CVGIP: GRAPHICAL MODELS AND IMAGE PROCESSING
Vol. 55, No. 6, November, pp. 532–537, 1993

NOTE

An Analysis of Histogram-Based Thresholding Algorithms

C. A. GLASBEY

Scottish Agricultural Statistics Service, JCMB, King's Buildings, Edinburgh EH9 3JZ, United Kingdom

Received October 1, 1991; revised June 8, 1993; accepted June 23, 1993

Eleven histogram-based global thresholding algorithms are presented in a common notational framework. Relationships among them are identified from 654 mixtures of two Gaussian distributions, plus effects of mixed pixels. The iterated version of Kittler and Illingworth's minimum error algorithm (*Pattern Recognition*, 19, 1986, 41–47) is found to be best. © 1993 Academic Press, Inc.

1. INTRODUCTION

Image thresholding converts a gray-level image into a binary one. The two binary levels may represent objects and background or, more generally, two classes in an image. Pixels whose value exceeds a critical value are assigned to one category, and the rest to the other. The threshold is global if the same critical value is used across the whole image. Many algorithms have been proposed for automatically selecting the threshold appropriate for a given image (see, for example, Sahoo *et al.* [1]). Some algorithms simply use the histogram of values in the image, that is the numbers of pixels at each gray-level, whereas others use contextual information such as gray-level occurrences in adjacent pixels.

Global, histogram-based algorithms are the most commonly used, despite the benefits that can accrue from using contextual information and allowing the threshold to vary over an image. They are simple to understand and implement, and computationally fast once the histogram has been obtained. However, within this restricted class a plethora of algorithms have been proposed in the last decade, which present a potential user with a bewildering choice. Lee, Chung, and Park [2] compared three histogram-based algorithms with two contextual ones. This paper attempts to be more comprehensive: 11 algorithms are presented in a common notational framework, and their results are compared in the simplest of applications, a mixture of two Gaussian distributions plus mixed pixels.

2. ALGORITHMS

The histogram will be denoted $y_0, y_1, \ldots, y_n$, where y_i is the number of pixels in the image with gray-level i, and n is the maximum gray-level attained (typically 255). The threshold value is an integer, denoted t. All pixel values less than or equal to t are allocated to one category, and those greater than t are allocated to the other.

Prewitt and Mendelsohn [3] suggested that t be chosen as the value of i at which y_i is minimized, in the valley between maxima of y. We denote this threshold MINIMUM. The algorithm assumes a bimodal histogram, which in practice will usually require the y's to be smoothed. A simple way to achieve this is to replace y_i by $(y_{i-1} + y_i + y_{i+1})/3$ for $i = 0, 1, \ldots, n$, with $y_{-1} = y_{n+1} = 0$, and repeat until the y's are bimodal. Then t is such that $y_{t-1} > y_t \leq y_{t+1}$. A simple alternative, INTERMODES, finds the two maxima, say y_j and y_k, and sets $t = (j + k)/2$.

Subsequent algorithms can be computed efficiently if the following partial sums are obtained:

$$A_j = \sum_{i=0}^{j} y_i, \quad B_j = \sum_{i=0}^{j} i y_i, \quad C_j = \sum_{i=0}^{j} i^2 y_i \quad \text{for } j = 0, \ldots, n.$$

These can be derived recursively, for example $A_j = A_{j-1} + y_j$.

In the absence of information on relative proportions in the two binary categories, one possible approach is to choose t such that 50% of pixels lie in each. Thus t is the MEDIAN of the distribution of pixel values, a special case of Doyle's [4] p-tile method: t is chosen so that A_t/A_n is as close as possible to 0.5. The MEAN pixel value is a similar statistic, calculated with t as the integer part of B_n/A_n.

An algorithm due to Tsai [5] chooses t such that the binary image has the same first three MOMENTS as the gray-level image. The threshold is such that A_t/A_n is the value of the fraction nearest to x_0, where

$$x_0 = \frac{1}{2} - \frac{B_n/A_n + x_2/2}{\sqrt{(x_2^2 - 4x_1)}}, \quad x_1 = \frac{B_n D_n - C_n^2}{A_n C_n - B_n^2},$$

$$x_2 = \frac{B_n C_n - A_n D_n}{A_n C_n - B_n^2}, \qquad D_n = \sum_{i=0}^{n} i^3 y_i.$$

1049-9652/93 $5.00

One of several maximum ENTROPY algorithms, due to Kapur *et al.* [6], requires the partial sums

$$E_j = \sum_{i=0}^{j} y_i \log(y_i) \quad \text{for } j = 0, \ldots, n.$$

The expression

$$\frac{E_j}{A_j} - \log A_j + \frac{E_n - E_j}{A_n - A_j} - \log(A_n - A_j)$$

is evaluated for $j = 0, \ldots, n-1$, and t is set to the value of j at which it is maximized.

Ridler and Calvard [7] and Trussell [8] proposed an iterative scheme. From an initial guess for t, say MEAN defined above, the mean gray-levels in the two classes defined by the threshold, that is, below and above it, are calculated as

$$\mu_t = \frac{B_t}{A_t}, \quad \nu_t = \frac{B_n - B_t}{A_n - A_t}.$$

The threshold t is recalculated to be half-way between these means, that is the integer part of $(\mu_t + \nu_t)/2$. Then μ and ν are recalculated, and a new value of t is obtained. This is repeated until convergence, that is a repeat of the same value of t on two consecutive iterations. We denote the result INTERMEANS (I), where (I) denotes iterated. A closely related algorithm, INTERMEANS, due to Otsu [9], requires the between-class sum of squares

$$A_j(A_n - A_j)\,(\mu_j - \nu_j)^2$$

to be evaluated for $j = 0, \ldots, n-1$. The threshold is set to the value of j at which the expression is maximised. Although it is not immediately apparent from the algebraic formulation, this has the effect of positioning the threshold midway between the means of the two classes.

Kittler and Illingworth [10] proposed a similar pair of algorithms, but which allow for possible differences in proportions and variances in the two categories when positioning the threshold. Additional statistics are required:

$$p_t = \frac{A_t}{A_n}, \qquad q_t = \frac{A_n - A_t}{A_n},$$

$$\sigma_t^2 = \frac{C_t}{A_t} - \mu_t^2, \quad \tau_t^2 = \frac{C_n - C_t}{A_n - A_t} - \nu_t^2.$$

In the iterated version of the algorithm, denoted MINERROR(I), an initial guess at t is required (MEAN in our case, although Ye and Danielsson [11] suggested using the result of INTERMEANS(I)). The integer part of the larger solution of the quadratic equation

$$x^2\left\{\frac{1}{\sigma^2} - \frac{1}{\tau^2}\right\} - 2x\left\{\frac{\mu}{\sigma^2} - \frac{\nu}{\tau^2}\right\} + \left\{\frac{\mu^2}{\sigma^2} - \frac{\nu^2}{\tau^2} + \log\left(\frac{\sigma^2 q^2}{\tau^2 p^2}\right)\right\} = 0, \tag{1}$$

then provides a new value for t. Let w_0, w_1, w_2 denote the three terms in curly brackets above, then t is reset to the integer part of

$$[w_1 + \sqrt{(w_1^2 - w_0 w_2)}]/w_0.$$

This minimizes the number of misclassifications between two Gaussian distributions with means, variances, and proportions given, respectively, by μ, ν, σ^2, τ^2, p, and q. (It is equivalent to INTERMEANS(I) only when $\sigma^2 = \tau^2$ and $p = q$.) All terms are recalculated using the new value of t, and then t is rederived. This is repeated until convergence, or until the algorithm fails because a quadratic equation is encountered with no real roots. The uniterated algorithm, MINERROR, is more computationally intensive and requires

$$p_j \log\left(\frac{\sigma_j}{p_j}\right) + q_j \log\left(\frac{\tau_j}{q_j}\right)$$

to be evaluated for $j = 0, \ldots, n-1$. The value of j at which the expression is minimized is chosen for t. Again, the two algorithms are almost equivalent: they both position the threshold to minimise the number of misclassifications.

More complicated methods have been proposed for estimating the terms in Eq. (1). The above estimators are biased because they do not allow for overlaps in distributions from the two classes. Chow and Kaneko [12] and Nagawa and Rosenfeld [13] fitted mixtures of two Gaussians to the y's, while Cho *et al.* [14] employed bias correction factors. The EM algorithm (Dempster *et al.* [15]) is a simple, general method for fitting mixtures of distributions. In this application, from initial estimates of the parameters μ, ν, σ^2, τ^2, p, and q, say from the first iteration of MINERROR(I) with t obtained from MINIMUM,

$$\phi_i = \frac{p}{\sigma}\exp\left[-\frac{(i-\mu)^2}{2\sigma^2}\right] \Big/ \left(\frac{p}{\sigma}\exp\left[-\frac{(i-\mu)^2}{2\sigma^2}\right] + \frac{q}{\tau}\exp\left[\frac{-(i-\nu)^2}{2\tau^2}\right]\right)$$

and $\gamma_i = 1 - \phi_i$ for $i = 0, \ldots, n$

$$F = \sum_{i=0}^{n} \phi_i y_i, \qquad G = \sum_{i=0}^{n} \gamma_i y_i,$$

$$p = F/A_n, \qquad q = G/A_n,$$

$$\mu = \sum_{i=0}^{n} i\,\phi_i y_i/F, \qquad \nu = \sum_{i=0}^{n} i\,\gamma_i y_i/G,$$

$$\sigma^2 = \sum_{i=0}^{n} i^2\,\phi_i y_i/F - \mu^2, \qquad \tau^2 = \sum_{i=0}^{n} i^2\,\gamma_i y_i/G - \nu^2.$$

This sequence of calculations is repeated until the parameters reach stable values. Then Eq. (1) is solved to determine t, denoted MAXLIK.

3. RESULTS

Initially, results were obtained using noise-free histograms, that is without any sampling variability. They were calculated for 972 different mixtures of two Gaussian distributions, together with some mixed pixels. Means were held fixed at

$$\mu = 100, \quad \nu = 151,$$

with $n = 255$. Standard deviations ranged over the values

$$\sigma, \tau = 1, 3, 5, 10, 15, 25,$$

except that cases where $\sigma + \tau \leq 10$ were omitted, because in such cases the two categories are so well separated that thresholding is a trivial operation. The proportions of pixels in the two component distributions were set to

$$p = \rho(1 - r), \quad q = (1 - \rho)(1 - r)$$

where

$$\rho = 0.005, 0.01, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9, 0.95, 0.99, 0.995,$$

and

$$r = 0, 0.1, 0.2.$$

The remaining pixels, proportion r, were allowed to be mixed and drawn from

$$\int_0^1 \frac{1}{\sqrt{(2\pi(z\sigma^2 + (1 - z)\tau^2))}} \exp\left[-\frac{(i - z\mu - (1 - z)\nu)^2}{2(z\sigma^2 + (1 - z)\tau^2)} \right] dz.$$

Of the 972 distributions, 318 were unimodal because the two classes overlapped substantially, such as when $\sigma = \tau = 25$. These were excluded as being unrealistic to threshold. The remaining 654 histograms were bimodal.

TABLE 1
Cases where MINERROR Failed to Find an Internal Threshold

σ	τ	ρ (%)	r (%)
1	25	0.5*, 1, 5	0, 10*, 20
		10	20
3	25	1*, 5	0*, 10*, 20
		10	10, 20
5	25	5, 10	0, 10, 20
10	25	10	0
		20	0, 10, 20
15	15	40	20
	25	40	0, 10, 20

Note. The other half of the cases are when σ and τ are intercharged and ρ is replaced by $1 - \rho$.

* Cases where MAXLIK also failed to find a threshold.

The 11 algorithms from Section 2 were applied to the 654 histograms. In 64 cases MINERROR chose t equal (or close) to 0 or 254, rather than an internal threshold value. Table 1 shows half the cases when this occurred, the other half, by symmetry, being when the values of σ and τ are interchanged and ρ is replaced by $1 - \rho$. To facilitate comparison with other algorithms, in these 64 cases MINERROR was replaced by the result from MINERROR(I), which always converged to an internal threshold value from the MEAN start. In 6 cases, MAXLIK encountered parameter values for which Eq. (1) had no real solution. The cases are starred in Table 1. Their thresholds were also set to the values from MINERROR(I). Results were always obtained from all other algorithms.

ENTROPY generated the widest spread of thresholds, ranging between 73 and 177. All other algorithms gave results between 100 and 150. The least variable method was INTERMODES, for which 75% of the thresholds were exactly 125. All methods had an average threshold of 125, as would be expected from the symmetry in choices of σ, τ, p, and q.

Table 2 shows root-mean-square differences between the methods, averaged over the 654 cases, ordered so that adjacent methods are most similar. Figure 1 is a representation of 74% of the variation in the table. A minimum spanning tree has been added, which connects methods which are most similar. The figure was obtained by a form of reduced-rank regression due to Glasbey [16]. An eigenvector decomposition of the matrix of differences between the methods and the average of all 11 was followed by a regression of these vectors on σ, τ, p, q, r, and products of them. Terms in $(\sigma - \tau)$ and $(p - q)$ were

TABLE 2
Root-Mean-Square Differences between Methods

1	MINIMUM									
2	MAXLIK	6								
3	MINERROR	8	7							
4	MINERROR(I)	14	12	11						
5	INTERMODES	15	15	14	15					
6	INTERMEANS	24	23	21	18	11				
7	INTERMEANS(I)	25	23	22	19	12	5			
8	MOMENTS	28	27	26	24	15	8	8		
9	ENTROPY	29	28	27	27	19	17	17	15	
10	MEAN	26	24	23	20	16	13	11	13	18	
11	MEDIAN	28	26	26	23	21	18	17	19	22	7
		1	2	3	4	5	6	7	8	9	10

Note. Numbers below the columns correspond to those methods listed at left.

found to explain most of the variability, hence their use as the axes in the figure. For example, MINIMUM has bivariate location (0.60, 12), whereas MINERROR(I) is located at (0.46, 2). Therefore the differences between thresholds selected by the two algorithms is approximately

$$(0.60 - 0.46)(\sigma - \tau) + (12 - 2)(p - q).$$

Figure 2 shows two typical histograms, which illustrate the axes in Fig. 1. In Fig. 2a standard deviations in the two Gaussians are different, but proportions are equal,

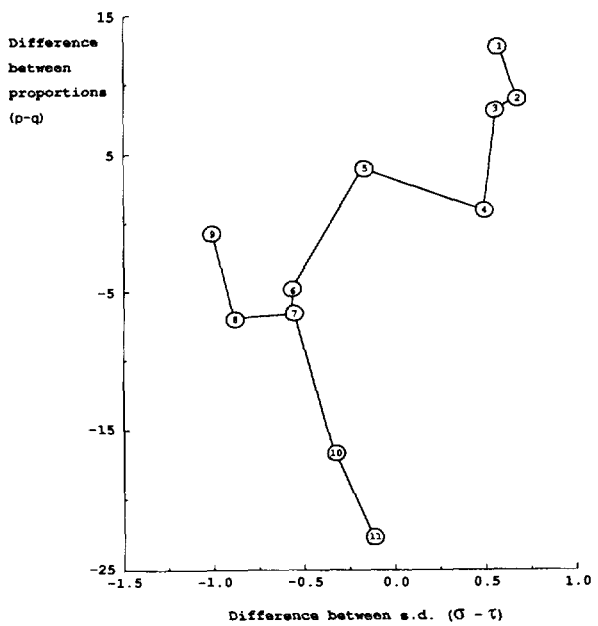

FIG. 1. A two-dimensional representation of the differences among the 11 thresholding algorithms, given in Table 1. Obtained by eigenvector decomposition and reduced rank regression. Lines connect most similar algorithms into a minimum spanning tree.

whereas in Fig. 2b proportions are different but standard deviations are the same. The thresholds selected by the 11 algorithms are ordered approximately the same as along the two axes of Fig. 1.

The inclusion of sampling variability in the histograms has little effect on the previous results. Sampled versions, of size 10,000, were generated for each of the 654 histograms. In one case, the smoothing used to determine MINIMUM and INTERMODES failed as the histogram changed from having three modes to having one in a single iteration. In five cases MINERROR(I) failed to converge, and was set to INTERMEANS. Table 3 shows the variability of each method about its noise-free estimate and the number of occasions on which each method other than ENTROPY gave a threshold outside the range 90 to 160. In order that the standard deviations in Table 3 should not be unduly influenced by these cases, when they occurred MINERROR was replaced by MINERROR(I), MINIMUM and MAXLIK were replaced by MINERROR, and INTERMODES was replaced by INTERMEANS. MEAN and MEDIAN show themselves to be least affected by noise, whereas MINIMUM, which is obtained directly from a smoothed histogram, is most sensitive.

4. DISCUSSION

Much of the pattern in Fig. 1 is as we would expect. INTERMEANS and INTERMEANS(I) are very similar. In fact, if the iterative process had been started at thresholds other than MEAN, then some of them would have produced the same result as INTERMEANS. This is similarly true for MINERROR and MINERROR(I), but in some cases that would have meant $t = 0$ or 254. MEDIAN and MEAN give similar results. INTERMODES gives results central to the others. MAXLIK is most similar to MINIMUM, but it is also close to MINERROR.

ENTROPY and MOMENTS give results on the oppo-

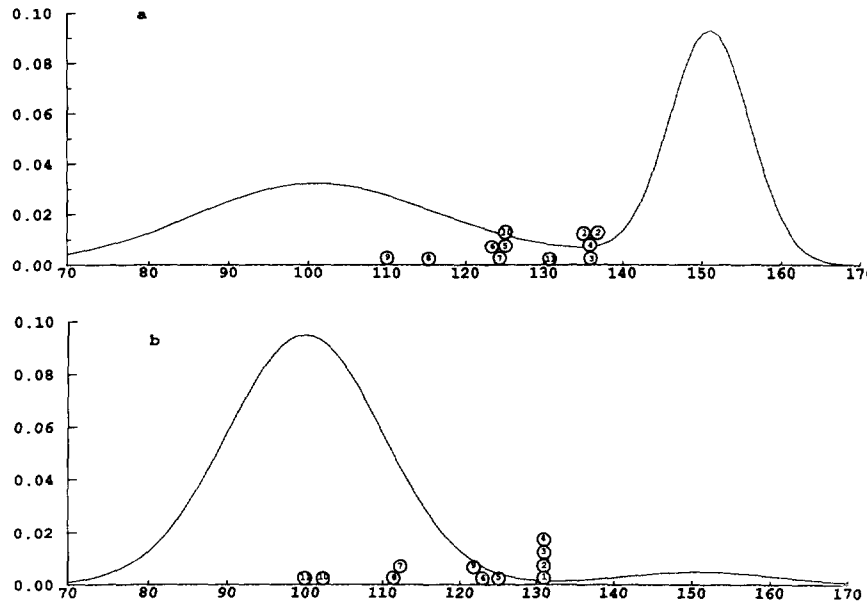

FIG. 2. Histograms of mixtures of two Gaussian distributions, with means 100 and 151, chosen to illustrate the axes in Fig. 1: (a) standard deviations are 15 and 5, proportions are equal, but 10% of values are mixed pixels; (b) standard deviations are both 10, proportions are 95% and 5%. Also shown are the threshold values selected by the 11 algorithms, indexed as in Table 2.

site side of INTERMODES to MINIMUM. Therefore, when MINIMUM lies above 125, they are below it, and conversely. To an extent, good thresholds are a matter for subjective judgement, but this characteristic seems undesirable. It would appear from Fig. 2a that ENTROPY and MOMENTS have chosen thresholds which are too low. Similar arguments would lead to the rejection of MEAN and MEDIAN in cases such as Fig. 2b when proportions are very unequal. INTERMEANS and INTERMEANS(I) tend to split the larger component when proportions are unequal, as noted by Kittler and Illingworth [17]. Algorithms MINIMUM, MAXLIK, MINERROR, and MINERROR(I) all give very similar results. MINERROR(I) would seem to be best. It is one of the simplest to compute, and appears to fail infrequently and to be relatively insensitive to the effects of sampling variability.

In practice, image histograms may be far from being mixtures of two Gaussian distributions. However, it may be somewhat optimistic to hope that a method which performs poorly with Gaussian mixtures would perform well in more complicated situations.

All methods, except MEAN, extend to the selection of two or more thresholds simultaneously. However, MINERROR and INTERMEANS become quite computer intensive and so are best replaced by their iterative forms, possibly with multiple start values. MINERROR(I) can also be extended to multivariate images, where it is equivalent to an adaptive clustering algorithm of Maronna and Jacovkis [18].

TABLE 3
Variability of Thresholds for Sample Size 10,000

		Standard deviation	Number out of range
1	MINIMUM	9.9	28
2	MAXLIK	5.8	23
3	MINERROR	3.2	65
4	MINERROR(I)	3.4	
5	INTERMODES	5.1	8
6	INTERMEANS	1.5	
7	INTERMEANS(I)	1.1	
8	MOMENTS	1.7	
9	ENTROPY	2.9	
10	MEAN	0.9	
11	MEDIAN	0.5	

REFERENCES

1. P. K. Sahoo, S. Soltani, A. K. C. Wong, and Y. C. Chen, A survey of thresholding techniques, *Comput. Vision Graphics Image Process.* **41**, 1988, 233–260.
2. S. U. Lee, S. Y. Chung, and R. H. Park, A comparative performance study of several global thresholding techniques for segmentation, *Comput. Vision Graphics Image Process.* **52**, 1990, 171–190.
3. J. M. S. Prewitt and M. L. Mendelsohn, The analysis of cell images, in *Ann. New York Acad. Sci.*, Vol. 128, pp. 1035–1053, New York Acad. Sci., New York, 1966.
4. W. Doyle, Operation useful for similarity-invariant pattern recognition, *J. Assoc. Comput. Mach.* **9**, 1962, 259–267.

5. W. Tsai, Moment-preserving thresholding: A new approach, *Comput. Vision Graphics Image Process.* **29,** 1985, 377–393.
6. J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, A new method for gray-level picture thresholding using the entropy of the histogram, *Comput. Vision Graphics Image Process.* **29,** 1985, 273–285.
7. T. Ridler and S. Calvard, Picture thresholding using an iterative selection method, *IEEE Trans. Systems Man Cybernet.* **SMC-8,** 1978, 630–632.
8. H. J. Trussell, Comments on "Picture thresholding using an iterative selection method," *IEEE Trans. Systems Man Cybernet.* **SMC-9,** 1979, 311.
9. N. Otsu, A threshold selection method from gray-level histogram, *IEEE Trans. Systems Man Cybernet.* **SMC-8,** 1978, 62–66.
10. J. Kittler and J. Illingworth, Minimum error thresholding, *Pattern Recognition* **19,** 1986, 41–47.
11. Q-Z. Ye and P-E. Danielsson, On minimum error thresholding and its implementations, *Pattern Recognition Lett.* **7,** 1988, 201–206.
12. C. K. Chow and T. Kaneko, Automatic boundary detection of left ventricle from cineangiograms, *Comput. Biomed. Res.* **5,** 1972, 338–410.
13. Y. Nakagawa and A. Rosenfeld, Some experiments on variable thresholding, *Pattern Recognition* **11,** 1979, 191–204.
14. S. Cho, R. Haralick, and S. Yi, Improvement of Kittler and Illingworth's minimum error thresholding, *Pattern Recognition* **22,** 1989, 609–617.
15. A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. Roy. Statist. Soc. Ser. B* **39,** 1977, 1–38.
16. C. A. Glasbey, A reduced-rank regression model for local variation in solar radiation, *Appl. Statist.* **41,** 1992, 381–387.
17. J. Kittler and J. Illingworth, On threshold selection using clustering criterion, *IEEE Trans. Systems Man Cybernet.* **SMC-15,** 1985, 652–655.
18. R. Maronna and P. M. Jacovkis, Multivariate clustering procedures with variable metrics, *Biometrics* **30,** 1974, 499–505.