Pattern Recognition, Vol. 26, No. 4, pp. 617 625, 1993
Printed in Great Britain

0031 3203/93 $6.00 + .00
Pergamon Press Ltd
Pattern Recognition Society

MINIMUM CROSS ENTROPY THRESHOLDING

C. H. Li and C. K. Lee†
Department of Electronic Engineering, Hong Kong Polytechnic, Hung Hom, Hong Kong

(*Received* 13 *February* 1992; *in revised form* 12 *August* 1992; *received for publication* 23 *September* 1992)

Abstract—The threshold selection problem is solved by minimizing the cross entropy between the image and its segmented version. The cross entropy is formulated in a pixel-to-pixel basis between the two images and a computationally attractive algorithm employing the histogram is developed. Without making a priori assumptions about the population distribution, this method provides an unbiased estimate of a binarized version of the image in an information theoretic sense.

Thresholding Image segmentation Minimum cross entropy Maximum entropy method

1. INTRODUCTION

Thresholding is an important preliminary step in many pattern recognition systems. The selection of a threshold will affect both the accuracy and the efficiency of the subsequent analysis of the segmented image. The principal assumption behind the approach is that the object and the background can be distinguished by comparing their gray level values with a suitably selected threshold value. The segmented image may not be suitable as the input for general image understanding routines due to its lack of consideration of the spatial context of the image pixels. However, the simplicity and speed of the thresholding algorithm make it one of the most widely used algorithms in automated systems ranging from medical applications to industrial manufacturing. The binarized image is especially suitable as the input for hardware implementation of template matching through correlation and moment-based recognition. Besides the application of global thresholding in image segmentation, it is also used in various classification problems in pattern recognition.

Research work in threshold selection is numerous and detailed reviews can be found in references (1–3). In general, threshold selection algorithms can be roughly classified as global or local according to the nature of the algorithms. Local thresholding algorithms select the threshold based on local properties in the histogram function such as the existence of maxima and minima. A typical example of a local thresholding algorithm is the bottom of the valley approach which involves locating local minima of the histogram. In principle, this approach relies on differentiation and is thus extremely prone to noise. Significant preprocessing such as smoothing the image and smoothing the histogram has to be done as a preliminary step in the procedure. Moreover, local extremal points are not guaranteed to exist and histogram enhancing or sharpening algorithms are often needed to overcome these difficulties. A particular advantage of the local thresholding algorithm is the determination of the number of classes without the need to be determined a priori. The global thresholding algorithms attempt to measure some global statistics of the histogram as the criteria for the selection. This kind of approach is less sensitive to noise and does not require elaborate enhancement which is usually sensitive to the individual image characteristics and requires a lot of supervision. Recently, Wilson and Spann[4] introduced a method of clustering combining the local and global thresholding approach which eliminates most of the above drawbacks. In this paper, we shall only discuss the global approach to thresholding and we use the two most well-known and most widely applied algorithms, the minimum error thresholding algorithm by Kittler and Illingworth[5] and Ostu's method[6] as the major comparisons to our method. In Ostu's method, the threshold is selected so as to maximize the class separability, which is based on the within-class variance, between-class variance, and the total variance of gray levels. This method is non-parametric, unsupervised and applies without a priori knowledge. This method has wide applicability and is often used as a standard algorithm with which other thresholding algorithms are compared. The major drawback of this algorithm is the existence of a bias in the threshold when the two underlying distributions have unequal variances or when the populations of the two distributions are very different.

In the minimum error approach, the sets of pixels which comprise the object and the background are both assumed to be normally distributed. A criterion function is constructed such that the selected threshold will minimize the average error in pixel classification. Besides assuming a normal distribution, the approach also assumes that the overlap between the two underlying distributions is small and the truncation error in the derivation of the algorithm can thus be ignored.

† Author to whom all correspondence should be addressed.

This method actually attempts to bypass the estimation of the mean, variance and standard deviation of the two distributions in the histogram. This algorithm gives a better estimate of the threshold when the distributions are in fact normally distributed and will not give a threshold when the distribution is a unimodal normal distribution. An improvement to the minimum error method is proposed by Cho *et al.*[7] which corrects the biased estimation of variances due to truncation. As the principle in modelling is identical, we shall use reference (5) as the basis for discussions and comparisons.

The above two algorithms are examples of thresholding algorithms which utilize the histogram as the only input data for the threshold selection. Another class of thresholding algorithms which are also developed solely on the histogram consists of the application of the maximum entropy principle to the problem of image segmentation. These approaches use the concept of entropy from information theory without explicit reference to the properties of an image as a two-dimensional distribution and have significant restrictions in practice. The details will be discussed in Section 3.

2. PRINCIPLE OF THE MAXIMUM ENTROPY

The maximum entropy principle was originally proposed by Jaynes[8] to the inference of unknown probability distribution under constraints. The role of the constraints is to limit the solution set to include only those solutions that are consistent with the data. While the inference problem is often underdetermined as a result of insufficient data, a number of feasible solutions often exist after applying all the constraints. The maximum entropy principle allows us to select the solution which gives the largest entropy. The original idea is that it will give the most unbiased estimate and allows a maximum freedom within the limit of the constraints. Throughout the years, the maximum entropy principle has undergone wide theoretical debates and has been applied with great success to various areas of science and engineering. With the use of the concentration theorem and the study of multiplicities, it has been shown that distributions of higher entropy have higher multiplicities and are thus more likely to be observed.[9] Axiomatic formulations have also shown that the maximum entropy method is the uniquely correct method for inductive inference when new information is given in the form of expected values.[10] It has been extended to be a general inference method that deals with the problem of recreation of positive and additive distributions including incoherent image intensity or power and spectral density. It is widely applied in various fields and is remarkably successful. For example, in the field of spectral estimation the maximum entropy method provides better resolution than other traditional methods such as the maximum likelihood method.[11]

The cross entropy was proposed by Kullback[12] under the name of directed divergence. The cross entropy measures the information theoretic distance between two distributions $P = \{p_1, p_2, \ldots, p_N\}$ and $Q = \{q_1, q_2, \ldots, q_N\}$ by

$$D(Q, P) = \sum_{k=1}^{N} q_k \log_2 \frac{q_k}{p_k}. \tag{1}$$

This measure $D(Q, P)$ is also studied by Renyi[13] as the information theoretical distance between the two distributions P and Q. Renyi also points out that the formula can be interpreted as the expectation of the change in the information content when we are using Q instead of P. The minimum cross entropy method can be seen as an extension of the maximum entropy method by setting equal initial estimates for all p_i when no prior information is available.

3. MINIMUM CROSS ENTROPY SEGMENTATION

For an experiment such as dice throwing, each toss is a separate trial and the outcome of each trial does not affect the others and the entropy maximization leads to independent probabilities for different trials. However, for the case of time-series analysis or image modelling, there is strong correlation in the data and in order to take account of the mutual information content, the modelling has to treat the entire time series or image as a single trial and the combinatorial argument of the maximum entropy principle applied to the collection of many different realizations of the experiment.

Former work in applying the maximum entropy method to image segmentation does not make the above distinction and all consider the pixel generation process as independent trials.[14–16] They use the normalized gray level histogram as the gray level probability distributions based on random trials of individual pixels having a certain gray level and measure the entropy of the pixel distributions. The maximum entropy thresholding method proposed by Kapur *et al.*[16] is the algorithm that is considered superior to other entropy thresholding algorithms.[17] However, this maximum entropy method is still not well accepted and performs poorly at times and the authors have various extensions of their work. Wong and Sahoo[18] used a combination of two measures concerning the spatial information in the image and the entropy as the criteria for selecting the threshold. While retaining the histogram entropy function in the maximum entropy thresholding method, Kapur introduces a set of additional heuristic principles to select the threshold.[17] Thus the formulation of a general histogram thresholding using the entropy principle without additional criteria has not been successful.

In the proposed scheme, the segmentation process is posed as one of reconstruction of the image distribution. Consider the image function f: $N \times N \rightarrow G$, where

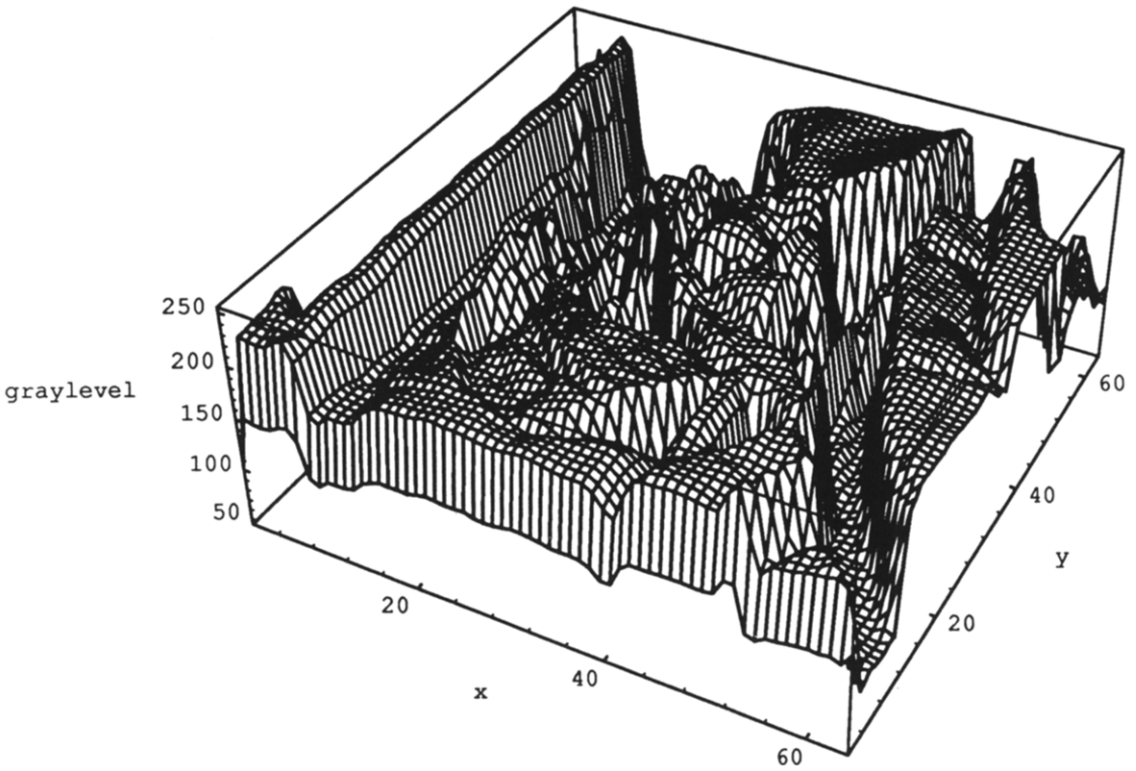

Fig. 1. Three-dimensional plot of a gray level image.

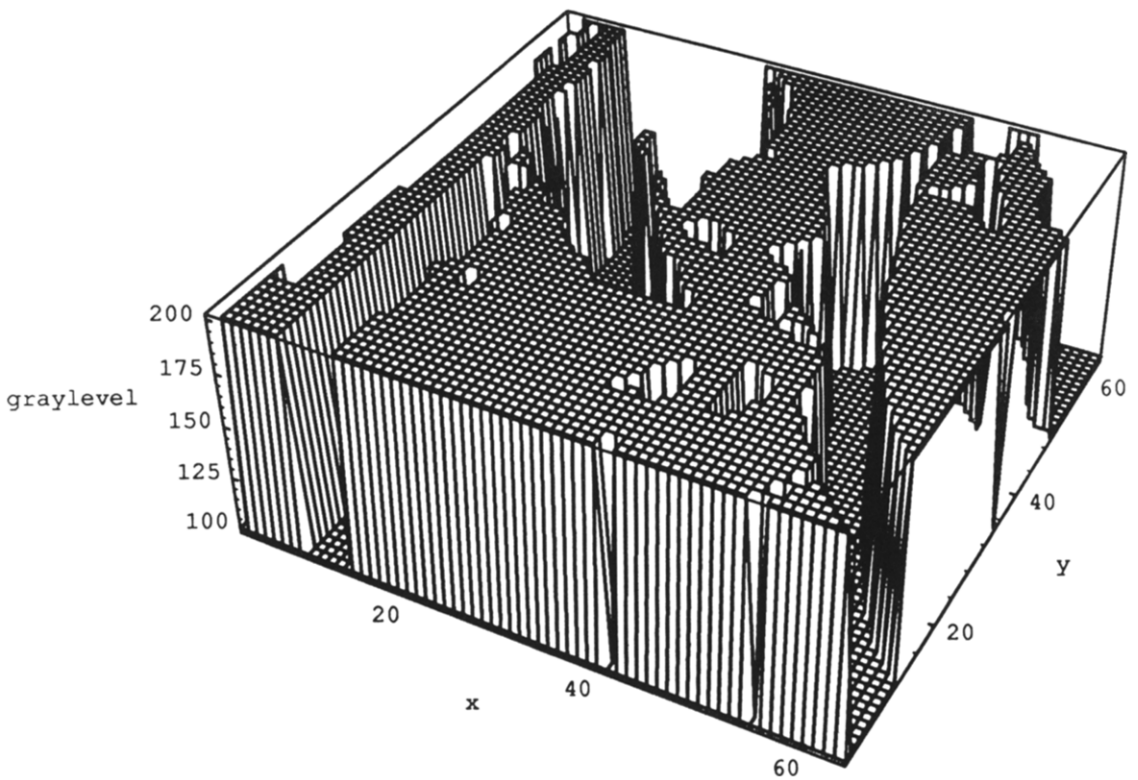

Fig. 2. Three-dimensional plot of the segmented image.

N is the set of natural integers and $G = \{1, \ldots, L\} \subset N$ the set of gray levels. The segmentation process is the construction of a function g: $N \times N \to S$, where $S = \{\mu_1, \mu_2\} \in \mathscr{R}^+ \times \mathscr{R}^+$, and $\mathscr{R}^+$ is the set of real positive numbers. Figures 1 and 2 show a sample $f(x, y)$ and a segmented image $g(x, y)$ in the coordinate space with the gray level displayed on the z-axis. The segmented image $g(x, y)$ will be constructed as follows:

$$g(x,y) = \begin{cases} \mu_1, & f(x,y) < t \\ \mu_2, & f(x,y) \geq t \end{cases}. \tag{2}$$

The segmented image $g(x, y)$ is uniquely determined from $f(x, y)$ by the specification of three unknown parameters t, μ_1 and μ_2. A criteria has to be constructed to enable us to find the optimal g, or equivalently the optimal set of parameters t, μ_1 and μ_2 that resembles f as close as possible. That is

$$\eta(g) \equiv \eta(t, \mu_1, \mu_2). \tag{3}$$

The criteria function is generally some sort of distortion measure, for example, the mean square difference of g from f is a general measure that can be used. The minimum error method and Ostu's method both belong to this category. At the end of this section, we will show that Ostu's method minimizes the mean square distance between the image and its segmented version while the proposed method minimizes the cross entropy. Instead of using the mean square differences, the cross entropy is the preferred measure for positive and additive distributions.

The above problem can be considered from a classical maximum entropy inference problem using constraints. Thus a set of values $\mathbf{G} = \{g_1, g_2, \ldots, g_N\}$, where N is the number of pixels in the image, has to be inferred from the observed image $\mathbf{F} = \{f_1, f_2, \ldots, f_N\}$ together with the use of suitable constraints. Here the distributions are obtained by linearizing the two-dimensional distributions in an identical way, so g_i and f_i come from the same location in the image space. And $\mathbf{G}$ contains elements having only two values, say μ_1 and μ_2, which are as yet unknown. Next, the intensity conservation constraint is applied. Since we want the reconstructed distribution to follow the data closely, the observed image intensity $\mathbf{F}$ gives the constraints on the values of μ_1 and μ_2 such that the total intensity in the reconstructed image is identical to the observed image in both categories. The constraints can be summarized as

$$\begin{aligned} &\text{(i)} \quad g_i \in \{\mu_1, \mu_2\} \\ &\text{(ii)} \quad \sum_{f_i < t} f_i = \sum_{f_i < t} \mu_1 \\ &\text{(iii)} \quad \sum_{f_i \geq t} f_i = \sum_{f_i \geq t} \mu_1 \end{aligned} \tag{4}$$

which allows the determination of μ_1 and μ_2 by the following equations:

$$\mu_1(t) = \frac{\sum_{f_i < t} f_i}{N_1}, \quad \mu_2(t) = \frac{\sum_{f_i \geq t} f_i}{N_2} \tag{5}$$

where N_1 and N_2 are the number of pixels smaller in the two regions. Combining equations (1), (2) and (5), we get

$$\eta(t) = \sum_{f_i < t} f_i \log\left(\frac{f_i}{\mu_1(t)}\right) + \sum_{f_i \geq t} f_i \log\left(\frac{f_i}{\mu_2(t)}\right). \tag{6}$$

The threshold is then selected by

$$t_0 = \min_t (\eta(t)) \tag{7}$$

where t_0 is the required threshold. The above summation is done on the entire image, however, there are repeated calculations that can be grouped. Thus we arrive at the following formulae:

$$\mu_1(t) = \frac{\sum_{j=1}^{j=t-1} jh_j}{\sum_{j=1}^{j=t-1} h_j}, \quad \mu_2(t) = \frac{\sum_{j=t}^{j=L} jh_j}{\sum_{j=t}^{j=L} h_j} \tag{8}$$

$$\eta(t) = \sum_{j=1}^{j=t-1} jh_j \log\left(\frac{j}{\mu_1(t)}\right) + \sum_{j=t}^{j=L} jh_j \log\left(\frac{j}{\mu_2(t)}\right).$$

The form of the cross entropy in our model bears similarity to the image entropy derived by Skilling[19] in which a set of four axioms has been employed to derive the following entropic functions:

$$s(f, m) = \int \mathrm{d}x (f(x) - m(x)) - f(x) \log (f(x)/m(x)) \tag{9}$$

where $f(x)$ is the image intensity distribution and $m(x)$ the model for the image. In fact, if the total intensity conservation constraints are included, the two equations are identical with a sign reversed, since the first two terms of the integral in equation (8) cancel out when integrated over all categories.

While the introduced method minimizes the cross entropy between the image and its segmented version, Ostu's method of minimization of the between-class variance can also be derived by the above method using mean square distance as the metric between the two images and the same set of constraints in equation (4). The criterion function in this case becomes

$$\theta(t) = \sum_{f_i < t} (f_i - \mu_1(t))^2 + \sum_{f_i \geq t} (f_i - \mu_2(t))^2. \tag{10}$$

Grouping the summation using the histogram, the criterion function becomes

$$\theta(t) = \sum_{j < t} h_j (j - \mu_1(t))^2 + \sum_{j \geq t} h_j (j - \mu_2(t))^2 \tag{11}$$

which is the within-class variance as defined in reference (6). Minimization of the function defined in equation (11) is equivalent to maximizing Ostu's criteria as the sum of the within-class variance and the between-class variance which is a constant independent of the threshold selected.

The use of the cross entropy function is not limited to the thresholding application. It can be extended to cover other areas of image segmentation when combined with other constraints. For example, a region-based segmentation method such as the split and merge

method using the cross entropy as a criterion for merging regions can be formulated with constraints on the labelling of spatial coordinates.

4. IMPLEMENTATIONS AND DISCUSSIONS

For the purpose of comparison, we adopt the same approach that was taken by Kittler and Illingworth. Ostu's method, the minimum error method, and the minimum cross entropy method are implemented on the normal distributions as discussed in reference (5). These histograms are shown in Figs 3–5. We also use the histogram of an actual image (Fig. 6) taken from an industrial setting. The thresholds selected by the four algorithms are summarized in Table 1.

The results of the first three histograms on the

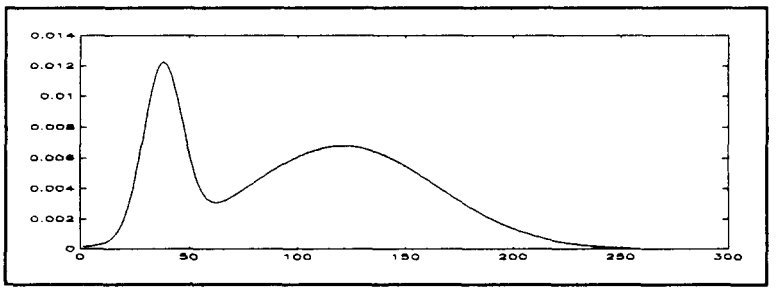

Fig. 3. Histogram (a).

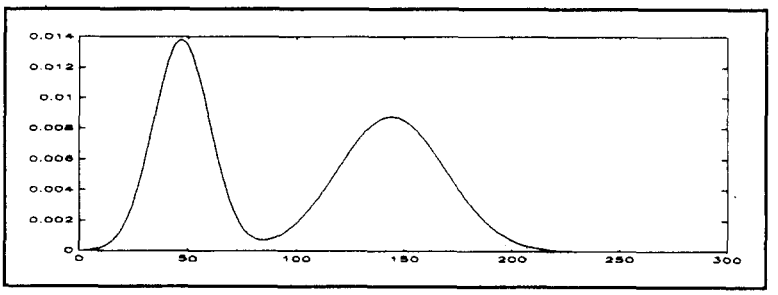

Fig. 4. Histogram (b).

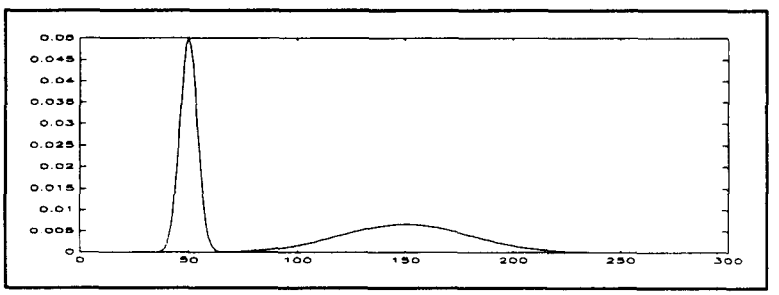

Fig. 5. Histogram (c).

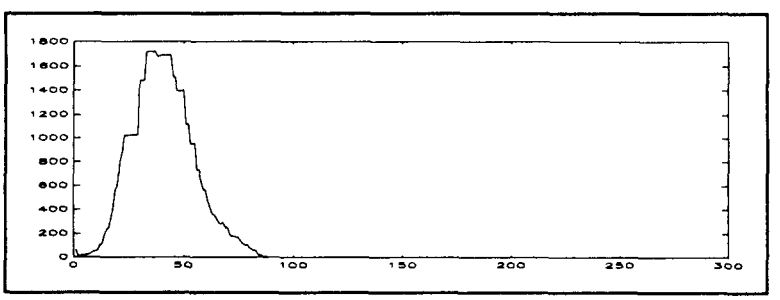

Fig. 6. Histogram (d).

Table 1. Threshold selected for the four trial histograms using Ostu's method, minimum error method, maximum entropy method, and the minimum cross entropy (MCE) method

Histogram	Ostu's method	Minimum error method	Maximum entropy method	MCE method
(a)	97	59	130	83
(b)	96	82	118	88
(c)	101	64	165	93
(d)	41	None	58	40

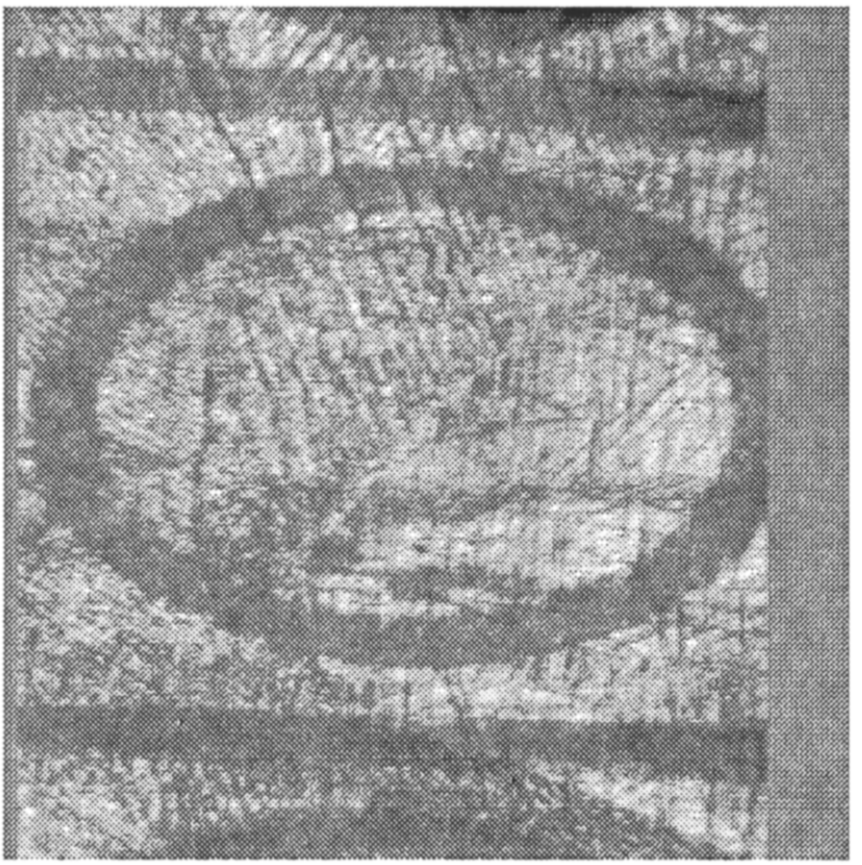

Fig. 7. Strain image.

Gaussian distributions are consistent as the fact that the minimum error method takes the normal distribution as the model and subsequently outperforms both the minimum cross entropy method and Ostu's method. However, the minimum cross entropy is able to give a threshold which is closer to the optimal threshold than both Ostu's method and the maximum entropy method.

For the strain measurement image in Fig. 7, the goal is to measure the change in the major and minor axes of the ellipse which will give information about the strain of the part of a piece of metal. The first step in this process is the segmentation of the parallel lines and the ellipse from the background. The histogram is extracted and shown in Fig. 6. The minimum error method does not give any threshold as there is no internal minimum for the criterion function. The histogram is thus wrongly inferred as unimodal while the actual situation is that the two distributions are highly overlapping with a significant amount of embedded noise. This shows that wrong assumptions may be disastrous. The minimum cross entropy method and Ostu's method successfully select a threshold which results in the segmented image with significantly less noise than thresholds which are selected otherwise.

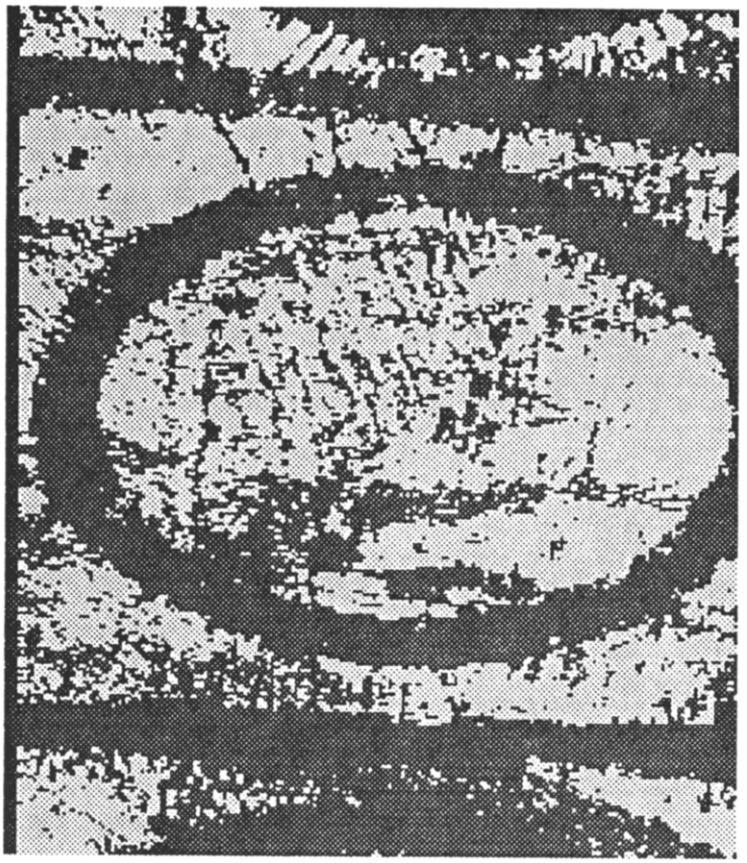

Fig. 8. Strain image with threshold at 40.

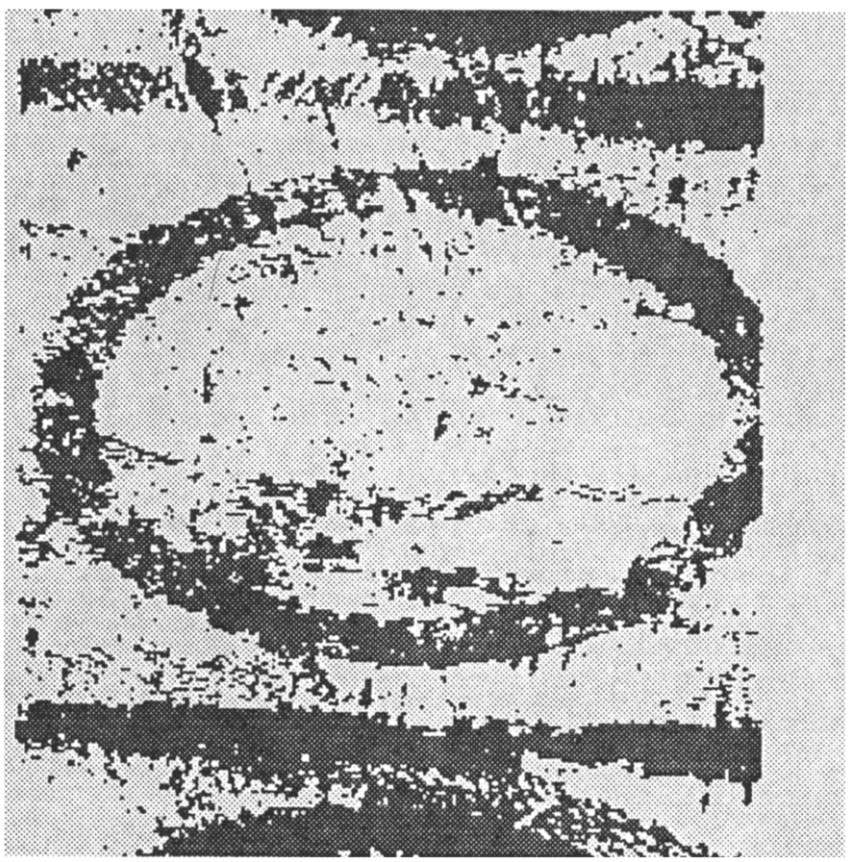

Fig. 9. Strain image with threshold less than 40.

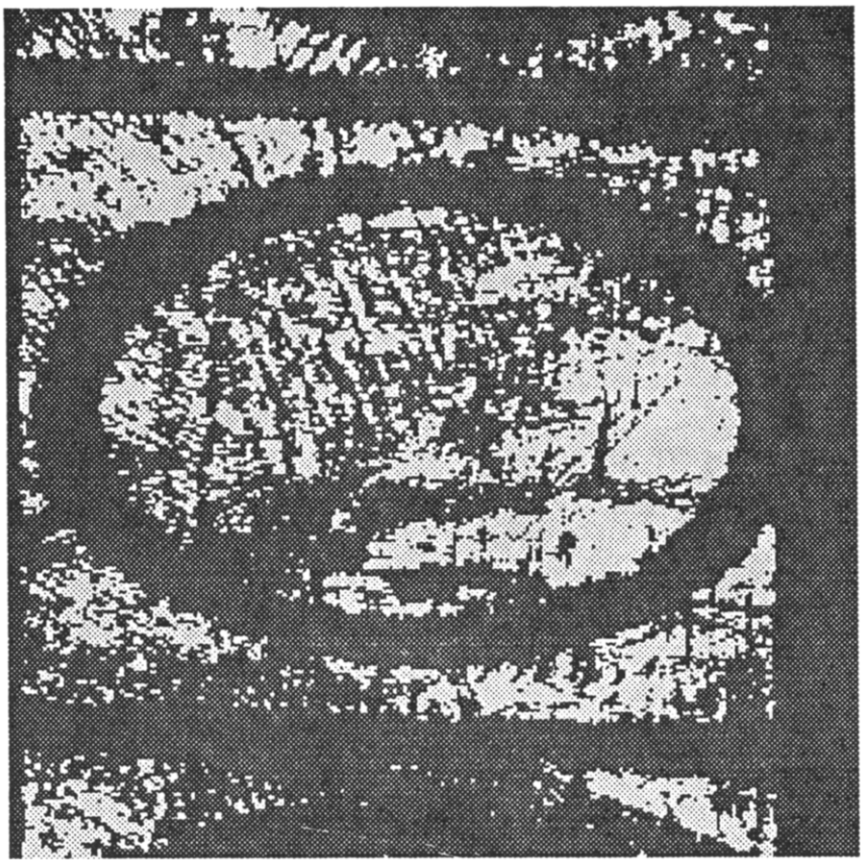

Fig. 10. Strain image with threshold larger than 40.

To illustrate this, compare Fig. 8 which shows the segmented image with threshold value 40 as selected by our method with Figs 9 and 10 which shows the segmented image with threshold slightly below and slightly above our estimated value. In Fig. 9 where a smaller threshold is used, there are more holes and openings in the lines and the ellipse. In Fig. 10, where a larger threshold is used, too much of the background is included. This illustrates the selection of the correct threshold is critical in this application as it is generally true with highly overlapping distributions.

The thresholds selected by the maximum entropy method are very far from the thresholds selected by the other three algorithms. This could be explained by the existence of the bias in the algorithms toward splitting the histogram into equal halves.[17]

5. CONCLUSIONS

The application of the minimum cross entropy method to the segmentation problem is developed and applied to both asymmetrically normal distributions and a real image with success. Our algorithm selects the threshold which minimizes the cross entropy between the thresholded image and the original image. Compared to Ostu's method, the use of cross entropy as the distortion measure instead of the variance, permits the selection of thresholds without the tendency of biasing when the two distributions have highly unequal populations and highly unequal variances. The minimum error method and its improved version provide a computationally efficient method of computing the threshold provided the population is normally distributed. However, for situations which we have no knowledge about the population model, such assumptions may not be necessarily correct. Without making a priori assumptions about the inherent distributions, the minimum cross entropy algorithm gives the most unbiased estimate of the binarized version of the image.

REFERENCES

1. P. K. Sahoo, S. Soltani and A. K. C. Wong, A survey of thresholding techniques, *Comput. Vision Graphics Image Process.* **41**, 233–260 (1988).
2. A. K. C. Wong and P. K. Sahoo, A gray-level threshold selection method based on maximum entropy principle, *IEEE Trans. Syst. Man Cybern.* **SMC-19**(4), 866–871 (July/ August 1989).
3. J. Kittler and J. Illingworth, Threshold selection based on a simple image statistics, *Comput. Vision Graphics Image Process.* **30**, 125–147 (1985).
4. R. Wilson and M. Spann, A new approach to clustering, *Pattern Recognition* **23**, 1413–1425 (1990).
5. J. Kittler and J. Illingworth, Minimum error thresholding, *Pattern Recognition* **19**, 41–47 (1986).

6. N. Ostu, A threshold selection method from gray-level histogram, *IEEE Trans. Syst. Man Cybern.* **SMC-9**(1), 62–66 (1979).
7. S. Cho, R. Haralick and S. Yi, Improvement of Kittler and Illingworth's minimum error thresholding, *Pattern Recognition* **22**, 609–617 (1989).
8. E. T. Jaynes, Information theory and statistical mechanics, *Phys. Rev.* **106**, 620–630 (1957).
9. E. T. Jaynes, On the rationale of maximum-entropy methods, *Proc. IEEE* **70**(9), (September 1982).
10. J. E. Shore and R. W. Johnson, Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross entropy, *IEEE Trans. Inf. Theory* **IT-26**, 26–37 (January 1980).
11. N. A. Malik and J. S. Lim, Properties of two-dimensional maximum entropy power spectral estimation, *IEEE Trans. Acoust. Speech Signal Proc.* **ASSP-30**, 788–798 (October 1982).
12. S. Kullback, *Information Theory and Statistics.* Wiley, New York (1959).
13. A. Renyi, *A Diary on Information Theory.* Akadememiai Kiado Wiley, Budapest (1984).
14. T. Pun, A new method for grey level thresholding using the entropy of the histogram, *Signal Process.* **2**, 223–237 (1980).
15. T. Pun, Entropic thresholding, a new approach, *Comput. Graphics Image Process.* **16**, 210–239 (1981).
16. J. N. Kapur, P. K. Sahoo and A. K. C. Wong, A new method for grey-level picture thresholding using the entropy of the histogram, *Comput. Graphics Vision Image Process.* **29**, 273–285 (1985).
17. J. N. Kapur, *Maximum Entropy Models in Science and Engineering.* Wiley Eastern, New Delhi (1989).
18. A. K. C. Wong and P. K. Sahoo, A gray-level threshold selection method based on maximum entropy principle, *IEEE Trans. Syst. Man Cybern.* **SMC-19**(4), 866–871 (July/August 1989).
19. J. Skilling, *Classic Maximum Entropy, Maximum Entropy and Bayesian Methods*, J. Skillings, ed., pp. 45–52. Kluwer Dordrecht (1988).

APPENDIX

For trial histograms (a), (b), and (c), the following normal distribution model

$$h(g)=\frac{p_1}{\sqrt{(2\pi)}\sigma_1}\exp\left(-\frac{(g-\mu_1)^2}{2\sigma_1^2}\right)+\frac{p_2}{\sqrt{(2\pi)}\sigma_2}\exp\left(-\frac{(g-\mu_2)^2}{2\sigma_2^2}\right)$$

with parameters from reference (4) are used:

(a) $p_1=0.25, \mu_1=38, \sigma_1=9, p_2=0.75, \mu_2=121, \sigma_2=44$;
(b) $p_1=0.45, \mu_1=47, \sigma_1=13, p_2=0.55, \mu_2=144, \sigma_2=25$;
(c) $p_1=0.5, \mu_1=50, \sigma_1=4, p_2=0.5, \mu_2=150, \sigma_2=30$.

Histogram (d) is obtained from the strain-gauge measurement image (Fig. 6) obtained from an industrial setting.

About the Author—CHUN-HUNG LI was born in Hong Kong in 1966. He received the B.Sc. degree in physics from the State University of New York at Stony Brook. He is currently a M.Phil. student at the Department of Electronic Engineering, Hong Kong Polytechnic. His research interests include computer vision and image processing.

About the Author—C. K. LEE received the B.Sc. and Ph.D. degrees in electronic engineering from the University of London and UCNW, University of Wales, in 1977 and 1984, respectively. He worked in the Hirst Research Center, GEC, in 1977. From 1981 to 1986, he was employed as Supervisory Engineer in ASM Assembly Automation, Hong Kong, to develop pattern recognition systems for wire-bonder machines. In 1986, he joined the Hong Kong Polytechnic. He is now a senior lecturer in the Department of Electronic Engineering. His research interests include digital signal processing, computer vision systems, and pattern recognition systems.