Object-background segmentation using new definitions of entropy

N.R. Pal
S.K. Pal

Indexing term: Algorithms

Abstract: The definition of Shannon's entropy in the context of information theory is critically examined and some of its applications to image processing problems are reviewed. A new definition of classical entropy based on the exponential behaviour of information-gain is proposed along with its justification. Its properties also include those of Shannon's entropy. The concept is then extended to fuzzy sets for defining a nonprobabilistic entropy and to grey tone image for defining its global, local and conditional entropy. Based on those definitions, three algorithms are developed for image segmentation. The superiority of these algorithms is experimentally demonstrated for a set of images having various types of histogram.

1 Introduction

The entropy of a system as defined by Shannon [1, 2] gives a measure of our ignorance about its actual structure. In the context of information theory, Shannon's function is based on the concept that information gain from an event is inversely related to its probability of occurrence. The logarithmic behaviour of entropy is considered to incorporate the additive property of information.

Several authors [3–5, 7–12] have used this concept to image processing and pattern recognition problems. Pun [3, 4] used Shannon's concept to define the entropy of an image assuming that an image is entirely represented by its grey level histogram only. Finally, he used this entropic measure for image segmentation into object and background. Kapur *et al.* [5] recently have also used the similar concept for image segmentation. They, instead of considering one probability distribution for the entire histogram, used two separate probability distributions for the object and the background. The total entropy of the image is then maximised to arrive at the threshold for segmentation.

Deluca and Termini [6] defined a nonprobabilistic entropy of a fuzzy set which is also based on the concept of Shannon's function. Instead of the probability function, the membership function is used here to give a measure of fuzziness (ambiguity) in a set.

Paper 6577E (E4), received 6th January 1988

N.R. Pal is with the Computer Science Unit, Indian Statistical Institute, Calcutta 700 035, India

S.K. Pal is with the Electronics and Communication Sciences Unit, Indian Statistical Institute, Calcutta 700 035, India

Pal and others [7–12] have implemented this nonprobabilistic entropy to formulate algorithms for image enhancement, thresholding, clustering, edge ambiguity and other information measures. They also defined intra and interset ambiguity (fuzziness) measures [11] to determine an index for feature evaluation problem.

It is to be mentioned here that the definition of Shannon's entropy which was formulated in the context of information theory was used by the above mentioned authors to image processing problems without highlighting the suitability of its concept in the case of a grey tone image.

The segmentation algorithms [3–5] using Shannon's function resulted in an unappealing result, namely, same entropy and threshold values for different images with identical histogram. Furthermore, in Reference 3 the maximisation of the upper bound of the *a-posteriori* entropy for threshold selection is not justified.

The present work attempts to introduce a new concept of entropy along with its applications. First of all, a new definition of classical entropy is proposed along with its justification. Unlike the logarithmic behaviour of Shannon's entropy, the gain function considered here is of exponential nature so that the gain in information from an event i with probability of occurrence p_i is defined at all points with bounds at both ends. All other properties except the additive property for independent event (which does not carry any extra weight for an image, as pixel intensities are normally dependent on each other) of Shannon's entropy are also proved,

In the second part, an extention is made to fuzzy sets for defining a nonprobabilistic entropy. This nonprobabilistic entropy is found to satisfy all the desired properties stated by De Luca and Termini [6] and Pal [12].

Based on the new concept, three definitions (e.g. global, local and conditional) of entropy of an image are then introduced. As an application of these definitions, three algorithms are developed for image segmentation.

The algorithms are finally implemented on a set of images with widely different types of histogram. Superiority of the proposed methods is established by comparing the results with those of Pun [3] and Kapur *et al.* [5].

2 Shannon's entropy

Shannon [1–2] defined the entropy of an n-state system as

$$H = -\sum_i p_i \log_2 p_i, \quad i = 1, 2, \ldots, n \tag{1}$$

where p_i is the probability of occurrence of the event i

and

$$\sum_i p_i = 1, \quad 0 \leqslant p_i \leqslant 1$$

In case of a binary system, the entropy becomes $H = -p \log_2 p - (1-p)\log_2(1-p)$.

The entropy H is claimed to express a measure of ignorance about the actual structure of the system. In order to explain why such an expression is taken as a measure of ignorance, let us critically examine the philosophy behind Shannon's entropic measure with an example given below.

Suppose a six-faced die, covered with a box, is placed on a table and someone is asked to guess the number on the top most face of the die. Since the exact state of the die is not known, he/she can describe the state of the die by the probability distribution of occurrences of different faces on the top. In otherwords, the state of the die can be expressed by specifying p_i, $i = 1, 2, \ldots, 6$; where p_i is the probability that the ith face is the topmost face. Obviously,

$$0 \leqslant p_i \leqslant 1 \quad \text{and} \quad \sum_{i=1}^{6} p_i = 1$$

When the box is opened, the state of the die becomes known to us and we gain some information. A very natural question arises, 'How much information did we gain?'

Let $p_k = \max_i \{p_i\}$: the most probable event and $p_m = \min_i \{p_i\}$: the least probable event. Now, if the kth face appears on the top, the gain in information would be minimum, whereas the occurrence of the mth face on the top would result in the maximum gain.

Thus we see that the gain in information from an event is inversely related to its probability of occurrence. This, of course, intuitively seems all right. For example, if somebody says, 'The sun rises in the east', the information content of the statement is practically nil. On the other hand if one says, 'He is ten feet in height', the information content of the statement is very high, as it is an unlikely event. A commonly used measure of such a gain is

$$\Delta I = \log_2(1/p_i) = -\log_2(p_i) \tag{2}$$

In order to justify the logarithmic function, the following points can be stated:

(*a*) It gives additive property of information. To make it more clear, suppose two independent events m and n with probabilities of occurrence p_m and p_n have occurred jointly, then the additive property says

$$\Delta I(p_m \cdot p_n) = \Delta I(p_m) + \Delta I(p_n) \tag{3}$$

where $(p_m \cdot p_n)$ is the probability of the joint occurrence of the events m and n. Thus the additive property can be stated as follows. The information gain from the joint occurrence of more than one event is equal to the sum of information gain from their individual occurrence.

(*b*) The gain in information from an absolutely certain event is zero, i.e., $\Delta I(p_i = 1) = 0$.

(*c*) As p_i increases, $\Delta I(p_i)$ decreases.

Referring back to our experiment of die, the expected gain in information from the experiment can be written as

$$H = E(\Delta I) = -\sum_{i=1}^{6} p_i \log_2 p_i.$$

The value of H denotes the entropy (Shannon's entropy) of the system. Thus for an n-state source, the entropy may be defined as in eqn. 1.

3 Entropic measures for image processing

Based on the concept of Shannon's entropy, different authors have defined entropy for an image and its extension to fuzzy sets. Let us discuss here those measures and the associated problems when applied to image processing and recognition problems.

Let $F = [f(x, y)]_{P \times Q}$ be an image of size $P \times Q$, where $f(x, y)$ is the grey value at (x, y); $f(x, y) \in G_L = \{0, 1, \ldots, L-1\}$, the set of grey levels. Let N_i be the frequency of the grey level i. Then

$$\sum_{i=0}^{L-1} N_i = P \times Q = N(\text{say}).$$

Pun [3–4] and Kapur *et al.* [5] considered the grey level histogram of F an L-symbol source, independently from the underlying image. In addition to this, they also assumed that these symbols are statistically independent.

Following Shannon's definition of entropy (eqn. 1), Pun [3] defined the entropy of the image (histogram) as

$$H = -\sum_{i=0}^{L-1} p_i \log_2 p_i \tag{4}$$

for an image segmentation problem.

3.1 *Evaluation function of Pun* [3]

Let s be the threshold which classifies the image into object and background. Let N_B and N_W be the number of pixels in the black and white portions of the image. Then the *a-posteriori* probability of a black pixel is $P_B = N_B/N$ and that of a white pixel is $P_W = N_W/N$. Thus, the *a-posteriori* entropy of the image is

$$\begin{aligned} H'_L(s) &= -P_B \log_2 P_B - P_W \log_2 P_W \\ &= -P_S \log_2 P_S - (1 - P_S)\log_2(1 - P_S) \end{aligned} \tag{5}$$

as

$$P_S = \sum_{i=0}^{s} p_i = P_B \quad \text{and} \quad P_W = 1 - P_S \tag{6}$$

Since the maximisation of H'_L gives the trivial result of $P_S = 1/2$, Pun [3] maximised an upper bound $g(s)$ of $H'_L(s)$, where

$$\begin{aligned} g(s) = &\frac{H_B^s \log_2 P_S}{H_L \log_2[\max(p_0, p_1, \ldots, p_s)]} \\ &+ \frac{(H_L - H_B^s)\log_2(1 - P_s)}{H_L \log_2[\max(p_{s+1}, p_{s+2}, \ldots, p_{L-1})]} \end{aligned} \tag{7}$$

where

$$H_L = -\sum_{i=0}^{L-1} p_i \log_2 p_i,$$

and

$$H_B^s = -\sum_{i=0}^{s} p_i \log_2 p_i.$$

The value of s which maximises $g(s)$ can be taken as the threshold for object and background classification.

3.2 *Method of Kapur, Sahoo and Wong* [5]

Recently, Kapur *et al.* have also used Shannon's concept of entropy but from a different point of view. They, instead of considering one probability distribution of the

entire image, considered two probability distributions; one for the object and the other for the background. The sum of the individual entropy of the object and background is then maximised.

If s is an assumed threshold, then the probability distribution of the grey levels over the black portion of the image is

$$\frac{p_0}{P_S}, \frac{p_1}{P_S}, \ldots, \frac{p_s}{P_S}$$

and that of the white portion is

$$\frac{p_{s+1}}{1-P_S}, \frac{p_{s+2}}{1-P_S}, \ldots, \frac{p_{L-1}}{1-P_S}$$

The entropy of the black portion (object) of the image is

$$H_B^{(s)} = -\sum_{i=0}^{s} \frac{p_i}{P_S} \log_2(p_i/P_S) \tag{8}$$

and that of the white portion is

$$H_W^{(s)} = -\sum_{i=s+1}^{L-1} \frac{p_i}{1-P_S} \log_2(p_i/(1-P_S)) \tag{9}$$

The total entropy of the image is then defined as

$$H_T^{(s)} = H_B^{(s)} + H_W^{(s)} \tag{10}$$

In order to select the threshold they maximised $H_T^{(s)}$. In other words, the value of s which maximises $H_T^{(s)}$ gives the threshold for object and background classification.

3.3 Entropy of fuzzy sets

The entropy so far we have considered is related only to the classical sets. There is another kind of entropy defined for a fuzzy set [6]. A fuzzy set A with its finite number of supports $x_1, x_2, \ldots, x_n$ in the universe of discourse U is formally defined as

$$A = \{(\mu_A(x_i), x_i)\}, \quad i = 1, 2, \ldots, n \tag{11}$$

where $\mu_A(x_i)$ is called the membership function of x_i with $0 \leqslant \mu_A(x_i) \leqslant 1$.

De Luca and Termini [6] defined the entropy of a fuzzy set A as

$$H(A) = K \sum_i S_n(\mu_A(x_i)), \quad i = 1, 2, \ldots, n \tag{12}$$

where S_n is Shannon's function having the form

$$S_n(x) = -x \log_2 x - (1-x) \log_2(1-x) \tag{13}$$

and K is a normalising constant.

The entropy $H(A)$ has the following properties:

$P1$: $H(A)$ is minimum if, and only if, $\mu_A(x_i) = 0$ or 1 for all i.

$P2$: $H(A)$ is maximum if, and only if, $\mu_A(x_i) = 0.5$ for all i.

$P3$: $H(A) \geqslant H(A^*)$, where A^* is any sharpened version of A.

A sharpened version of A is defined as

$$\mu_{A^*}(x_i) \geqslant \mu_A(x_i) \quad \text{if} \quad \mu_A(x_i) \geqslant 0.5$$

and

$$\mu_{A^*}(x_i) \leqslant \mu_A(x_i) \quad \text{if} \quad \mu_A(x_i) \leqslant 0.5$$

$P4$: $H(A) = H(\bar{A})$ with $\bar{A}$ = complement set of A.

It is very easy to see that with proper choice of K properties $P1$ to $P4$ are satisfied by $H(A)$ of eqn. 12.

$H(A)$ is thus seen to use Shannon's function but its meaning is quite different from classical entropy (eqn. 1), because no probabilistic concept is needed to define it. $H(A)$ provides the degree of fuzziness which expresses, on a global level, the average amount of difficulty (or ambiguity) in deciding whether an element would be considered to be a member of A or not.

Pal and others [7–12] have used this concept for image enhancement, fuzzy thresholding, edge ambiguity measure, feature selection and other information measures by optimising $H(A)$, with respect to S and π membership functions.

3.4 Some remarks

All the methods [3–5] discussed so far virtually assume that an image is entirely represented only by its histogram. Thus, different images with identical histograms will result in same entropic value in spite of their different spatial distributions of grey levels. This is, of course, not intuitively appealing. For example, consider Fig. 1 and Fig. 2. Both of Fig. 1 and Fig. 2 have identical histograms but different spatial distributions of grey levels. As a result, the entropy (information content) of Fig. 1 and Fig. 2 are expected to be different.

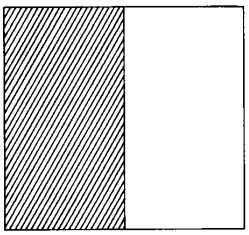


Fig. 1 *Two tone image*

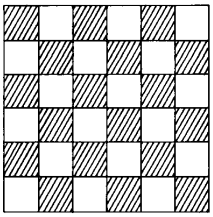


Fig. 2 *Two tone image*
Histogram identical to that of Fig. 1 but different spatial distribution

Under those definitions all images with identical histograms but different spatial distributions of grey levels will therefore give rise to same threshold value. Our experience and intuition also do not support this.

In the algorithm of Pun [3], the concept of maximisation of the upper bound of the evaluation function $g(s)$ (eqn. 7) for object background classification is not justified. For example, the maximum value of eqn. 7 may even correspond to a minimum value of the *a-posteriori* entropy (eqn. 5).

Moreover, all these methods have used Shannon's concept of entropy in image processing without highlighting its adequateness in the case of an image.

4 New definition of entropy

4.1 Justification

Before introducing the new definition of entropy the following points are in order.

(*a*) It is to be noted from the logarithmic entropic measure that as $p_i \to 0$, $\Delta I(p_i) \to \infty$ but $\Delta I(p_i = 1) = 0$

and $\Delta I(p_i) = -\log_2(p_i)$ is not defined for $p_i = 0$. Thus we see that information gain from an event is neither bounded at both ends nor defined at all points. In practice, the gain in information from an event, whether highly probable or highly unlikely, is expected to lie between two finite limits. For example, as more and more pixels in an image are analysed, the gain in information increases and when all the pixels are inspected the gain attains its maximum value, irrespective of the content of the image.

(*b*) The additive property for independent events does not carry any extra weight for an image, as pixel intensities in an image are normally dependent on each other.

(*c*) In Shannon's theory the measure of ignorance or the gain in information is taken as $\log_2(1/p_i)$ i.e., ignorance is inversely related to p_i. But mathematically, a more sound expression is possible to arrive at. If u_i is the uncertainty of the ith event then using the knowledge of probability one can write that $u_i = 1 - p_i$. Since u_i is the unlikeliness (i.e., probability of nonoccurrence), statistically ignorance can be better represented by $(1 - p_i)$ than $1/p_i$.

Now if we define the gain in information corresponding to the occurrence of the ith event as

$$\Delta I(p_i) = \log (1 - p_i)$$

then $\Delta I \leqslant 0$ which is intuitively unappealing. Furthermore, taking $-\log (1 - p_i)$ as gain in information leads to the fact that $\Delta I(p_i)$ increases with p_i; this is again not desirable.

The above problem can be circumvented by considering exponential function of $(1 - p_i)$ instead of the logarithmic behaviour. This is also appropriate while considering the concept of information gain in an image.

For example, consider the Figs. 3*a–e*. Suppose the images have only two grey levels; one corresponding to the lines (black portion) and the other corresponding to the white portion. In the case of the first image we have analysed only few black pixels and from this image we cannot say firmly about the content of the image. At this stage we see that it can be either a curtain or the hair of a face or something else. From the image *b* we can say that it is not a curtain (i.e., some gain in knowledge) while, from image *c* one can realise that it is a face. The image *d* says that it is a face with the mouth. However, image *e* does not say anything more than what is described by inage *d*, though the number of black pixels (hence probability) has increased.

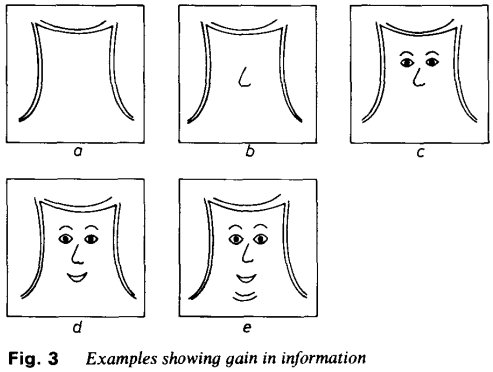


Fig. 3 *Examples showing gain in information*

Let $\Delta I(a)$, $\Delta I(b)$, $\Delta I(c)$, $\Delta I(d)$ and $\Delta I(e)$ be the information content of the images *a–e*, respectively. Now define the following quantities, representing change in gain:

$$\begin{aligned} G_1 &= \Delta I(b) - \Delta I(a) \\ G_2 &= \Delta I(c) - \Delta I(b) \\ G_3 &= \Delta I(d) - \Delta I(c) \\ G_4 &= \Delta I(e) - \Delta I(d). \end{aligned} \tag{14}$$

Obviously, $G_1 > G_2 > G_3 > G_4 \simeq 0$.

The above analysis and the fact that information gain approaches a finite limit when more and more pixels (increase in N_i and hence p_i) are analysed strengthen the assertion that the gain in information (i.e., increase in knowledge or decrease in ignorance) is exponential in nature.

4.2 Definition

The previously mentioned analysis led us to the following properties for the new entropic function.

*P*1: $\Delta I(p_i)$ is defined at all points in [0, 1].

*P*2: $\lim_{p_i \to 0} \Delta I(p_i) = \Delta I(p_i = 0) = k_1$, $k_1 \geqslant 0$ and finite.

*P*3: $\Delta I(p_i = 1) = k_2$, $k_2 \geqslant 0$ and finite

*P*4: $k_2 < k_1$

*P*5: With increase in p_i, $\Delta I(p_i)$ decreases exponentially. In other words, with increase in the uncertainty (u_i) the gain in information increases exponentially.

*P*6: $\Delta I(p)$ and H, the entropy, are continuous for $0 \leqslant p \leqslant 1$

*P*7: H is maximum when all p_is are equal.

Under the above framework let us define the gain in information from an event with probability p_i as

$$\Delta I(p_i) = e^{u_i} = e^{1 - p_i} \tag{15}$$

and the entropy H

$$H = E(\Delta I) = \sum_i p_i e^{1 - p_i} \tag{16}$$

It is easy to see that the properties *P*1 to *P*6 are satisfied where k_1 and k_2 take the values e and 1, respectively. The proof of *P*7 is given in Appendix 11.1.

In the case of a binary source, the entropy H takes the form

$$H = pe^{1-p} + (1 - p)e^{p}.$$

It is proved in Appendix 11.2 that H monotonically increases in [0, 0, 5] and monotonically decreases in [0.5, 1] with a maximum at $p = 0.5$.

4.3 Extension to fuzzy sets

Based on the aforesaid analysis, let us define a new expression for the entropy of a fuzzy set A as

$$H'(A) = \frac{1}{n} \sum_{i=1}^{n} [\mu_A(x_i)e^{(1 - \mu_A(x_i))} + \{1 - \mu_A(x_i)\}e^{\mu_A(x_i)}] \tag{17}$$

Like eqn. 12, $H'(A)$ also satisfies all the properties *P*1 to *P*4 of Section 3.3. Proofs are given in Appendix 11.3. Therefore, eqn. 17 can be regarded as a measure of fuzziness in a set which gives the average amount of difficulty (ambiguity) in deciding whether an element would be considered to be a member of a set A or not.

5 Entropy of an image

5.1 Global and local entropy

We know that in an image pixel intensities are not independent of each other. This dependency of pixel intensities can be incorporated by considering sequences of pixels to estimate the entropy. In order to arrive at the

expression of entropy of an image the following theorem can be stated based on the idea of Shannon [1, 2, 13].

Theorem: Let $p(s_i)$ be the probability of a sequence s_i of grey levels of length q. Let us define

$$H^{(q)} = \frac{1}{q} \sum_i p(s_i) e^{\{1 - p(s_i)\}} \tag{18}$$

where the summation is taken over all grey level sequences of length q. Then $H^{(q)}$ is a monotonic decreasing function of (q) and

$$\lim_{q \to \infty} H^{(q)} = H,$$

the entropy of the image.

For different values of q we get various orders of entropy.

Case 1: $q = 1$, i.e., sequence of length one. If $q = 1$ we get

$$H^{(1)} = \sum_{i=0}^{L-1} p_i e^{\{1 - p_i\}} \tag{19}$$

where p_i is the probability of occurrence of the grey level i.

Such an entropy is a function of the histogram only and it may be called the 'global entropy' of the image. Therefore, different images with identical histogram would have same $H^{(1)}$ value irrespective of their contents. The definitions used by Pun [3] and Kapur *et al.* [5], in fact, belong to Case 1.

Case 2: $q = 2$, i.e., sequences of length two. Hence,

$$H^{(2)} = \frac{1}{2} \sum_i p(s_i) e^{\{1 - p(s_i)\}},$$

where s_i is a sequence of grey level of length two.

$$= \frac{1}{2} \sum_i \sum_j p_{ij} e^{\{1 - p_{ij}\}} \tag{20}$$

where p_{ij} is the probability of co-occurrence of the grey levels i and j. Therefore, $H^{(2)}$ can be obtained from the co-occurrence matrix.

$H^{(2)}$ takes into account the spatial distribution of grey levels. Therefore, two images with identical histogram but different spatial distributions will result in different entropy, $H^{(2)}$ values. Expressions for higher order entropies $(q > 2)$ can also be deduced in a similar manner. $H^{(i)}$, $i \geqslant 2$ may be called 'local entropy' of order i of an image.

5.2 Conditional entropy

Suppose an image has two distinct portions, the object X and the background Y. Suppose the object consists of the grey levels $\{x_i\}$ and the background contains the grey levels $\{y_i\}$. The conditional entropy of the object X given the background Y i.e., the average amount of information that may be obtained from X given that one has viewed the background Y, can be defined as

$$H(X/Y) = \sum_{x_i \in X} \sum_{y_j \in Y} p(x_i/y_j) e^{\{1 - p(x_i/y_j)\}} \tag{21}$$

Similarly, the conditional entropy of the background Y given the object X is defined as

$$H(Y/X) = \sum_{y_j \in Y} \sum_{x_i \in X} p(y_j/x_i) e^{\{1 - p(y_j/x_i)\}} \tag{22}$$

The pixel y_j, in general, can be an mth order neighbour of the pixel x_i, i.e., y_j can be the mth pixel after x_i. Since the estimation of such a probability is very difficult, we impose another constraint on x_i and y_j of equations (21) and (22). In addition to $x_i \in X$ and $y_j \in Y$, we also impose the restriction that x_i and y_j must be adjacent pixels. Thus eqns. 21–22 can be rewritten as

$$H(X/Y) = \sum_{x_i \in X} \sum_{\substack{y_j \in Y \\ (x_i, y_j) \text{ adjacent}}} p(x_i/y_j) e^{\{1 - p(x_i/y_j)\}} \tag{23}$$

and

$$H(Y/X) = \sum_{y_j \in Y} \sum_{\substack{x_i \in X \\ (y_j, x_i) \text{ adjacent}}} p(y_j/x_i) e^{\{1 - p(y_j/x_i)\}} \tag{24}$$

The conditional entropy of the image can, therefore, be defined as

$$H^{(C)} = (H(X/Y) + H(Y/X))/2 \tag{25}$$

when X and Y represent object and background, respectively, of an image.

6 Application to image segmentation

Based on the new definitions of entropy of an image, the following three algorithms for object-background classification are proposed.

6.1 Algorithm 1

Following the concept of Kapur *et al.* and making use of eqn. 19 we can find an expression for thresholding as follows. If s is an assumed threshold then s partitions the image into object (black) and background (white). Using eqn. 19, the global entropy of the object or the black portion of the image can be defined as

$$H_B^{(G)}(s) = \sum_{i=0}^{s} \frac{p_i}{P_S} e^{\{1 - p_i/P_S\}} \tag{26}$$

where

$$P_S = \sum_{i=0}^{s} p_i$$

and

$$\sum_{i=0}^{L-1} p_i = 1$$

and the global entropy of the background or the white portion of the image as

$$H_W^{(G)}(s) = \sum_{i=s+1}^{L-1} \frac{p_i}{(1 - P_S)} e^{\{1 - p_i/(1 - P_s)\}} \tag{27}$$

Thus the total global entropy of the image can be defined as

$$H_T^{(G)}(s) = H_B^{(G)}(s) + H_W^{(G)}(s) \tag{28}$$

Let

$$\max_s \{H_T^{(G)}(s)\} = H_T^{(G)}(k), \quad 0 \leqslant k, s \leqslant L - 1.$$

Then the level k can be taken as a threshold for object-background classification of the image.

The threshold so obtained will classify the object and background in such a way that the sum of information in background and object is maximised, i.e., the resulting distribution of grey level in object and background would be uniform in the best possible way. However, like the entropic measures used by Pun [3] and Kapur *et al.* [5],

eqn. 28 is also a function of the grey level histogram of the image only. In other words, different images with identical histogram would result in same threshold level irrespective of the content of the image.

6.2 Algorithm 2

We are now going to describe another algorithm based on eqn. 20, which takes into account the spatial details of an image. Since such a method is dependent on the probability of co-occurrence of pixel intensities, let us define first of all the co-occurrence matrix before proceeding further.

Co-occurrence matrix: The co-occurrence matrix of the image F is an $L \times L$ dimensional matrix $T = [t_{ij}]_{L \times L}$ that gives an idea about the transition of intensities between adjacent pixels. In other words, t_{ij}, the (i, j)th entry of the matrix gives the number of times the grey level j follows the grey level i in some particular fashion.

The probability of co-occurrence p_{ij} of grey levels i and j can be written as

$$p_{ij} = t_{ij} \Big/ \left(\sum_i \sum_j t_{ij} \right);$$

obviously $0 \leqslant p_{ij} \leqslant 1$. If s, $0 \leqslant s \leqslant L-1$, is a threshold, then s partitions the co-occurrence matrix into four quadrants, namely A, B, C and D (Fig. 4).

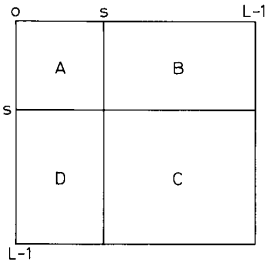


Fig. 4 *Quadrants of co-occurrence matrix*

Let us define the following quantities:

$$P_A = \sum_{i=0}^{s} \sum_{j=0}^{s} p_{ij},$$

$$P_B = \sum_{i=0}^{s} \sum_{j=s+1}^{L-1} p_{ij}, \tag{29}$$

$$P_C = \sum_{i=s+1}^{L-1} \sum_{j=s+1}^{L-1} p_{ij}$$

and

$$P_D = \sum_{i=s+1}^{L-1} \sum_{j=0}^{s} p_{ij}.$$

Normalising the probabilities within individual quadrant, such that the sum of the probabilities of each quadrant equals to one, we get the following cell probabilities for different quadrants.

$$p_{ij}^A = \frac{p_{ij}}{P_A} = \frac{t_{ij} \Big/ \left(\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} t_{ij} \right)}{\sum_{i=0}^{s} \sum_{j=0}^{s} \left\{ t_{ij} \Big/ \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} t_{ij} \right\}} = \frac{t_{ij}}{\sum_{i=0}^{s} \sum_{j=0}^{s} t_{ij}}$$

$$\text{for } 0 \leqslant i \leqslant s, \text{ and } 0 \leqslant j \leqslant s. \tag{30}$$

Similarly,

$$p_{ij}^B = \frac{p_{ij}}{P_B} = \frac{t_{ij}}{\sum_{i=0}^{s} \sum_{j=s+1}^{L-1} t_{ij}}$$

$$\text{for } 0 \leqslant i \leqslant s, \text{ and } s+1 \leqslant j \leqslant L-1 \tag{31}$$

$$p_{ij}^C = \frac{p_{ij}}{P_C} = \frac{t_{ij}}{\sum_{i=s+1}^{L-1} \sum_{j=s+1}^{L-1} t_{ij}} \quad \text{for } \begin{matrix} s+1 \leqslant i \leqslant L-1 \\ s+1 \leqslant j \leqslant L-1 \end{matrix} \tag{32}$$

and

$$p_{ij}^D = \frac{p_{ij}}{P_D} = \frac{t_{ij}}{\sum_{i=s+1}^{L-1} \sum_{j=0}^{s} t_{ij}} \quad \text{for } \begin{matrix} s+1 \leqslant i \leqslant L-1 \\ 0 \leqslant j \leqslant s \end{matrix} \tag{33}$$

Now with the help of eqns. 20 and 30, the second order local entropy of the object can be defined as

$$H_A^{(2)}(s) = \frac{1}{2} \sum_{i=0}^{s} \sum_{j=0}^{s} p_{ij}^A e^{(1 - p_{ij}^A)} \tag{34}$$

Similarly, the second order entropy of the background can be written as

$$H_C^{(2)}(s) = \frac{1}{2} \sum_{i=s+1}^{L-1} \sum_{j=s+1}^{L-1} p_{ij}^C e^{(1 - p_{ij}^C)} \tag{35}$$

Hence the total second order local entropy of the object and the background can be written as

$$H_T^{(2)}(s) = H_A^{(2)}(s) + H_C^{(2)}(s) \tag{36}$$

The grey level corresponding to the maximum of $H_T^{(2)}(s)$ gives the threshold for object-background classification.

6.3 Algorithm 3

This algorithm is based on the concept of conditional entropy (eqns. 23–25). Suppose s is an assumed threshold. Then pixels with grey level values ranging from 0 to s constitute the object while the remaining pixels with grey values lying between $s+1$ to $L-1$ correspond to the background. Let t_{ij} be an entry of the quadrant B (Fig. 4), then t_{ij} gives the number of transitions, such that i belongs to the object and j belongs to the background, and i and j are adjacent. Therefore, p_{ij}^B as defined in eqn. 31 gives the probability that grey level i and j belong to the object and background, respectively, and they are adjacent. Thus, $p_{ij}^B s$ of eqn. 31 give the probabilities required by eqn. 23. Similarly, $p_{ij}^D s$ of eqn. 33 correspond to the probabilities of eqn. 24.

Therefore

$$H(\text{object/background}) = H(O/B)$$

$$= \sum_{i=0}^{s} \sum_{j=s+1}^{L-1} p_{ij}^B e^{(1 - p_{ij}^B)} \tag{37}$$

and

$$H(\text{background/object}) = H(B/O)$$

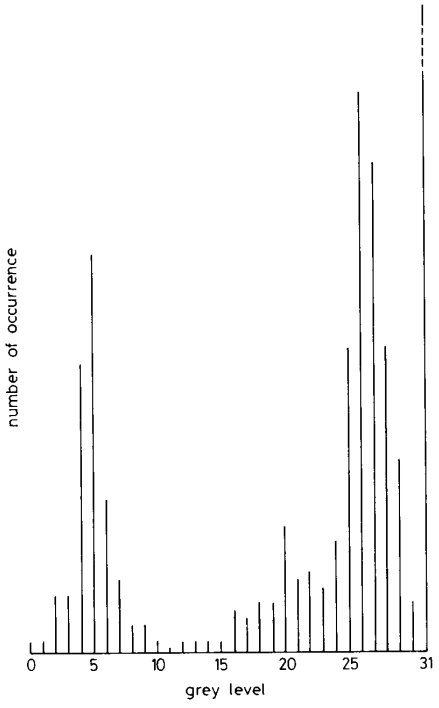
$$= \sum_{i=s+1}^{L-1} \sum_{j=0}^{s} p_{ij}^D e^{(1 - p_{ij}^D)} \tag{38}$$

Now the conditional entropy of the image is

$$H_T^{(C)} = (H(O/B) + H(B/O))/2 \tag{39}$$

In order to get the threshold for object-background classification $H_T^{(C)}$ is maximised with respect to s.

a



b

c

d and g

e

f

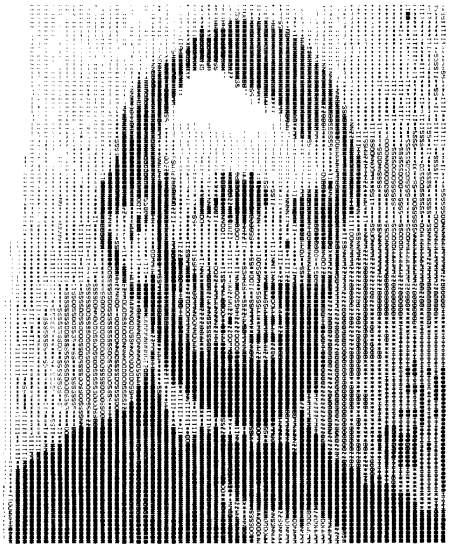
Fig. 5 *Biplane image*

a Input
b Histogram
c Proposed algorithm 1
d Proposed algorithm 2
e Proposed algorithm 3
f Algorithm of Pun
g Algorithm of Kapur *et al.*

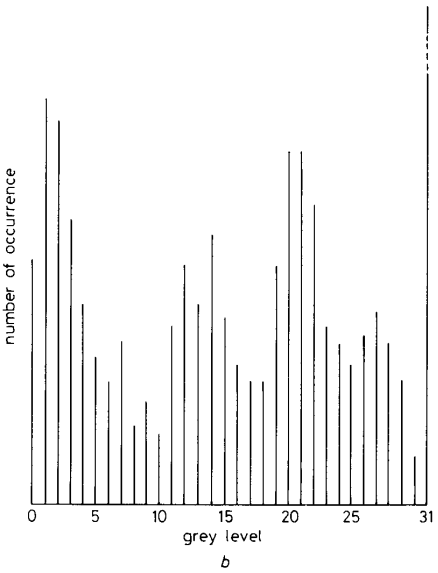
7 Implementation and results

The segmentation (object-background classification) algorithms described in Sections 3 and 6 are implemented on a set of four images with widely different types of histogram. Figs. 5*a*, 6*a*, 7*a* and 8*a* represent the input images while Figs. 5*b*, 6*b*, 7*b* and 8*b* represent the corresponding grey level histograms. The input images are produced on a line printer by over printing different character combinations for different grey levels. The threshold levels produced by different methods are presented in Table 1.

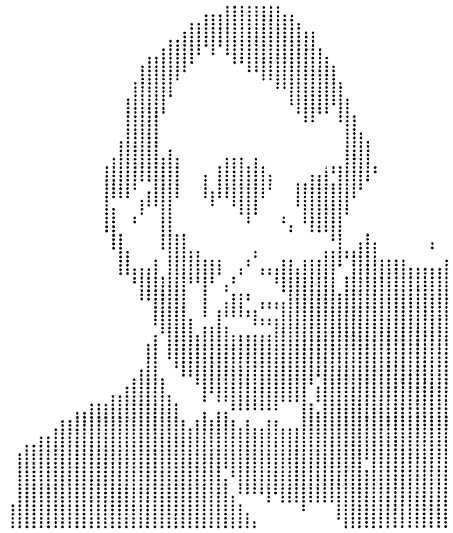
Fig. 5*a* represents the image of a biplane with two dominant modes in its grey level histogram (Fig. 5*b*). The segmented images produced by different methods are



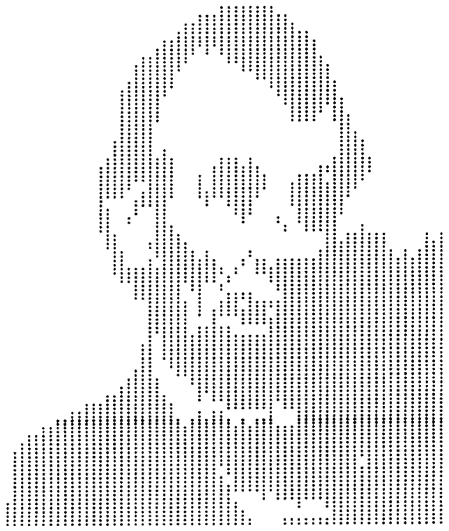
a



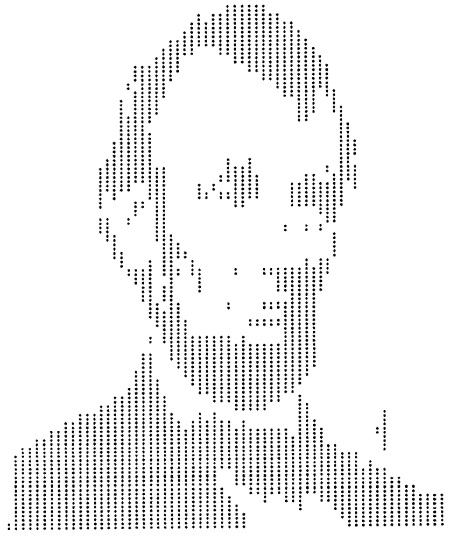
b



c and *g*



d and *f*



e

Fig. 6 *Lincoln image*

a Input
b Histogram
c Proposed algorithm 1
d Proposed algorithm 2
e Proposed algorithm 3
f Algorithm of Pun
g Algorithm of Kapur *et al.*

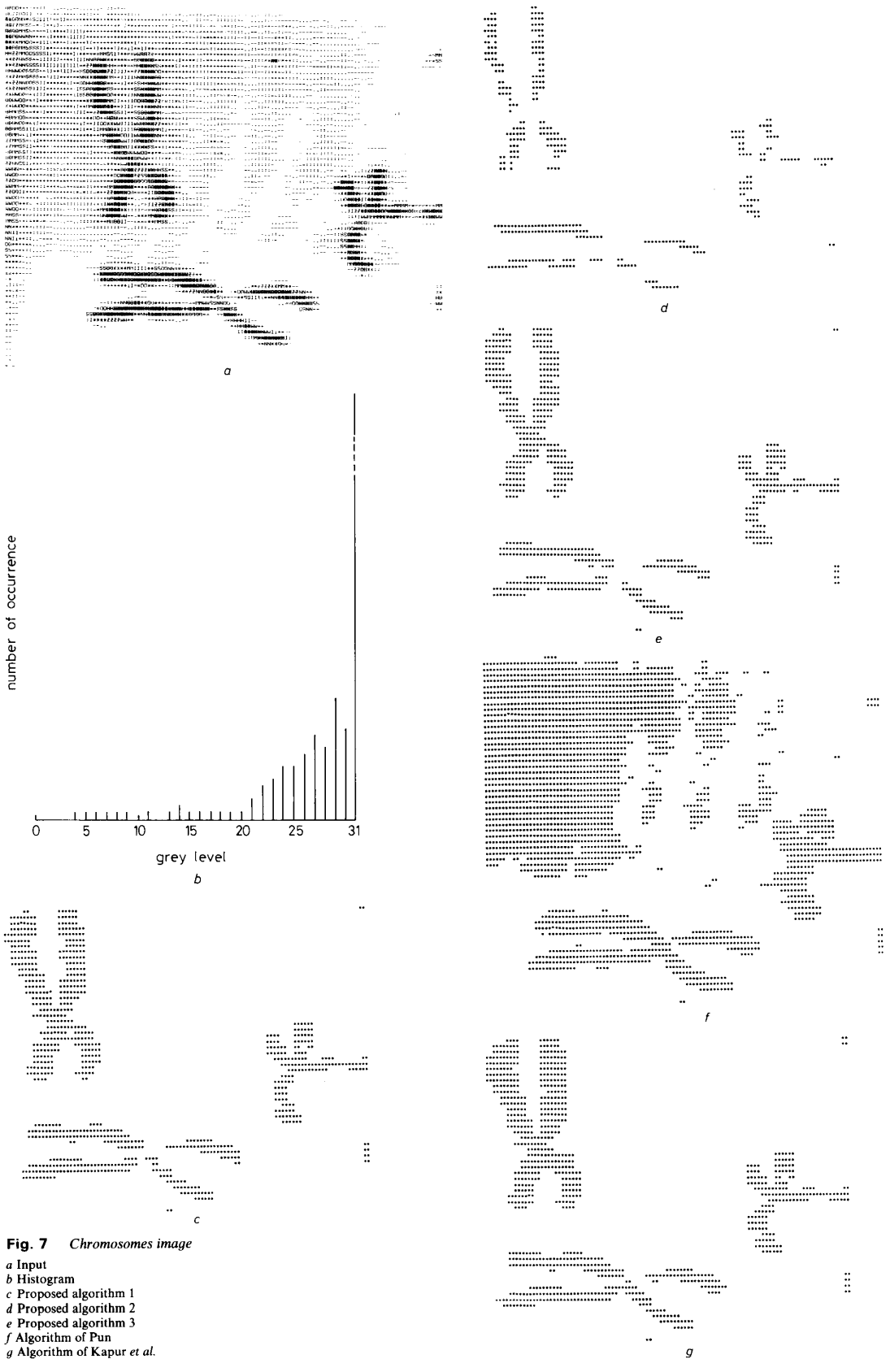


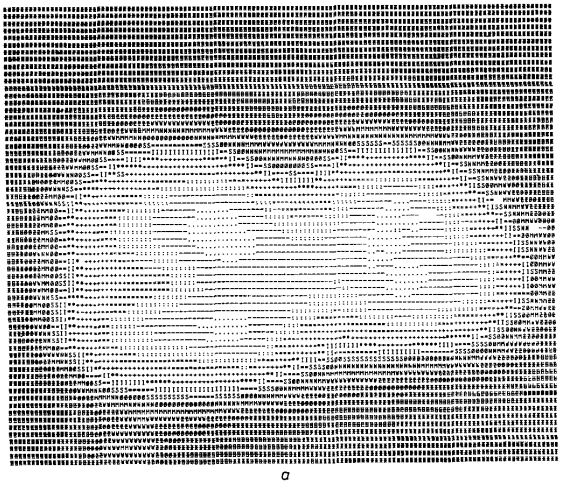
Fig. 7 *Chromosomes image*

a Input
b Histogram
c Proposed algorithm 1
d Proposed algorithm 2
e Proposed algorithm 3
f Algorithm of Pun
g Algorithm of Kapur *et al.*

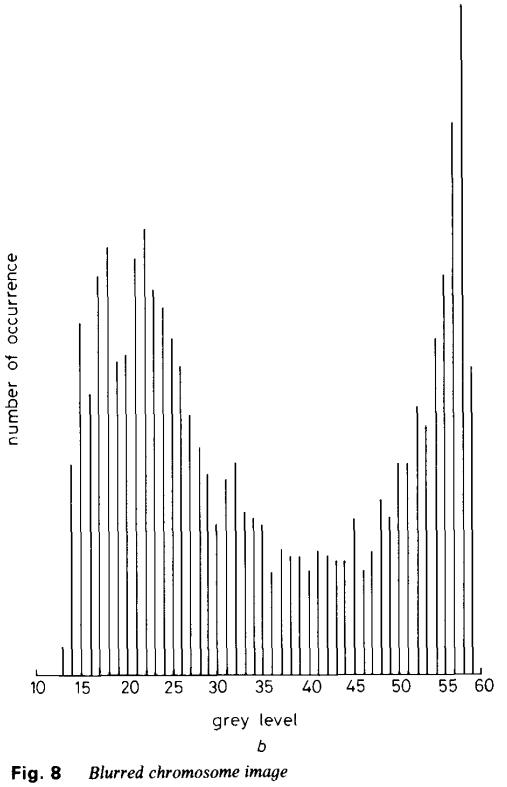
shown in Figs. 5*c–g*. From the results one can see that except for the conditional entropic method (eqn. 39), the propeller in front of the biplane is lost. In all but algorithm 3, some portion of the background became mixed up with the object, though the image has two dominant modes. The methods of Pun [3] and of Kapur [5] have produced comparable results to those of eqns. 28 and 36.

Figs. 6*a* and *b* represent the input image of Abraham Lincoln and its grey level histogram, respectively. The histogram has a number of deep valleys. The thresholds produced by different methods are shown in Table 1 and the corresponding segmented images are shown in Figs. 6*c–g*. In this case too, all the methods except the conditional entropic method (algorithm 3) have produced comparable result. The best result is produced by algorithm 3 (eqn. 39) which has clearly separated the object from the background. All other methods failed to discriminate between the beard and the background at the bottom left-hand corner of the image.

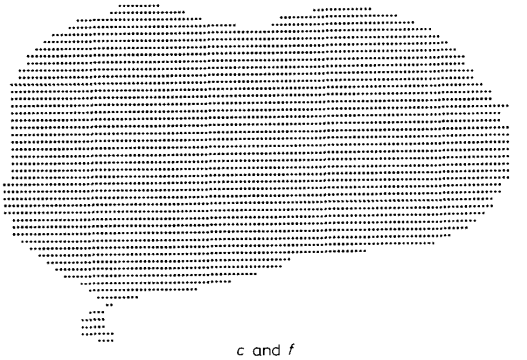
To demonstrate the effectiveness of the algorithms for



a

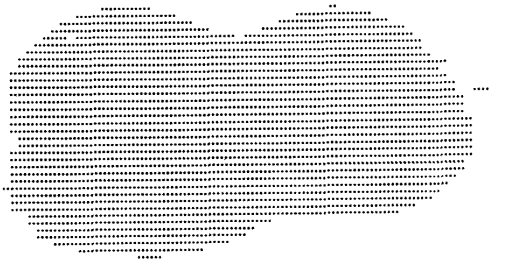


b

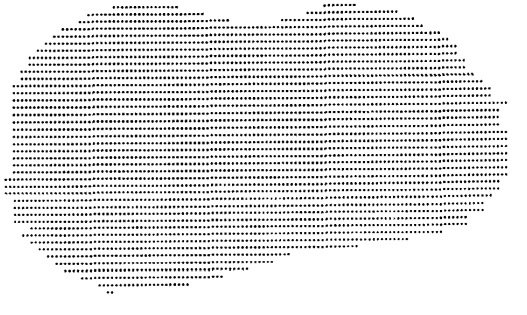


c and *f*

d



e



g

Fig. 8 *Blurred chromosome image*

a Input
b Histogram
c Proposed algorithm 1
d Proposed algorithm 2
e Proposed algorithm 3
f Algorithm of Pun
g Algorithm of Kapur *et al.*

Table 1. Thresholds for object-background classification

Images	Thresholds				
	Proposed algorithm 1 (eqn. 28)	Proposed algorithm 2 (eqn. 36)	Proposed algorithm 3 (eqn. 39)	Algorithm of Pun (eqn. 7)	Algorithm of Kapur et al. (eqn. 10)
Biplane (Fig. 5)	22	21	12	24	21
Lincoln (Fig. 6)	15	16	9	16	15
Chromosomes (Fig. 7)	19	10	17	27	20
Blurred chromosome (Fig. 8)	32	31	41	32	33

images with unimodal histogram, an image of a set of three chromosomes (Fig. 7*a*) has been considered. Fig. 7*b* depicts its grey level histogram. In this case we find that the proposed algorithms 1 (eqn. 28) and 3 (eqn. 39) give good results, while the second order local entropy (algorithm 2) gives rise to a thinned version of the chromosomes. The method of Kapur *et al.* (eqn. 10) is found to generate results comparable to that of algorithm 1 and 3. The worst result is produced by the method of Pun (eqn. 7) which could not extract one of the chromosomes at all.

The algorithms are also tested on an image of blurred chromosome (Fig. 8*a*) having a bimodal histogram (Fig. 8*b*). Here too, all the methods except the conditional entropic method (algorithm 3) have produced similar results. However, the best classification is done by Algorithm 3. This also conforms well to the recent work of Pal and Rosenfeld [14].

8 Conclusion

A new definition of probabilistic entropy based on the exponential behaviour of information-gain is proposed along with its justification. Its properties are also found to include those of Shannon's entropy. Based on this concept, various definitions of entropy (namely, global, local and conditional) for an image are introduced. The idea is also found to be extendable for defining non-probabilistic entropy of a fuzzy set.

Three algorithms for object-background classification (as an example of application of the new concept) are proposed whereby it is found to be able to segment/extract object from background. The results are compared with those of the existing entropic thresholding methods and are found to be superior for a wide class of images.

9 Acknowledgment

The authors thank Mr. J. Dasgupta for typing the manuscript, Mr. S.C. Nandy for his valuable discussion, and Prof. D. Dutta Majumder for his interest in this work.

10 References

1 SHANNON, C.E.: 'A mathematical theory of communication', *Bell Syst. Tech. J.*, 1948, **27**, pp. 379–423
2 SHANNON, C.E., and WEAVER, W.: 'The mathematical theory of communication' (Urbana: The University of Illinois Press, 1949)
3 PUN, T.: 'A new method for grey-level picture thresholding using the entropy of the histogram', *Signal Processing*, 1980, **2**, pp. 223–237
4 PUN, T.: 'Entropic thresholding, a new approach', *Comput. Graphics & Image Process.*, 1981, **16**, pp. 210–239
5 KAPUR, J.N., SAHOO, P.K., and WONG, A.K.C.: 'A new method for grey-level picture thresholding using the entropy of the histogram', *Comput. Graphics, Vision & Image Process.*, 1985, **29**, pp 273–285
6 DE LUCA, A., and TERMINI, S.: 'A definition of a non-probabilistic entropy in the setting of fuzzy sets theory', *Inf. & Control*, 1972, **20**, pp. 301–312
7 PAL, S.K.: 'A note on the quantitative measure of image-enhancement through fuzziness', *IEEE Trans. Pattern Anal. Machine Intell.*, 1982, **PAMI-4**, (2), pp. 204–206
8 PAL, S.K., KING, R.A., and HASHIM, A.A.: 'Automatic grey-level thresholding through index of fuzziness and entropy', *Patt. Recog. Lett.*, 1983, **1**, pp. 141–146
9 PAL, S.K.: 'A measure of edge ambiguity using fuzzy sets', *ibid.*, 1986, **4**, (1), pp. 51–56
10 PAL, S.K., and PRAMANIK, P.K.: 'Fuzzy measures in determining seed points in clustering', *Patt. Recog. Lett.*, 1986, **4**, (3), pp. 159–164
11 PAL, S.K., and CHAKRABORTY, B.: 'Fuzzy set theoretic measure for automatic feature evaluation', *IEEE Trans. Syst. Man & Cyberns*, 1986, **16**, (5), pp. 754–760
12 PAL, S.K., and DUTTA MAJUMBER, D.: 'Fuzzy mathematical approach to pattern recognition' (John Wiley and Sons, USA, 1986)
13 HALL, E.L.: 'Computer image processing and recognition' (Academic Press, USA, 1979)
14 PAL, S.K., and ROSENFELD, A.: 'Image enhancement and thresholding by optimisation of fuzzy compactness', *Tech. Report*, CAR-TR-252, 7 Centre for Automation Research, University of Maryland, USA

11 Appendix

11.1 Proof of P7 (Section 4.2)

$$H = \sum_{i=1}^{n} p_i e^{(1-p_i)} \quad 0 \leqslant p_i \leqslant 1 \quad \text{and} \quad \sum_{i=1}^{n} p_i = 1$$

$$= p_1 e^{(1-p_1)} + p_2 e^{(1-p_2)} + \cdots + p_{n-1} e^{(1-p_{n-1})} + (1 - p_1 - p_2 \cdots - p_{n-1}) e^{(p_1+p_2+\cdots+p_{n-1})}$$

as $p_n = 1 - p_1 - p_2 \cdots - p_{n-1}$

Now taking the partial derivative of H with respect to p_i, $1 \leqslant i \leqslant n$ and equating it to zero we get

$$\frac{\delta H}{\delta p_i} = 0, \quad i = 1, 2, \ldots, n-1$$

or

$$e^{(1-p_i)} - p_i e^{(1-p_i)} - e^{(p_1+p_2+\cdots+p_{n-1})} + (1 - p_1 - p_2 \cdots - p_{n-1}) e^{(p_1+p_2+\cdots+p_{n-1})} = 0$$

or

$$(1 - p_i) e^{(1-p_i)} = (p_1 + p_2 + \cdots + p_{n-1}) e^{(p_1+p_2+\cdots+p_{n-1})}. \tag{40}$$

Writing $1 - p_i = x_i$ and $p_1 + p_2 + \cdots + p_{n-1} = y$ we get

$$x_i e^{x_i} = y e^{y}. \tag{41}$$

Now define a function

$$f(x) = x e^{x} \quad 0 \leqslant x \leqslant 1 \tag{42}$$

Claim: $f(x)$ is a bijection, i.e. $f(x)$ maps x uniquely. Let x_1 and x_2 be the two points in [0, 1] i.e. $0 \leqslant x_1, x_2 \leqslant 1$. Then $f(x_1) = x_1 e^{x_1}$ and $f(x_2) = x_2 e^{x_2}$.

If $x_1 > x_2$ then $e^{x_1} > e^{x_2} \Rightarrow f(x_1) > f(x_2)$ and if $x_1 < x_2$ then $e^{x_1} < e^{x_2} \Rightarrow f(x_1) < f(x_2)$ thus, $f(x_1) = f(x_2)$ if, and only if, $x_1 = x_2$. Therefore, $f(x) = xe^x$, $0 \leqslant x \leqslant 1$ is a bijection. Using this fact and equation 41, one can write

$$x_i = y \quad \text{for } i = 1, 2, \ldots, n-1$$

or

$$1 - p_i = p_1 + p_2 + \cdots + p_{n-1} \quad \text{for } i = 1, 2, \ldots, n-1$$

Now taking summation on both sides over $i = 1, 2, \ldots, n-1$.

We get $\sum_i (1 - p_i) = \sum_i (p_1 + p_2 + \cdots + p_{n-1})$

or $(n-1) - (p_1 + p_2 + \cdots + p_{n-1})$

$$= (n-1)(p_1 + p_2 + \cdots + p_{n-1})$$

or $(n-1) = n(p_1 + p_2 + \cdots + p_{n-1})$

or $n - 1 = n(1 - p_n)$

or $n \cdot p_n = 1$

or $p_n = \dfrac{1}{n}$

Similarly, expressing other p_i, $i = 1, 2, \ldots, n-1$, in terms of the remaining probabilities one can prove the condition of maximality of H as $p_i = 1/n$ for $i = 1, 2, \ldots, n$.

11.2 Claim:

$$H = pe^{1-p} + (1-p)e^p \quad 0 \leqslant p \leqslant 1$$

monotonically increases in (0, 0.5) and monotonically decreases in (0.5, 1) and attains the maximum at $p = 0.5$.

Proof:

$$\frac{dH}{dp} = \frac{d}{dp}(pe^{1-p} + (1-p)e^p)$$

$$= e^{1-p} - pe^{1-p} - e^p + (1-p)e^p$$

$$= (1-p)e^{1-p} - pe^p$$

if $p \in (0, 0.5)$ then $(1-p)e^{1-p} > pe^p$ and if $p \in (0.5, 1)$ then $(1-p)e^{1-p} < pe^p$. Therefore,

$$\frac{dH}{dp} > 0 \quad \text{if } p \in (0, 0.5)$$

$$\frac{dH}{dp} < 0 \quad \text{if } p \in (0.5, 1)$$

and

$$\frac{dH}{dp} = 0 \quad \text{if } p = 0.5$$

Hence the proof.

11.3 Proof of P1 (Section 4.3)

If $\mu_A(x_i) = 0$ or 1, then

$$\mu_A(x_i)e^{\{1-\mu_A(x_i)\}} + \{1 - \mu_A(x_i)\}e^{\mu_A(x_i)} = 1$$

Therefore, if $\mu_A(x_i) = 0$ or 1 for all i, then $H'(A) = (1/n)\sum 1 = n/n = 1$, the minimum value of $H'(A)$ for $\mu_A(x_i) \in [0, 1]$ (taking the result of Appendix 11.2 into consideration).

11.4 Proof of P2 (Section 4.3)

Differentiating $H'(A)$ with respect to $\mu_A(x_i)$ and equating to zero we get

$$\frac{\delta H'(A)}{\delta \mu_A(x_i)} = 0$$

or

$$\frac{1}{n}\{e^{\{1-\mu_A(x_i)\}} - \mu_A(x_i)e^{\{1-\mu_A(x_i)\}} - e^{\mu_A(x_i)} + \{1 - \mu_A(x_i)\}e^{\{\mu_A(x_i)\}} = 0$$

or

$$\{1 - \mu_A(x_i)\}e^{\{1-\mu_A(x_i)\}} = \mu_A(x_i)e^{\mu_A(x_i)} \quad \text{for } i = 1, 2, \ldots, n.$$

Using the fact that $f(x) = xe^x$ is a bijection we can write that

$$1 - \mu_A(x_i) = \mu_A(x_i) \quad \text{for } i = 1, 2, \ldots, n$$

or

$$\mu_A(x_i) = \tfrac{1}{2} \quad \text{for } i = 1, 2, \ldots, n$$

Thus we see that $H'(A)$ attains the maximum value when all $\mu_A(x_i) = \frac{1}{2}$ for $i = 1, 2, \ldots, n$.

11.5 Proof of P3 (Section 4.3)

$$H'(A) = \frac{1}{n}\sum_i \{\mu_A(x_i)e^{\{1-\mu_A(x_i)\}} + (1 - \mu_A(x_i))e^{\mu_A(x_i)}\}$$

$$= \frac{1}{n}\sum_i C_{x_i}$$

where

$$C_{x_i} = \{\mu_A(x_i)e^{\{1-\mu_A(x_i)\}} + (1 - \mu_A(x_i))e^{\mu_A(x_i)}\}$$

In order to prove $P3$ it is enough to show that C_{x_i} monotonically increases for $\mu_A(x_i) \in [0, 0.5)$ monotonically decreases for $\mu_A(x_i) \in (0.5, 1]$ and attains the maximum value for $\mu_A(x_i) = 0.5$

$$\frac{dC_{x_i}}{d\mu_A(x_i)} = e^{\{1-\mu_A(x_i)\}} - \mu_A(x_i)e^{\{1-\mu_A(x_i)\}} - e^{\mu_A(x_i)} + (1 - \mu_A(x_i))e^{\mu_A(x_i)}$$

$$= (1 - \mu_A(x_i))e^{\{1-\mu_A(x_i)\}} - \mu_A(x_i)e^{\mu_A(x_i)}$$

If $\mu_A(x_i) \in [0, 0.5)$, then $(1 - \mu_A(x_i))e^{\{1-\mu_A(x_i)\}} > \mu_A(x_i)e^{\mu_A(x_i)}$; if $\mu_A(x_i) \in (0.5, 1]$ then $(1 - \mu_A(x_i))e^{\{1-\mu_A(x_i)\}} < \mu_A(x_i)e^{\mu_A(x_i)}$ and if $\mu_A(x_i) = 0.5$ then $(1 - \mu_A(x_i))e^{\{1-\mu_A(x_i)\}} = \mu_A(x_i)e^{\mu_A(x_i)}$. Therefore,

$$\frac{dC_{x_i}}{d\mu_A(x_i)} > 0 \quad \text{if } \mu_A(x_i) \in [0, 0.5)$$

$$< 0 \quad \text{if } \mu_A(x_i) \in (0.5, 1]$$

$$= 0 \quad \text{if } \mu_A(x_i) = 0.5.$$

Hence the proof.

11.6 Proof of P4 (Section 4.3)

$$H'(A) = \frac{1}{n}\sum_{i=1}^{n} [\mu_A(x_i)e^{\{1-\mu_A(x_i)\}} + \{1 - \mu_A(x_i)\}e^{\mu_A(x_i)}]$$

$$= \frac{1}{n}\sum_{i=1}^{n} [\{1 - \mu_A(x_i)\}e^{\mu_A(x_i)} + \mu_A(x_i)e^{\{1-\mu_A(x_i)\}}]$$

$$= \frac{1}{n}\sum_{i=1}^{n} [\mu_{\bar{A}}(x_i)e^{\{1-\mu_{\bar{A}}(x_i)\}} + \{1 - \mu_{\bar{A}}(x_i)\}e^{\mu_{\bar{A}}(x_i)}]$$

$$= H'(\bar{A})$$

Hence the proof.