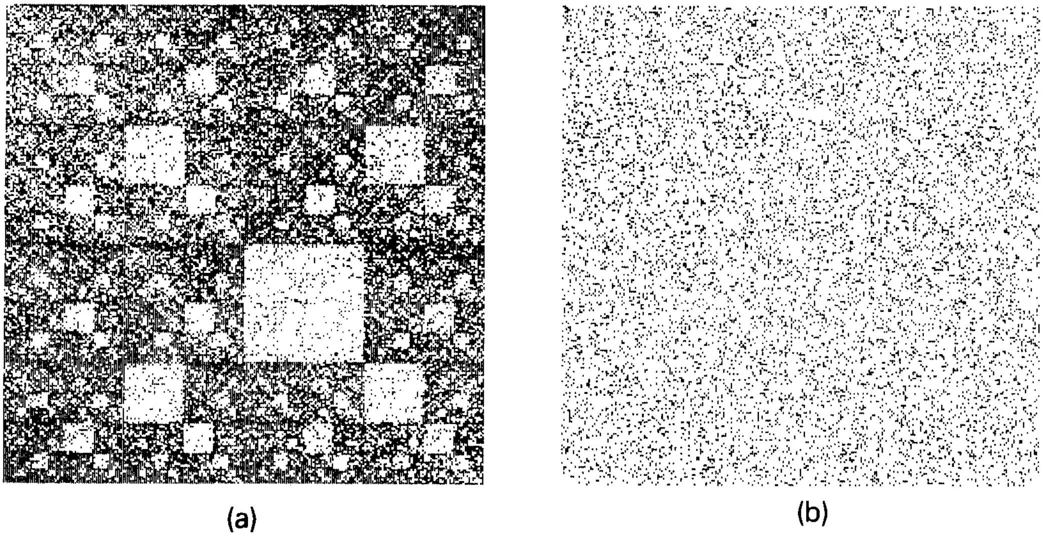

Fig. 1. CSR for (a) a DNA chain; (b) a random sequence computer generated.

the cells. By normalizing this series, we obtain the set of relative frequencies or densities (the type) $\mathscr{Q}=\{q_1, q_2, ..., q_i, ..., q_L\}$, which is also called the *frequency vector* by Mansuripur (1987).

3.2. The CSR histogram

We shall next consider the set $\{N_1, N_2..., N_i, ..., N_L\}$ as a new sequence. By scanning it and counting the number of occurrences of each N_i, and by normalization, we have a new type. We denote it by $\mathscr{Z}=\{z_0, z_1, ..., z_k, ..., z_N\}$, $z_k=n_k/4^l$, n_k being the number of l-squares with a given number of points k, so that z_k measures the probability of finding a cell with n_k points when one is randomly picked. $\mathscr{Z}$ stands for the histogram of densities in the CSR, the *CSR-histogram* for short. We shall keep in mind that it is also the histogram of the l-lengthened subsequences that are contained in a overlapping way in the whole sequence.

We argue next in favor of the CSR-histogram, instead of the type $\mathscr{Q}$, from the point of view of extracting a feature from the sequence. If two sequences were to differ only in the interchange of any pair (N_i, N_j), both must in fact have the same feature, since such a difference is usually irrelevant. The same may be said if these were to interchange the pair (n_s, n_t), as long as the ordering of cells remains quite arbitrary. Of course, a relevant feature must be insensitive to these irrelevant differences. As will soon be seen, entropic measures of the CSR-histogram fulfill this requirement by the symmetry of such measures in respect to their arguments.

4. Entropic feature

4.1. Entropic measures

Shannon's entropy. For the resolution l, Shannon's entropies of the CSR ($\mathscr{Q}^l$) and the histogram $\mathscr{Z}^{(l)}$ are, respectively,

$$H(\mathscr{Q}^{(l)})=\sum_i q_i^{(l)} \log q_i^{(l)},$$

$$H(\mathscr{Z}^{(l)})=\sum_k z_k^{(l)} \log z_k^{(l)}.$$

Both entropies could be taken to be absolute features (no reference is involved) of the sequence. However, $H(\mathscr{Z})$ seems to be a better, more significant feature than $H(\mathscr{Q})$, since it defers irrelevant differences between sequences, as seen in the preceding paragraph. It means the uncertainty about the number of points in an l-square randomly chosen from the CSR. Alternatively, this means the uncertainty about the number of l-subsequences of a randomly chosen l-pattern in the sequence.

Kullback's divergence. It may happen that a certain histogram is expected under some hypotheses. It may also occur that a set of sequences is proposed as patterns in a classification problem. In both cases a feature is needed either to test the hypotheses or to measure how far the given sequence is from the expected one(s). The *relative entropy* (also called *directed divergence*) of the histogram is a good candidate to account for the deviation of the current sequence from a reference:

$$D(\mathscr{Z}\|\mathscr{R})=\sum_k z_k \log(z_k/r_k),$$

where the histogram of reference is $\mathscr{R}=\{r_0, r_1, ..., r_k, ..., r_N\}$, and the superindexes l have been dropped in the probabilities and histogram notations for simplicity.

4.2. Entropic feature for non-randomness

In spite of the several uses that a feature may be given, in this work we restrict ourselves to considering the entropic measures of the CSR-histograms as features of non-randomness. Specifically, the relative entropy should compare the histogram $\mathscr{Z}$ to a reference $\mathscr{R}$ which is given by the histogram expected from a fully random sequence. This subject is left for further development, although the result for $\mathscr{R}$ is briefly described next.

A fully random sequence is expected to be produced by a uniform i.i.d. source. All of these are equally probable, and therefore all possible subsequences must also be equally probable for each length l. As a result, the most probable histogram is not the degenerated one ($z_{k=4-l}=1$, $z_{k\neq 4-l}=0$), since many different CSRs have the same histogram, but rather the binomial one. The question may be posed in terms of the probabilities of allocating one point after the other in the cells of the diagram, and then asking for the probability of a given number of points appearing in any l-square. The probability of a point falling in a given l-square is 4^{-l}, and $(1-4^{-l})$ out of it. This leads to the binomial distribution

$$P(N_{j,l})=\binom{N}{N_j}(4^{-l})^{N_j}(1-4^{-l})^{(N-N_j)},$$

where $P(N_{j,l})$ is the probability of having N_j points in an l-square from the total number N. Thus, the expected sequence histogram from a fully random source is the binomial one $\mathscr{B}(p,\mu)$ with elemental probability $p=4^{-l}$ and mean $\mu=N\cdot 4^{-l}$. Consequently, an appropriate entropic feature for non-randomness in sequences is the relative entropy with respect to the binomial distribution as reference, $D(\mathscr{X}\|\mathscr{B})$. It should be noticed that the condition (B) in Section 2.2 demands an alphabet with cardinal $N=2^M$. This is a constraint imposed upon the numbers generated by the source, which generally produces numbers in the interval $[0,1)$. A transformation has to be made, based on to the number of vertices selected. On the other hand, the procedure converts the number of intervals N into dimensions of the CSR, $M=\log N$, as well as subsequence length l in spatial resolution.

4.3. The underlying meaning of the entropic features

The entropic features may be theoretically grounded. For this, complete randomness (if it exists; see Knuth, 1981, p. 142) is supposed to be void of information at all. Conversely, a practical concept of information requires that it rely on the deviation from randomness. In this context, sequences can be thought of as messages emitted by generators, this information being a natural measure of the partial lack of randomness.

Let us notice that a uniform CSR has the maximum entropy $H(\mathscr{Q})$ at each resolution l, yet such a CSR is not informative at all in the exposed sense. However, the entropy of the histogram $H(\mathscr{X})$ goes to zero when $N\to\infty$. Thus, the entropy of the histogram is justified from the Information Theory point of view, as well as by heuristic considerations.

The usual understanding of the relative entropy as a measure of the gain of information is fully applicable here. It represents the variation of information when passing from a very random sequence to the present one, which is an appealing meaning indeed.

The same argument has been proposed (Román-Roldán et al., 1991) by looking for meaning, information conveying in images by means of the entropy of the histogram instead of the conventional entropy of the image (similar to $H(\mathrm{Q})$ for a sequence).

5. Examples

We present next some examples in which the entropic profile is extracted as a feature from certain kinds of sequences. In all three examples, both the entropy and the relative entropy is drawn versus the resolution level l. The binomial profile is included in the entropy representation, as a reference for full randomness. An algorithm for computations is straightforwardly derived from the IFS structure itself, as exposed in Sections 2.1. and 2.2. For all $l<l_{\max}$, the mean histogram has been considered from a number of CSRs in such a way that the same number of cells are computed, namely $4^{l_{\max}}$. For this, the total number of symbols in the sequence must be equally spread in $4^{l_{\max}-l}$ CSRs, resulting in the same mean (points per cell) for all l.

Fig. 2 shows the entropic profile extracted from sequences generated by three different random number generators. $N=4$ and $M=2$ have been used. Therefore, each random number belonging to the interval $[0,\ 1)$ has been classified in the subintervals $[0,\ 0.25)$, $[0.25, 0.5)$, and so on. The CSR has $(2^M)^l=4^l$ squares. A sequence of 122880 numbers has been analysed, allowing for $l=1$ to 6 with mean $\mu=30$ points per l-square. The A-carry generator looks much better than others, as its entropy is very close to the binomial distribution, which has been omitted for this reason.

Fig. 3 presents the results for four different DNA sequences. Since there are 4 constituent bases, $N=4$ and $M=2$ have been used again. The profiles show that the two human sequences (HUMFIXG and HUMAFP) clearly differ from both the bacterial (ECUNC) and the viral one (PT7CG) at all resolution levels but the third (trimere).

Fig. 4 shows the results for three binary images. The pixels have been read in an appropriate ordering so

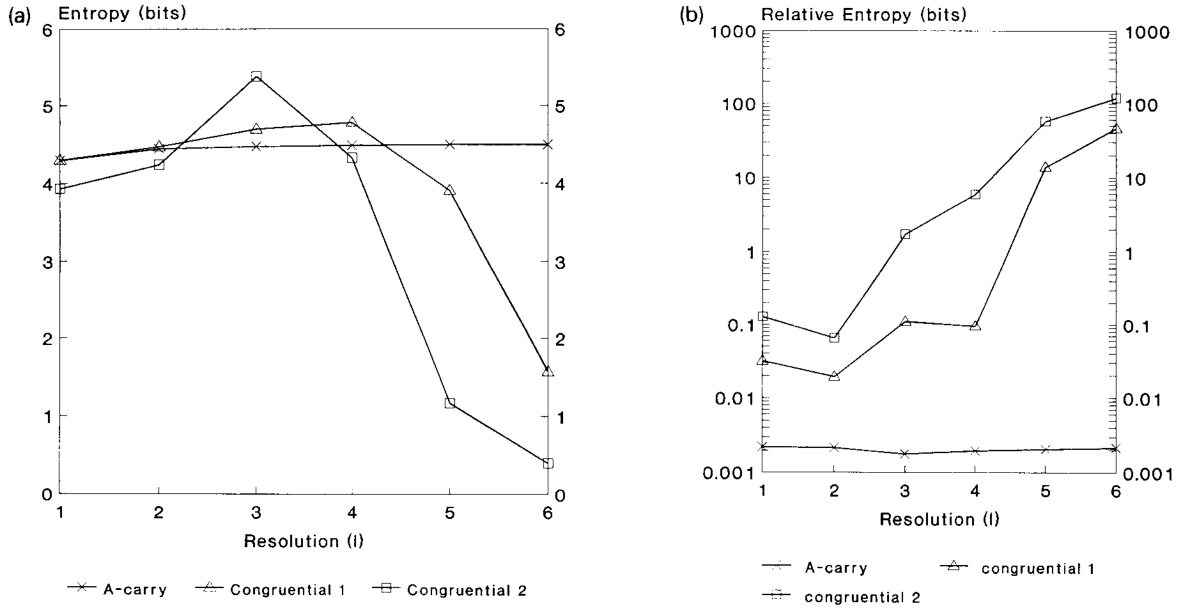

Fig. 2. Entropic profiles for three different random number generators.

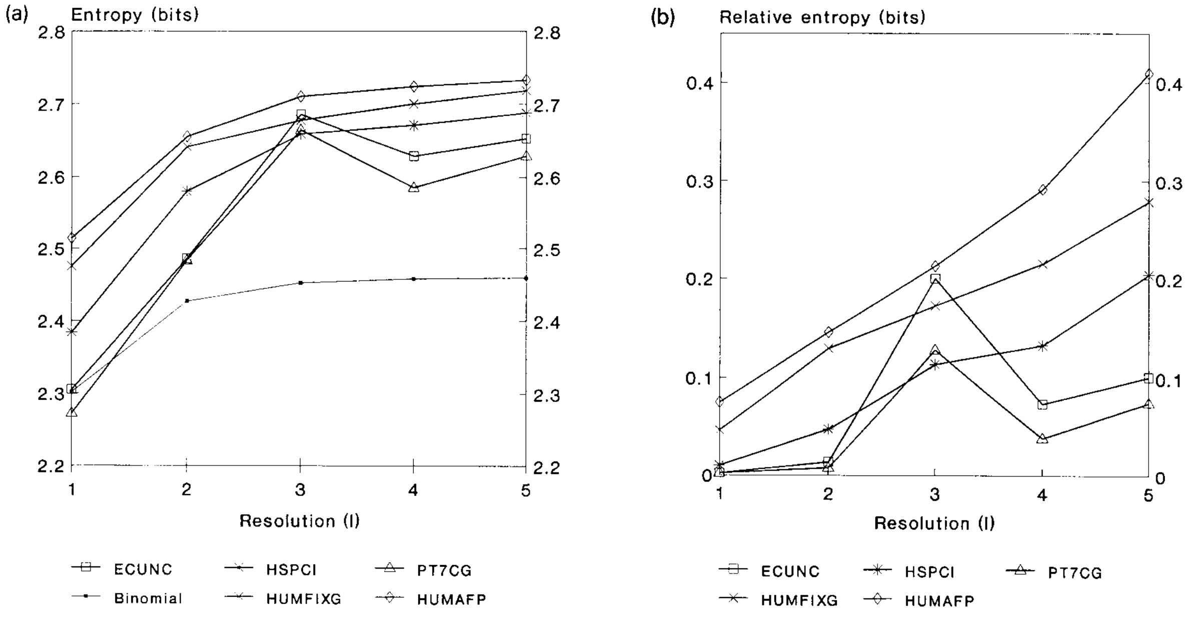

Fig. 3. Entropic profiles for four DNA chains.

that, for all spatial resolutions, black or white regions in the image lead to uniform substrings in the sequence. Because of the binary character of this kind of sequence, $N=2$ and $M=1$ have been used here, thus the CSR lies in the straight line.

Some general observations can be made about the entropic profiles. The entropy measure alone does not reveal how far the histogram is from the reference. For instance, the entropy of the image *Photo* 1 is about equal to the entropy of the binomial distribution for $l=7$, yet their relative entropies are quite different.

However, the entropy profiles supply complemen-

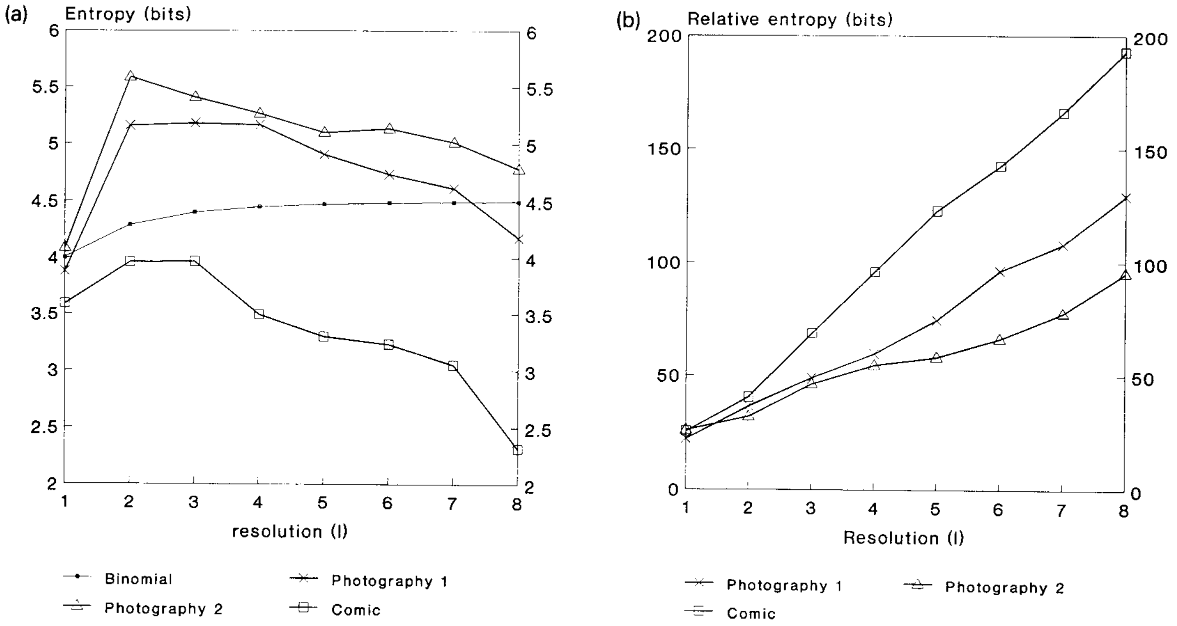

Fig. 4. Entropic profiles for three binary images.

tary information with respect to the relative entropy profiles. A binomial instead of a degenerated histogram accounts for the expected fluctuations from a fully random source, which would have zero entropy. Therefore, an entropy higher than the binomial suggests a $\mathscr{Q}$-type more grouped, thus some subsequences occur more frequently than others. Conversely, a lower entropy of the histogram corresponds to a $\mathscr{Q}$-type closer to a plain one than expected; this fact points to a Markovian dependence or periodic, deterministic sources.

6. Conclusions

The entropy and the relative entropy of the histogram of the Chaos Sequence Representation has been suggested as features extracted from sequences as a function of the subsequence length, so given place to entropic profiles.

The Chaos Sequence Representation, derived from the Iterated Function Systems, has proved to allow for low computational complexity in scanning the sequence for large subsequences.

The entropic features of the histogram exhibit high discrimination capability in representing patterns in sequences, while they are theoretically grounded through Information Theory.

A source of random sequences have been considered as a reference in an example for detecting and quantifying non-randomness.

References

Barnsley, M. (1988). *Fractals Everywhere.* Academic Press, London.

Cover, T.M. and J.A. Thomas (1991). *Elements of Information Theory.* Wiley, New York.

Jeffrey, H.J. (1990). Chaos game representation of gene structure. *Nucleic Acids Research* 18, 2163–2170.

Kittler, J. (1986). Feature selection and extraction. In: *Handbook of Pattern Recognition and Image Processing.* Academic Press, London.

Knuth, D.E. (1981). *The Art of Computer Programming.* Addison-Wesley, Reading, MA, 2nd ed.

Mansuripur, M. (1987). *Introduction to Information Theory,* Prentice-Hall, Englewood Cliffs, NJ.

Oliver, J.L., P. Bernaola-Galván, J. Guerrero-Garcia and R. Román-Roldán (1992). Entropic profiles of DNA sequences through chaos-game derived images. *J. Theoretical Biology,* to appear.

Román-Roldán, R., J.J. Quesada Molina and J. Martinez Aroza (1991). Multiresolution-information analysis for images. *Signal Processing* 24, 77–91.

Valiveti, R.S. and B.J. Oommen (1991). Recognizing sources of random strings. *IEEE Trans. Pattern Anal. Machine Intell.* 13, 386–394.