ELSEVIER

Pattern Recognition Letters 17 (1996) 29–36

Pattern Recognition Letters

Using spatial information as an aid to maximum entropy image threshold selection

A.D. Brink *

Department of Physics, University of Pretoria, Pretoria 0002, South Africa

Received 14 September 1994; revised 17 May 1995

Abstract

Many image grey-level thresholding methods based on the theory of maximum entropy have been proposed in the past. However, the exact definition of what is meant by "image entropy" has varied considerably, with measures based on the image itself, its histogram or other related distributions being proposed. Most of these ignore the spatial component of the image, assuming pixels within the image to be independent of each other. This is convenient, but it is also counter-intuitive. The inclusion of context-related information can be a simple matter, as is demonstrated in this article.

A common measure of image entropy is based on the notion of viewing the image itself as a probability distribution (often referred to as the "monkey model" of the image). Direct application of this measure as a criterion function for grey-level threshold selection yields generally unsatisfactory results. With the inclusion of spatial information in the entropy measure, the results improve dramatically.

Keywords: Entropy; Spatial information; Segmentation; Thresholding; Image processing

1. Introduction

Threshold selection is one of the simplest and most widely used methods for image region segmentation (Gonzalez and Woods, 1992; Haralick and Shapiro, 1985; Sahoo et al., 1988). Briefly, it is based on the assumption that illumination of the scene is relatively uniform and that the regions of interest (objects) in the scene differ significantly in brightness (grey-level) from their background. Other features, such as texture, may be also be used either to complement the simple grey-level thresholding approach or as an alternative to it. As such features can themselves be expressed numerically (effectively as "grey-levels"), we will concern ourselves only with grey-level thresholding here. Ideally, a threshold selection method should be locally variable to make allowance for non-uniformities in illumination (dynamic thresholding). We also require at times the selection of multiple thresholds to segment the image into more than two classes. However, such methods can be computationally expensive and can in any case be obtained as relatively simple extensions of the global techniques (Nakagawa and Rosenfeld, 1979; Brink, 1991a). This article thus focuses on the criteria for selection of a single, global, grey-level threshold.

The selection of a grey-level threshold can be, and often is, *ad hoc*, based on our observation and expectations of the scene and the applications we may have in mind. A more scientific approach is based on what we know about the image and what we are effectively

* Email: abrink@scientia.up.ac.za

SSDI 0167-8655(95)00096-8

doing to it by segmenting it. A common approach is to base the choice of threshold on the optimization of some *criterion function*. This is a function of the threshold grey-level value, reflecting some property either of the segmented image or of a distribution, such as the grey-level histogram, related to the image. Brink's (1989) correlation criterion or, equivalently, Otsu's (1979) "measure of goodness of threshold" are typical of this approach, maximizing in this case the magnitude of the correlation between the segmented (binary) image and the original grey-scale input.

A measure which has gained a lot of attention lately is the information-theoretic measure of entropy (Shannon and Weaver, 1949) and the theory of maximum entropy (MaxEnt) put forward as a method (or, in fact, *the* method) for solving problems of Bayesian inference (Skilling, 1989). It is usually applied to problems such as image restoration and reconstruction, but has enjoyed limited use as a criterion function for threshold selection (Kapur et al., 1985; Abutaleb, 1989; Pal and Pal, 1989).

In Section 2 we briefly describe the concept of entropy and its relationship to the "monkey model" of the image. The problem of the tacit assumption of spatial independence is addressed, and the "corrected" form of the entropy expression is presented. In Section 3 the threshold selection scheme using entropy as a criterion function is described. Section 4 shows some typical results, showing the effect of taking spatial information into account, and in Section 5 these results and their evaluation are discussed.

2. Image entropy and the monkey model

2.1. The Shannon entropy

The most generally accepted form of entropy was derived by Shannon (Shannon and Weaver, 1949) in connection with information theory. Given a discrete probability distribution p_i, $i = 1, \ldots, N$, the entropy is given by

$$H = -\sum_{i=1}^{N} p_i \log p_i, \quad \sum_{i=1}^{N} p_i = 1 \tag{1}$$

where $i = 1, \ldots, N$ are a set of possible outcomes or states of a discrete information source modelled as a Markov process. Shannon's measure is used as a measure of information gain, choice and uncertainty.

Shannon points out a number of interesting properties of (1), including:

(i) $H = 0$ if and only if $p_i = 0 \quad \forall\, i \neq j$, $p_j = 1$, where j can indicate any position in the distribution (note that $0 \log 0$ is defined to be zero). Otherwise H is positive. This makes intuitive sense, since $p_j = 1$ indicates certainty of the outcome, and the information gained by the occurrence of event j is thus zero.

(ii) For a given number of discrete states N, H is a maximum when all the p_i are equal, i.e. $p_i = 1/N \quad \forall\, i = 1, \ldots, N$. Intuitively this is the most uncertain situation: all outcomes are equally likely, making accurate prediction impossible.

In the case of continuous distributions, Jaynes (1968) has pointed out that the continuous form of (1),

$$H = -\int p(x) \log p(x)\, \mathrm{d}x$$

does not hold, as it lacks invariance under a change of variables $x \rightarrow y(x)$ and it ultimately turns out not to be the correct information measure for a continuous distribution. The corresponding expression in the continuous case was shown to be

$$H = -\int p(x) \log \frac{p(x)}{m(x)}\, \mathrm{d}x$$

where $m(x)$ is in general an "invariant measure" function proportional to the limiting density of discrete points, i.e. related to our sampling of x. The integrals in both expressions are assumed to cover the full range over which the distributions are defined. Skilling (1986) has since discussed a "discretized" form of this expression

$$H = -\sum_{i=1}^{N} p_i \log \frac{p_i}{m_i} \tag{2}$$

The entropy of the distribution p is determined relative to a measure m over the same domain. In image processing m can usually be interpreted as a model based on known constraints (prior information), relative to which the entropy, given "data" (posterior information) yielding a new distribution p, is measured. This is a more general expression, (1) representing the special case where m is a uniform prior distribution.

The maximum entropy formalism attempts to maximize (1) or (2), subject to known constraints about the parameters we are trying to estimate (p_i, $i = 1, \ldots, N$). Effectively, by maximizing the entropy, we are attempting to maximize our information gain or, equivalently, arrive at a solution which fits our prior knowledge but makes no assumptions beyond what is known. In the absence of any prior information, the maximum entropy distribution is simply the uniform distribution, as indicated by Shannon's point (ii) above. It can thus be aptly viewed as an application of Laplace's principle of insufficient reason, which states that the uniform distribution is the most unbiased when one has no prior knowledge regarding a probabilistic event.

2.2. *The monkey model and the entropy of an image*

A digital image can be regarded as a probability distribution with each pixel's grey-level representing an estimate of the number (or probability) of photons reaching that point (Frieden, 1980). We imagine the scene to be imaged as being made up of a fixed number of photons (unit grey-levels) G and our initial image as an empty grid or raster consisting of N cells (pixels). The G photons are allocated one at a time among the N cells with uniform spatial probability (Frieden, 1972). This model has been dealt with in some depth by various authors and is often referred to as the "monkey model" due to the analogy of a group of monkeys randomly throwing G balls at a 2-dimensional array of N boxes to form the image (Gull and Daniell, 1978; Skilling, 1986; Jaynes, 1986). Each ball represents a unit grey-level and each box a pixel. A diagram of the model for $G = 20$ photons is shown in Fig. 1: for clarity it is represented in one dimension.

Jaynes' *principle of maximum degeneracy* (Jaynes, 1968) states that, subject to *a priori* constraints, the most degenerate image to be formed in this fashion is the most likely to occur. In other words, subject to any prior knowledge we may possess about the system, we wish to maximize the degeneracy of our estimate of the image. Our only *a priori* constraint for this model is that of non-negativity, i.e. $g_i \geqslant 0, 1 \leqslant i \leqslant N$, where g_i is the grey-level of pixel i. The photons can be considered to be indistinguishable and any number can occupy a given cell. Any given image has a corresponding non-unique grey-level histogram. The number of ways W that a general image $(g_1, g_2, \ldots, g_N)$ can be formed is given by the Boltzmann law

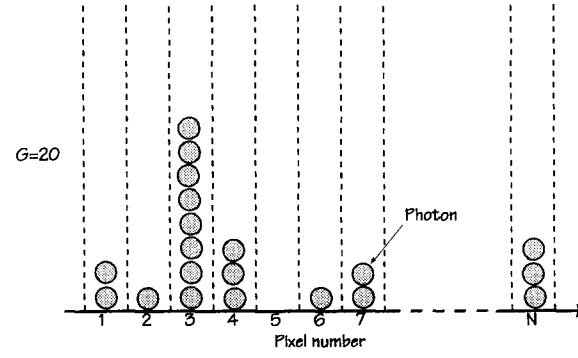

Fig. 1. Photon/unit grey-level allocation model ("monkey model") in 1 dimension. G photons are dealt out among N cells (pixels) with uniform spatial probability. The model obeys a simple positivity constraint.

$$W(g_1, \ldots, g_N) = \frac{G!}{g_1! \cdots g_N!} \tag{3}$$

Following Jaynes' rationale, we set W = max or, equivalently, $\log W$ = max. Using Stirling's approximation

$$\log m! \simeq (0.5 + m) \log m - m + 0.5 \log 2\pi$$

we find that the quantity to be maximized has the same form as the Shannon entropy (1). We maximize the expression

$$H(g_1, \ldots, g_N) = -\sum_{i=1}^{N} g_i \log g_i \tag{4}$$

Although the g_i are not probability values as such, the entropy expression has been shown to hold for positive, additive distributions in general (Skilling, 1988). Note that in order to compare the entropies of different images the grey-levels should be normalized to "probability" values by dividing each pixel value by $G = \sum_i g_i$ or, equivalently, imposing the constraint that $\sum_i g_i = G$ = constant $\forall$ images. In terms of probabilities, the expression for entropy (Shannon and Weaver, 1949) is given (from (1)) by

$$H = -\sum_{i=1}^{N} p_i \log p_i \quad , \; p_i = \frac{g_i}{G}.$$

It was indicated in Section 2.1 that, more correctly, entropy is a *relative* measure, given by

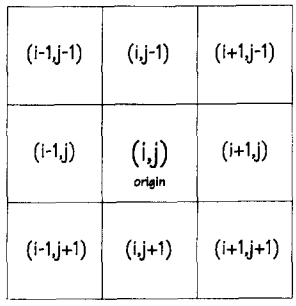

Fig. 2. 3×3 neighbourhood of a pixel centred at image coordinates (i, j).

$$H = -\sum_{i=1}^{N} p_i \log \frac{p_i}{m_i}.$$

The standard Shannon entropy (1) or (4) assumes that the individual events (pixels), $i = 1, \ldots, N$, are independent of each other or, rather, it implies that *dependence* is not assumed. This means that the pixels of the image may be randomly rearranged without affecting the entropy as measured by this formula, which is clearly counter-intuitive. The measure m may be used to address this problem, as suggested by Skilling in his reply to Titterington (Titterington, 1984) who raised precisely this criticism of the maximum entropy method. While the exact nature of the pixel inter-dependence may vary from image to image (consider, for example, an astronomical image as opposed to a standard holiday snapshot), we *can* easily measure local grey-level variations within the image. A simple measure is the local grey-level variance: this is usually determined over some specified neighbourhood of each pixel, e.g. the 3×3 neighbourhood (N_3) as shown in Fig. 2.

The grey-level variance over the neighbourhood of Fig. 2 is given by

$$\sigma_i^2 = \sum_{i \in N_3} \frac{(g_i - \mu_{N_3})^2}{9}$$

where μ_{N_3} is the mean grey-level in the pixel neighbourhood. The relative entropy (2) of the image is then given by

$$H = -\sum_{i=1}^{N} p_i \log \frac{p_i}{m_i},$$

$$p_i = \frac{g_i}{G}, \quad m_i = 1 + \sigma_i^2. \tag{5}$$

To simplify matters the measure m_i is set to $1 + \sigma_i^2$ as there is no guarantee that the variance will be non-zero in all neighbourhoods.

The interpretation of m_i here is thus as a "weighting" factor on the pixel illumination probabilities p_i of the image. The choice of what exact form the m_i should take is based on qualitative prior knowledge of the human visual system: edges, borders and similar areas of detail are essential to human visual understanding and it is thus desirable, for both visual and automated interpretation, that such information should be preserved in the segmented image. To this end, m_i may be selected to emphasize such features in the threshold selection process by using a measure of local grey-level variation. The 3×3 neighbourhood variance above has been chosen here as one feasible measure which should yield the desired result: other measures such as gradient operators or texture and busyness measures (Gonzalez and Woods, 1992) can be expected to yield similar results.

The selection of thresholds based on maximization of the entropy forms (1) and (2) is dealt with in Section 3.

3. Threshold selection

The simplest method of threshold selection, given the relatively small number of discrete grey-levels making up most images (usually 256 grey-levels for a 1 byte-per-pixel image), is a simple iterative search through all possible threshold values to find the one which optimizes the criterion function (in our case, the value that maximizes the entropy). The process becomes more complex when a variable threshold or multiple thresholds are required: more sophisticated optimization schemes (conjugate gradients, the isodata algorithm, etc.) should then be considered. For our purposes an iterative search will be adequate.

Following the approach of Kapur et al. (1985), we view the image as being split into two distributions by the threshold. This allows us to define a class entropy for each of the distributions. While Kapur et al. used the entropy of the image histogram (a distribution showing the relative frequencies of occurrence of the grey-levels in the image), we will use the entropy based on the image itself, as defined in the previous section. In the case where no context related informa-

tion is used the entropy (1) can be completely determined from the image histogram alone:

$$H = -\sum_{i=1}^{N} p_i \log p_i, \quad p_i = \frac{g_i}{G}$$

$$= -\sum_{g=0}^{n-1} f_g \frac{g}{G} \log \frac{g}{G} \tag{6}$$

where

$$G = \sum_{i=1}^{N} g_i = \sum_{g=0}^{n-1} g f_g$$

is the total illumination of the image, n is the number of grey-levels in the image and f_g is the number of pixels having grey-level g. Thresholding the image results in two classes. We can define corresponding class entropies

$$H_0(T) = -\sum_{i=1 \ (i\in 0)}^{N} \frac{g_i}{G_0(T)} \log \frac{g_i}{G_0(T)}$$

$$= -\sum_{g=0}^{T} f_g \frac{g}{G_0(T)} \log \frac{g}{G_0(T)}, \tag{7a}$$

$$H_1(T) = -\sum_{i=1 \ (i\in 1)}^{N} \frac{g_i}{G_1(T)} \log \frac{g_i}{G_1(T)}$$

$$= -\sum_{g=T+1}^{n-1} f_g \frac{g}{G_1(T)} \log \frac{g}{G_1(T)} \tag{7b}$$

where

$$G_0(T) = \sum_{g=0}^{T} g f_g \quad \text{and} \quad G_1(T) = \sum_{g=T+1}^{n-1} g f_g.$$

When we include spatial information, such as the local variances discussed above (5), the image entropy can no longer be expressed in terms of the histogram. The class entropies become

$$H_0(T) = -\sum_{i=1 \ (i\in 0)}^{N} g_i/G_0(T) \log \frac{g_i/G_0(T)}{m_i}, \tag{8a}$$

$$H_1(T) = -\sum_{i=1 \ (i\in 1)}^{N} g_i/G_1(T) \log \frac{g_i/G_1(T)}{m_i}. \tag{8b}$$

Kapur et al. (1985) proposed a criterion function made up of the sum of the class entropies. This "total segmentation entropy" is then maximized to determine the optimum threshold, τ, i.e.

$$\tau = \arg\left\{ \max_{0\leqslant T<n-1} \{H_0(T) + H_1(T)\} \right\}. \tag{9}$$

Brink (1991b), on the other hand, suggests that we aim to maximize each class entropy: thus a trade-off is necessary whereby both entropies are as nearly maximized as possible by the choice of threshold. This effectively requires the determination of the *maximin* of the class entropies, i.e. we find that threshold, τ, where the smaller class entropy value is maximized:

$$\tau = \arg\left\{ \max_{0\leqslant T<n-1} \{\min(H_0(T), H_1(T))\} \right\}. \tag{10}$$

At first glance it would appear that (10) and, to some extent, (9) attempt to force the solution to equalize the background and foreground areas. Certainly (10) will in most practical cases effectively determine the threshold as that point which yields equal background and foreground entropies, by virtue of the maximin process. It is necessary to briefly analyze how this works and what effect the use of the weighted class entropies (8) will have, as opposed to the more common measure (7).

In the unweighted case (7) the process is in fact unbiased: (10) attempts to find a threshold at which both background and foreground entropies are maximized as far as possible. The entropy form (7) is maximized when the distribution of p_i is uniform. While the class entropies are independent of the relative background and foreground areas, the process (10) will effectively strive for a solution where the background and foreground are equally uniform, or at least have equal levels of busyness, regardless of the spatial arrangement of the pixels in the resulting regions.

When the weights m_i are introduced (8), the effect is to tailor this process. With the m_i related to local grey-level variation, the individual local contributions to each overall class entropy value are effectively increased in areas of high activity while remaining relatively unchanged in more uniform regions. The overall result when used in (10), and to a slightly lesser extent in (9), is that the image is thresholded in such a way that both regions are maximally uniform while containing similar amounts of gradient or variance in-

Fig. 3. The use of words I, 1928–29, a painting by René Magritte. 1 byte per pixel, 300 × 236 pixels, grey-level range $0 \leqslant g \leqslant 200$.

formation. This acts to "fix" the region boundaries in the segmented image to correspond to areas of high grey-level activity, such as edges, in the original and hence we would expect an aesthetic and, hopefully, practical improvement over the unbiased results. This can be viewed as an attempt to use the power of the maximum entropy method to include essentially *qualitative* prior information.

Some representative results appear in the next section.

4. Results

Two images where bi-level segmentation seemed reasonable are shown in Figs. 3 and 4. For the image of Fig. 3, the results of using the class entropy definitions (7) and (8) with the threshold selection scheme (9) appear respectively in Fig. 5 (a) and (b), while those based on the selection scheme (10) are shown in Figs. 5(c) and 5(d). The corresponding results for the image of Fig. 4 appear in Figs. 6(a)–(d).

5. Discussion

While a quantitative performance evaluation would be desirable, even at a glance the improvement in the bi-level results shown in Figs. 5/6 (b) and (d) over those of Figs. 5/6 (a) and (c) is remarkable. The comparison of the results using the two different threshold selection schemes (9) and (10) is more difficult. Aesthetically the results using (10), as expected, are more "recognizable", but, for example, the use of (9)

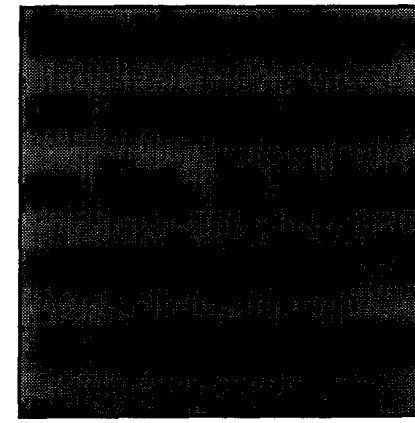

Fig. 4. Video image of a portion of a printed page. 1 byte per pixel, 126 × 126 pixels, grey-level range $33 \leqslant g \leqslant 63$.

on the image of Fig. 3 results in a better segmentation of the region corresponding to the pipe ("object") from its background (Fig. 5 (b)): the light reflected off the pipe is not misclassified as "background" as in Fig. 5 (d). This problem could possibly be rectified by using a different measure of local grey-level activity. For example, when the common Sobel gradient measure (Gonzalez and Woods, 1992) is substituted for the variance in the m_i, the results yielded by (9) and (10) used on the image of Fig. 3 are both identical to Fig. 5 (b). If a thresholded gradient measure were used, whereby strictly high gradient values (edges), as opposed to lower non-zero gradients, were considered, one would expect such details as the light reflected off the pipe in Fig. 3 to be ignored: of course, insofar as the actual grey-levels in such a region may match those of the background, one could still not expect perfect segmentation using a simple global thresholding process.

Quantitative evaluation poses a problem since results such as these must be evaluated using various different and sometimes contradictory criteria: aesthetics, usefulness for a specific application, computational complexity, etc. Albregtsen's (1993) evaluation based on the partitioning of synthetic histograms appears to be useful for the evaluation of methods using histogram-based criterion functions (although one could argue that histogram partitioning is not really the aim of image segmentation). Threshold selection techniques based on the image itself, such as those described here, are more difficult to evaluate. The use of synthetic images where the "correct" result is known

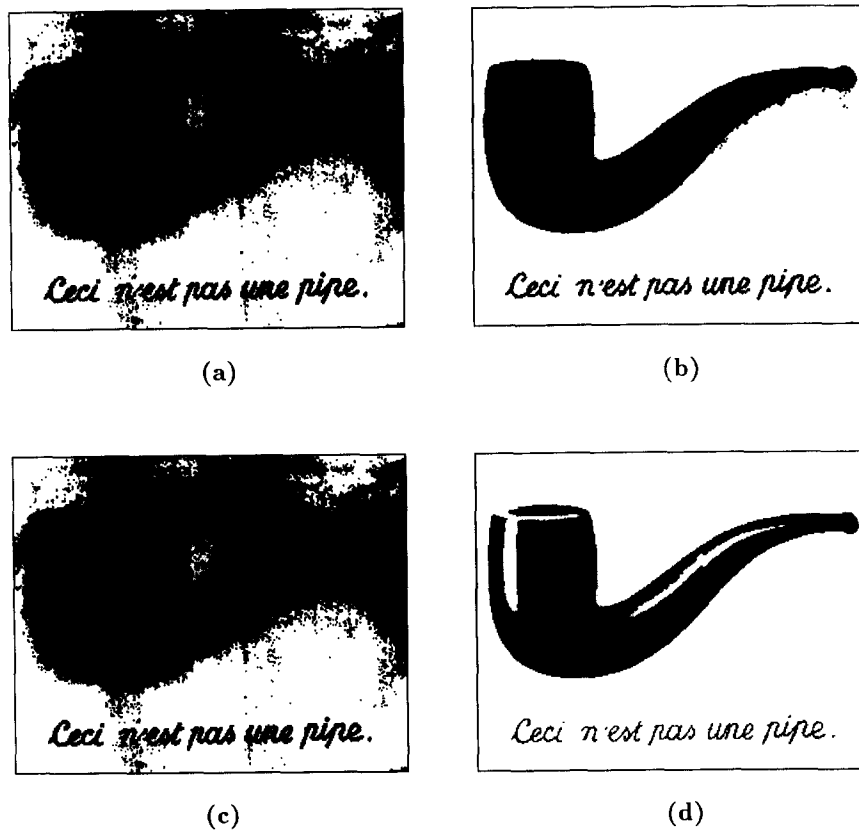

Fig. 5. Results of thresholding the image of Fig. 3: Using the thresholding scheme (9), (a) with class entropies (7) ($\tau = 153$) and (b) with class entropies (8) ($\tau = 120$). Using the scheme (10), (c) with class entropies (7) ($\tau = 153$) and (d) with class entropies (8) ($\tau = 65$).

a priori may be the best route to take: however, care should be taken in the design of the synthetic image and the choice of quality measure used.

What can be concluded from the results is that use of context-related information can drastically improve our results. Such information is available at little additional computational cost: it should not be ignored.

References

Abutaleb, A.S. (1989). Automatic thresholding of gray-level pictures using two-dimensional entropy. *Computer Vision, Graphics, and Image Processing* 47, 22–32.

Albregtsen, F. (1993). Non-parametric histogram thresholding methods – error versus relative object area. *Proc. 8th Scandinavian Conf. on Image Analysis*, Tromsø, Norway, 273–280.

Brink, A.D. (1989). Grey-level thresholding of images using a correlation criterion. *Pattern Recognition Lett.* 9, 335–341.

Brink, A.D. (1991a). Edge-based dynamic thresholding of digital images. *Proc. IEEE COMSIG 91 – A Symposium on Communications and Signal Processing*, Pretoria, South Africa, 63.

Brink, A.D. (1991b). Thresholding of digital images using the entropy of the histogram. *Proc. Second South African Workshop on Pattern Recognition*, Stellenbosch, South Africa, 48–53.

Frieden, B.R. (1972). Restoring with maximum likelihood and maximum entropy. *J. Opt. Soc. Amer.* 62 (4), 511–518.

Frieden, B.R. (1980). Statistical models for the image restoration problem. *Computer Graphics and Image Processing* 12, 40–59.

Gonzalez, R.C. and R.E. Woods (1992). *Digital Image Processing*, Addison-Wesley, Reading, MA.

Gull, S.F. and G.J. Daniell (1978). Image reconstruction from incomplete and noisy data. *Nature* 272, 686–690.

Haralick, R.M. and L.G. Shapiro (1985). Survey: image segmentation techniques. *Computer Vision, Graphics, and Image Processing* 29, 100–132.

Jaynes, E.T. (1968). Prior probabilities. *IEEE Trans. Syst. Sci. Cybernet.* 4 (3), 227–241.

Jaynes, E.T. (1986). Monkeys, kangaroos and *N*. In: J.H. Justice, Ed., *Maximum Entropy and Bayesian Methods in Applied*

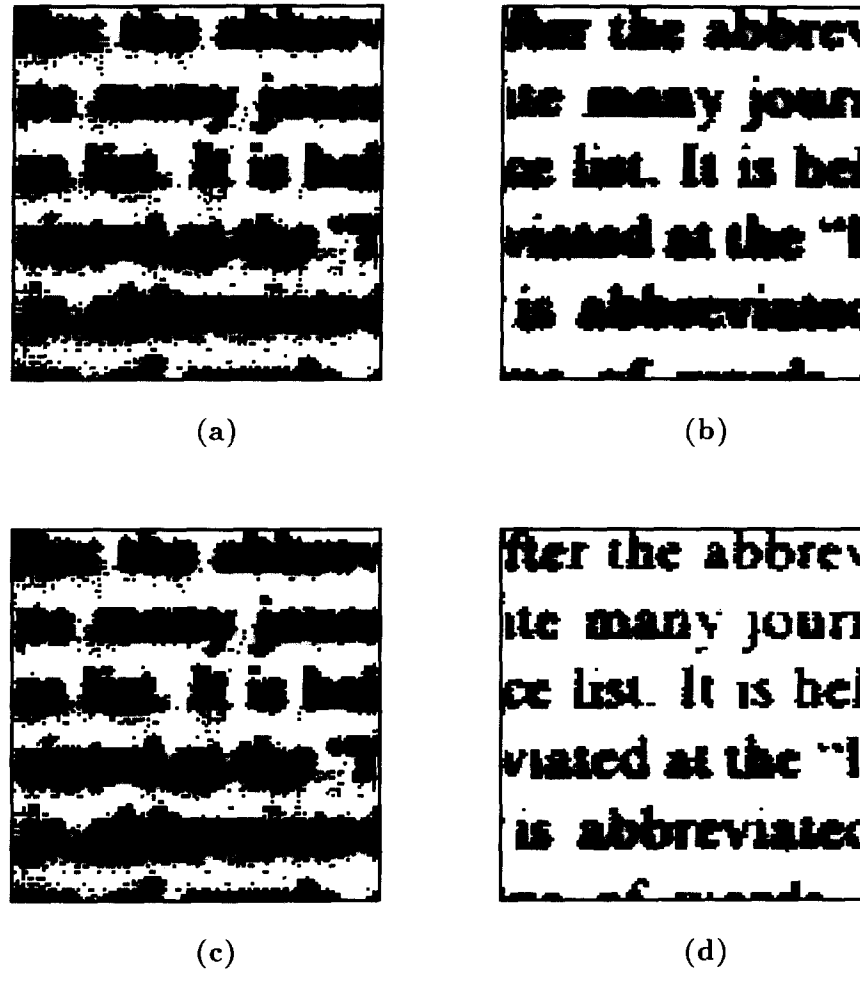

Fig. 6. Results of thresholding the image of Fig. 4: Using the thresholding scheme (9), (a) with entropy definitions (7) ($\tau = 56$) and (b) with entropies (8) ($\tau = 51$). Using the selection scheme (10), (c) with class entropies (7) ($\tau = 56$) and (d) with class entropies (8) ($\tau = 48$).

Statistics. Cambridge Univ. Press, UK, 26–58.

Kapur, J.N., P.K. Sahoo and A.K.C. Wong (1985). A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image Processing* 29, 273–285.

Nakagawa, Y. and A. Rosenfeld (1979). Some experiments on variable thresholding. *Pattern Recognition* 11, 191–204.

Otsu, N. (1979). A threshold selection method from grey-level histograms. *IEEE Trans. Syst. Man Cybernet.* 9 (1), 62–66.

Pal, N.R. and S.K. Pal (1989). Entropic thresholding. *Signal Processing* 16, 97–108.

Sahoo, P.K., S. Soltani, A.K.C. Wong and Y.C. Chen (1988). A survey of thresholding techniques. *Computer Vision, Graphics, and Image Processing* 41, 233–260.

Shannon, C.E. and W. Weaver (1949). *The Mathematical Theory of Communication*. Univ. of Illinois Press, Urbana, IL.

Skilling, J. (1986). Theory of maximum entropy image reconstruction. In: J.H. Justice, Ed., *Maximum Entropy and Bayesian Methods in Applied Statistics*. Cambridge Univ. Press, Cambridge, UK, 156–178.

Skilling, J. (1988). The axioms of maximum entropy. In: G.J. Erickson and C.R. Smith, Eds., *Maximum Entropy and Bayesian Methods in Science and Engineering*, Vol. 1. Kluwer Academic Publishers, Dordrecht, Netherlands, 173–187.

Skilling, J. (1989). Classic maximum entropy. In: J. Skilling, Ed., *Maximum Entropy and Bayesian Methods*. Kluwer Academic Publishers, Dordrecht, Netherlands, 45–52.

Titterington, D.M. (1984). The maximum entropy method for data analysis. *Nature* 312, 381–382.