Survey and comparative analysis of entropy and relative entropy thresholding techniques

C.-I Chang, Y. Du, J. Wang, S.-M. Guo and P.D. Thouin

Abstract: Entropy-based image thresholding has received considerable interest in recent years. Two types of entropy are generally used as thresholding criteria: Shannon's entropy and relative entropy, also known as Kullback–Leibler information distance, where the former measures uncertainty in an information source with an optimal threshold obtained by maximising Shannon's entropy, whereas the latter measures the information discrepancy between two different sources with an optimal threshold obtained by minimising relative entropy. Many thresholding methods have been developed for both criteria and reported in the literature. These two entropy-based thresholding criteria have been investigated and the relationship among entropy and relative entropy thresholding methods has been explored. In particular, a survey and comparative analysis is conducted among several widely used methods that include Pun and Kapur's maximum entropy, Kittler and Illingworth's minimum error thresholding, Pal and Pal's entropy thresholding and Chang *et al.*'s relative entropy thresholding methods. In order to objectively assess these methods, two measures, uniformity and shape, are used for performance evaluation.

1 Introduction

Thresholding is an important technique in image segmentation, enhancement and object detection. Many methods have been reported in the literature [1–5]. Of particular interest is an information theoretic approach that is based on the concept of entropy introduced by Shannon in information theory [6]. The principle of entropy is to use uncertainty as a measure to describe the information contained in a source. The maximum information is achieved when no a priori knowledge is available, in which case, it results in maximum uncertainty. For instance, if an experiment is conducted in an unknown environment that cannot be estimated a priori, a reasonable approach is to assume that all outcomes of the experiment are equally likely to avoid introduction of any possible biased knowledge. Under this situation, the ME is achieved by the maximum uncertainty. This is intuitively appealing from an information theory point of view. In other words, if one has no preference among samples resulting from an experiment, the best decision is not to introduce any biased knowledge into the decision process. Instead, all samples must be treated equally important. In this case, the probability distribution that describes the experiment is either uniformly distributed in continuous probability space or equally likely in discrete probability space, both of which yield the ME.

Using ME as an optimal criterion for image thresholding was first proposed by Pun [7, 8]. It was later corrected and improved by Kapur *et al.* [9]. The concept was further generalised to Renyi's entropy [10]. Basically, the entropy-based thresholding considers an image histogram as a probability distribution, and then selects as an optimal threshold value that yields the ME. More specifically, a best entropy-thresholded image is the one that preserves as much information as possible that is contained in the original unthresholded image in terms of Shannon's entropy. Although such entropy thresholding seems promising, it also suffers from one drawback. It does not take into account the image spatial correlation. Therefore different images with an identical histogram will result in the same threshold value. In order to mitigate this problem, two approaches were proposed in the past. Both extended a one-dimensional (1-D) image histogram to two-dimensional (2-D) image histograms, both of which had taken care of inter-pixel spatial correlation in different ways. One approach was first proposed by Abutaleb [11] who used the original 1-D histogram and its local average to form a 2-D histogram from which a pair of optimal threshold values can be derived. Several extensions to Abutaleb's approach have been investigated [12–16]. Another approach considers the grey-level co-occurrence matrix as a means to capture transitions between grey levels [17]. Unlike Abutaleb's approach that makes use of two separate threshold values, the co-occurrence matrix-based approach requires only one single threshold value. It is known that the co-occurrence matrices are often used in texture analysis. Haralick *et al.* [18] proposed 14 co-occurrence matrix-based texture measures to extract information for texture analysis. On the basis of the concept of the co-occurrence matrix Pal and Pal [19] recently developed two entropy-based thresholding techniques, called local entropy (LE) and joint entropy (JE). They can be viewed as an extension of Pun and Kapur *et al.*'s ME approach where the LE and the JE maximise entropies of local quadrants

IEE Proceedings online no. 20050032

doi:10.1049/ip-vis:20050032

Paper first received 29th January 2005 and in revised form 22nd June 2006

C.-I Chang is with the Remote Sensing Signal and Image Processing Laboratory, Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, MD 21250, USA and is also with the Department of Electrical Engineering, National Chung Hsing University, Taiwan, Republic of China

J. Wang is with the Remote Sensing Signal and Image Processing Laboratory, Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, MD 21250, USA

Y. Du is with the Department of Electrical and Computer Engineering, Purdue School of Engineering and Technology, Indiana University-Purdue University, Indianapolis, IN 46202, USA

S.-M. Guo is with the Department of Information Engineering, National Cheng Kung University, Tainan, Taiwan, Republic of China

P.D. Thouin is with the Department of Defense, Fort Meade, MD, USA

E-mail: cchang@umbc.edu

and joint quadrants resulting from thresholding the co-occurrence matrix, respectively. So, if we consider Pun and Kapur *et al.*'s approach as a first-order entropy thresholding method, Abutaleb's method and Pal and Pal's method can be thought of as second-order entropy thresholding methods.

The entropy-based thresholding methods discussed earlier are derived from maximisation of Shannon's entropy. Relative entropy, also known as Kullback–Leibler information distance, direct divergence or cross entropy, has been also proposed as an alternative thresholding criterion. Two early approaches were minimum error thresholding (MET) developed by Kittler and Illingworth [20] and minimum cross entropy (MCE) developed by Li and Lee [21]. The underlying assumption of Kittler and Illingworth's method is that the image to be thresholded can be modelled by a mixture of two Gaussian distributions with appropriate weights, in which the two Gaussian distributions are used to describe the image background and foreground, respectively, and the weights are determined by the threshold. The desired optimal threshold that produces a two-member Gaussian mixture best matches the original 1-D image histogram where the relative entropy is used as such a matching measure. Minimising relative entropy is equivalent to finding a two-member Gaussian mixture which has the minimal discrepancy between the histogram of thresholded image and the original histogram. This concept was further generalised by Pal and Pal [22], in which the relative entropy and Gaussian mixture model were replaced by the divergence and a Poisson model, respectively. In contrast, Li and Lee's approach considered a constrained thresholding problem with cross entropy used as an optimal criterion. It minimised the cross entropy subject to two constraints that the means of foreground and background must remain unchanged before and after thresholding. Unfortunately, it was shown that the MCE used in Li and Lee's method was not actually cross entropy [23].

More recently, Chang *et al.* [24] developed an alternative relative entropy thresholding method that also used the relative entropy as a threshold criterion. Instead of using the image histogram as the way considered in Kittler and Illingworth's MET and Li and Lee's MCE, their approach used the co-occurrence matrix and minimised the discrepancy of grey-level transitions in the co-occurrence matrix before and after an image was thresholded. Conceptually, what Pal and Pal's approach was to Pun and Kapur *et al.*'s entropy thresholding method is exactly what Chang *et al.*'s relative entropic thresholding method was to Kittler and Illingworth's MET and Li and Lee's MCE. In other words, Kittler and Illingworth's MET, Li and Lee's MCE and Pun and Kapur *et al.*'s method can be considered as first-order entropy-based thresholding methods, as they only deal with the 1-D image histogram as opposed to Pal and Pal's and Chang *et al.*'s methods which can be considered as second-order entropy-based methods due to the use of the 2-D co-occurrence matrix. The crucial difference between entropy thresholding and relative entropy thresholding is that the former maximises Shannon's entropy, whereas the latter minimises relative entropy. Chang *et al.*'s approach was further improved in the work of Lee *et al.* [25] and was also extended to Ali-Silvey distance measures in the work of Ramac and Varshney [26]. In analogy with the idea that Pal and Pal extended Pun's ME approach to local entropy and joint entropy methods, Lee *et al.*'s also extended Chang *et al.*'s relative entropy approach to local relative entropy (LRE) and joint relative entropy (JRE) methods. Interestingly, their derived LRE and JRE were not actually relative entropy, a similar error that was made in Li and Lee's MCE [23]. Nonetheless, like Li and Lee's MCE, the LRE and JRE proposed by Lee *et al.* [27] were also demonstrated to be reasonable good criteria.

In this paper, we investigate the entropy-based and relative entropy-based thresholding criteria and explore relationship among entropy and relative entropy thresholding methods. In particular, we conduct a comparative study and analysis between entropy-based and relative entropy-based thresholding methods. Three new thresholding methods, global entropy (GE), LRE and JRE are also introduced, where the LRE and JRE are correct versions of the LRE and JRE proposed by Lee *et al.* [27]. Interestingly, Chang *et al.*'s [24] method can be reinterpreted in this paper as global relative entropy (GRE), which complements the LRE and JRE. With these interpretations, the three relative entropy thresholding methods, GRE, LRE and JRE can be considered as counterparts of GE, LE and JE in entropy thresholding methods. As many popular first-order thresholding methods have been surveyed and compared in the work of Yang *et al.* [1] as well as Abutaleb's 2-D histogram-based approaches were discussed in the work of Yang *et al.* [14], there is no need to repeat their work here. Instead, this paper is primarily focused on a comparative study and analysis among Kittler and Illingworth's MET, the three co-occurrence matrix-based entropy thresholding techniques and three relative entropy thresholding methods plus Otsu's [28] method. The reason of including Otsu's method in our study is because this method has been widely used and proved to be one of the most successful techniques in image thresholding. It should be noted that Pun and Kapur *et al.*'s methods and Li and Lee's MCE are not included in our study. The former was shown not comparable with Pal and Pal's method and the latter performed very poorly in most of our experiments. In addition, two objective measures, uniformity and shape, suggested in Sahoo *et al.* [1] are introduced to evaluate their comparative performance.

2 Entropy thresholding

The concept of entropy has been widely used in data compression to measure information content of an information source. Suppose that a source X has L source alphabets denoted by $\{x_i\}_{i=1}^{L}$ and the probability of the ith source alphabet x_i is given by p_i. In this case, a source can be specified by a probability vector $\boldsymbol{p} = (p_1, \ldots, p_L)$, where p_i is the probability of x_i. For each source symbol x_i for $1 \leq i \leq L$, we can define the so-called self-information of x_i as $I(x_i) = -\log(p_i)$ [29, 30]. Such self-information $I(x_i)$ describes how much information or uncertainty produced by a particular source alphabet x_i. Furthermore, because the significance of each source alphabet is also determined by its occurrence generated by the source X, the probability of each source alphabet must be factored in the description of the information for X. As a consequence, an effective means to describe the information for the source X is the mean of self-information over the L source alphabets $\{x_i\}_{i=1}^{L}$, which turns out to be $E_X[I(X)]$. However, if we expand the expression of $E_X[I(X)]$ as follows, $E_X[I(X)]$ becomes the well-known entropy.

$$H(X) = E_X[I(X)] = E_X[-\log(X)] = \sum_{i=1}^{L} p(x_i)I(x_i)$$

$$= \sum_{i=1}^{L} p_i[-\log(p_i)] = -\sum_{j=1}^{L} p_j \log p_j \qquad (1)$$

As an image can be viewed as an information source with a probability vector described by its grey-level image histogram, the entropy of the histogram can be used to represent a certain level of information contained in the image. Pun [7, 8] and Kapur *et al.* [9] had taken this concept to derive entropy thresholding methods, that will be referred to as ME methods. However, their approaches did not take into account the correlation among grey levels. As a result, two different images with an identical image histogram will result in the same threshold value. One way to resolve this problem is to consider the grey-level co-occurrence matrix defined in the following section, which contains the information of grey-level transitions in an image. Two approaches have been investigated in the past. One is Abutaleb's 2-D histogram, that takes advantage of the correlation between a grey level value and its local average to capture image spatial correlation. Another is the co-occurrence matrix that records transitions between every pair of grey levels in an image histogram. As Abutaleb's 2-D histogram-based approaches require two separate threshold values that are not in our scope, they will not be discussed here. Instead, we will primarily focus in this paper on single threshold value-based approaches.

2.1 Grey-level co-occurrence matrix

Assume that an image has a size of $M \times N$ with L grey levels denoted by $G = \{0, 1, \ldots, L-1\}$. Let $f(x, y)$ be the grey level of the pixel at the spatial location (x, y). Then the image can be represented by an $M \times N$ matrix $\boldsymbol{F} = [f(x, y)]_{M \times N}$. A 1-D image histogram resulting from $f(x, y)$ and the image matrix $\boldsymbol{F}$ is a distribution of the L grey levels $G = \{0, 1, \ldots, L-1\}$ in accordance with the frequency of their occurrence. Unfortunately, such a 1-D histogram discards the correlation among grey levels, which is crucial in image thresholding and segmentation. In order to resolve this issue, a 2-D histogram that can describe and capture image correlation is necessary to improve thresholding performance. One such approach is the use of co-occurrence matrix.

A co-occurrence matrix of an image is an $L \times L$ square matrix, denoted by $\boldsymbol{W} = [t_{ij}]_{L \times L}$ whose elements are specified by the numbers of transitions between all pairs of grey levels in $G = \{0, 1, \ldots, L-1\}$ in a particular way. For each image pixel at spatial co-ordinate (m, n) with its grey level specified by $f(m, n)$, it considers its nearest four neighbouring pixels at locations of $(m-1, n)$, $(m+1, n)$, $(m, n-1)$, $(m, n+1)$ and referred to as the 4-adjacency in the work of Gonzalez and Woods [17]. The co-occurrence matrix developed by Haralick *et al.* [18] is designed to dictate the grey level changes by comparing its grey level $f(m, n)$ to their corresponding grey levels, $f(m-1, n)$, $f(m+1, n)$, $f(m, n-1)$, $f(m, n+1)$. It has been shown that there is no significant difference between considering all the four neighbouring pixels and using only two neighbouring pixels at the horizontal and vertical directions in the 4-adjacency of a pixel. One widely used co-occurrence matrix is an asymmetric matrix that only considers the grey level transitions between two adjacent pixels. More specifically, let t_{ij} be the (i, j)th element of the co-occurrence matrix $\boldsymbol{W}$. Following the definition given in the work of Chang *et al.* [24]

$$t_{ij} = \sum_{m=1}^{M} \sum_{n=1}^{N} \delta_{mn} \tag{2}$$

and

$$\delta_{mn} = 1 \quad \text{if} \begin{cases} f(m, n) = i \text{ and } f(m+1, n) = j \\ \text{and/or} \\ f(m, n) = i \text{ and } f(m, n+1) = j \end{cases}$$
$$= 0; \quad \text{otherwise}$$

where 'and/or' used in the δ_{mn} defined earlier implies 'either or both'.

Normalising the total number of transitions in the co-occurrence matrix, a desired transition probability from grey level i to grey level j is obtained by

$$p_{ij} = \frac{t_{ij}}{\sum_{k=0}^{L-1} \sum_{l=0}^{L-1} t_{kl}} \tag{3}$$

For more details on co-occurrence matrix, we refer to previous studies [3, 17, 18].

2.2 Quadrants of the co-occurrence matrix

Let t be a value used to threshold an image. It partitions a co-occurrence matrix into four quadrants, namely, A, B, C and D, shown in Fig. 1. These four quadrants can be further grouped into two classes, referred to as local quadrants and joint quadrants. We assume that pixels with grey levels above the threshold are assigned to the foreground (corresponding to objects), and those equal to or below the threshold are assigned to the background. Then quadrants A and C correspond to local transitions within background and foreground, respectively, whereas quadrants B and D are joint quadrants which represent joint transitions across boundaries between background and foreground. The probabilities associated with each quadrant are then given by

$$P_A^t = \sum_{i=0}^{t} \sum_{j=0}^{t} p_{ij}, \quad P_B^t = \sum_{i=0}^{t} \sum_{j=t+1}^{L-1} p_{ij},$$
$$P_C^t = \sum_{i=t+1}^{L-1} \sum_{j=0}^{t} p_{ij}, \quad P_D^t = \sum_{i=t+1}^{L-1} \sum_{j=t+1}^{L-1} p_{ij} \tag{4}$$

The probabilities of grey-level transition within each particular quadrant can be further obtained by so called 'cell probabilities'

$$p_{ij|A}^t = \frac{p_{ij}}{P_A^t}, \quad p_{ij|B}^t = \frac{p_{ij}}{P_B^t}, \quad p_{ij|C}^t = \frac{p_{ij}}{P_C^t}, \quad p_{ij|D}^t = \frac{p_{ij}}{P_D^t} \tag{5}$$

2.3 LE, JE and GE methods

Three entropies can be derived on the basis of the cell probabilities defined by (4) and (5), each of which yields a different measure. The first two were proposed by Pal and Pal [19], which are called LE and JE. The third one is a new definition, which will be referred to as GE.

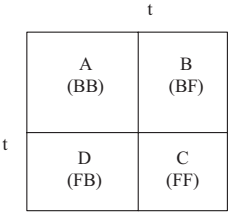

Fig. 1 *Four quadrants of a co-occurrence matrix*

2.3.1 Local entropy: As local quadrants A and C contain local transitions from background to background (BB) and foreground to foreground (FF), respectively, the local transition entropy of BB, denoted by $H_{BB}(t)$ and the local transition entropy of FF, denoted by $H_{FF}(t)$ can be defined, respectively.

$$H_{BB}(t) = -\sum_{i=0}^{t}\sum_{j=0}^{t} p_{ij|A}^{t} \log p_{ij|A}^{t} \tag{6}$$

$$H_{FF}(t) = -\sum_{i=t+1}^{L-1}\sum_{j=t+1}^{L-1} p_{ij|C}^{t} \log p_{ij|C}^{t} \tag{7}$$

where both $H_{BB}(t)$ and $H_{FF}(t)$ are determined by the threshold t, thus they are function of t.

By summing up the local transition entropies of foreground and background, Pal and Pal derived so-called LE, denoted by $H_{LE}(t)$ as follows.

$$H_{LE}(t) = H_{BB}(t) + H_{FF}(t) \tag{8}$$

Obviously, $H_{LE}(t)$ describes the grey-level transitions entropy of the local quadrants A and C. Thus, it is more precisely to be called 'local transition entropy' to reflect the characteristic of quadrants A and C. The LE method proposed by Pal and Pal [19] finds a grey level value specified by

$$t_{LE} = \arg\left\{ \max_{t\in G=\{0,1,\ldots,L-1\}} H_{LE}(t) \right\} \tag{9}$$

which maximises $H_{LE}(t)$ defined by (8) over t.

2.3.2 Joint entropy: Alternatively, the joint quadrants B and D provide edge information about joint transitions from background to foreground (BF) and foreground to background (FB). In analogy with LE, another entropy, called JE, $H_{JE}(t)$ was also derived by Pal and Pal, which is the sum of the joint transition entropy $H_{FB}(t)$ resulting from the joint quadrant B, and the joint transition entropy $H_{BF}(t)$ from the joint quadrant D and is defined as follows.

$$H_{FB}(t) = -\sum_{i=t+1}^{L-1}\sum_{j=0}^{t} p_{ij|B}^{t} \log p_{ij|B}^{t} \tag{10}$$

$$H_{BF}(t) = -\sum_{i=0}^{t}\sum_{j=t+1}^{L-1} p_{ij|D}^{t} \log p_{ij|D}^{t} \tag{11}$$

$$H_{JE}(t) = H_{BF}(t) + H_{FB}(t) \tag{12}$$

Similarly, $H_{JE}(t)$ is more accurately to be called 'joint transition entropy' to reflect the grey-level transition activities in the joint quadrants B and D. A method of finding t_{JE} that maximises $H_{JE}(t)$ defined by (12) over t is called the JE method, which is

$$t_{JE} = \arg\left\{ \max_{t\in G=\{0,1,\ldots,L-1\}} H_{JE}(t) \right\} \tag{13}$$

2.3.3 Global entropy: The GE $H_{GE}(t)$ defined below is simply the sum of the LE $H_{LE}(t)$ and the JE $H_{JE}(t)$, that is

$$\begin{aligned} H_{GE}(t) &= H_{LE}(t) + H_{JE}(t) = H_{BB}(t) + H_{FF}(t) \\ &\quad + H_{BF}(t) + H_{FB}(t) \end{aligned} \tag{14}$$

Finding a value, t_{GE} that maximises $H_{GE}(t)$ defined by (14) over t via the following equation

$$t_{GE} = \arg\left\{ \max_{t\in G=\{0,1,\ldots,L-1\}} H_{GE}(t) \right\} \tag{15}$$

is called the GE threshold method. It should be noted that the GE defined by (14) was not defined by Pal and Pal [19]. However, it turns out to be a counterpart of Chang *et al.*'s [24] GRE. Because GE is the sum of LE and JE, it can be expected that the performance based on GE will be moderate between LE and JE. The experiments seem to justify our claim. So, when it is uncertain about which one should be chosen, GE may be a good candidate for a compromise.

3 Relative entropy thresholding

Relative entropy has been used to measure the information distance between two information sources. The smaller the relative entropy is, the closer the two sources are in terms of their probability distributions. As described in the beginning of Section 2, a source can be specified by a probability vector. Now, assume that there are two sources, X and Y, each of which has L source alphabets. Let X and Y be specified by the probability vectors $\boldsymbol{p} = (p_1, \ldots, p_L)$ and $\boldsymbol{h} = (h_1, \ldots, h_L)$, respectively. The relative entropy between two sources X and Y via their respective probability vectors $\boldsymbol{p}$ and $\boldsymbol{h}$ (or the entropy of $\boldsymbol{h}$ relative to $\boldsymbol{p}$), denoted by $J(\boldsymbol{p}; \boldsymbol{h})$ is defined by

$$J(\boldsymbol{p}; \boldsymbol{h}) = \sum_{j=0}^{L-1} p_j \log \frac{p_j}{h_j} \tag{16}$$

The definition given by (16) was first introduced by Kullback [29] as an information distance measure between two probability distributions. It is called Kullback–Leiber's information discriminant measure, and is also known as cross entropy and directed divergence. It implies that the smaller the relative entropy, the less the discrepancy between $\boldsymbol{p}$ and $\boldsymbol{h}$, thus, the better the match between the two probability vectors. Relative entropy can be used to measure the distance between an image and a thresholded image. It is worth noting that the relative entropy is not symmetric, that is $J(\boldsymbol{p}; \boldsymbol{h}) \neq J(\boldsymbol{h}; \boldsymbol{p})$. In this paper, the original image is always designated as the nominal image $\boldsymbol{p}$, and the thresholded image is $\boldsymbol{h}$, the one which tries to match the original image.

3.1 Kittler and Illingworth's MET

The concept of using relative entropy as a thresholding criterion was first suggested by Kittler and Illingworth [20], in which they assumed that an image could be modelled by a mixture of two Gaussian distributions, which can be used to describe background and foreground, respectively. More specifically, let $\boldsymbol{p}_{true} = (p_{0|true}, p_{1|true}, \ldots, p_{L-1|true})$ be an image histogram. Assume that t is a threshold value used to segment the image into background and foreground, both of which are also modelled by Gaussian distributions, $\boldsymbol{p}_B(t)$ and $\boldsymbol{p}_F(t)$, respectively. Define $\boldsymbol{p}_{mix}(t)$ as a mixture of these two Gaussian distributions by

$$\boldsymbol{p}_{mix}(t) = \alpha \boldsymbol{p}_B(t) + (1-\alpha)\boldsymbol{p}_F(t) \tag{17}$$

where α is determined by the portions of background and foreground in the image. Kittler and Illingworth's MET finds a grey level value t_{MET} that minimises the mismatch between $\boldsymbol{p}_{true}$ and $\boldsymbol{p}_{mix}(t)$ over t, that is

$$t_{mix} = \arg\left\{ \min_{t\in G=\{0,1,\ldots,L-1\}} J(\boldsymbol{p}_{true}; \boldsymbol{p}_{mix}(t)) \right\} \tag{18}$$

where $J(\boldsymbol{p}; \boldsymbol{p}_{mix}(t))$ is the relative entropy between $\boldsymbol{p}$ and $\boldsymbol{p}_{true}$ defined by (16) to measure the discrepancy between the two probability vectors, $\boldsymbol{p}$ and $\boldsymbol{p}_{true}$. As expected, if the background

and foreground are well separated in terms of grey levels, Kittler and Illingworth's MET may work well. Unfortunately, this assumption is generally not true in many practical applications. Pal and Pal [22] also proposed a Poisson model approach to improve the Gaussian model in MET.

3.2 Grey-level co-occurrence matrix used for relative entropic thresholding

As noticed in Kittler and Illingworth's MET, their method is based solely on the grey level histogram of an image which has not taken into consideration the correlation among grey levels. This leads to an idea of using co-occurrence matrix to extend the MET, called second-order relative entropy as opposed to the MET referred to as first-order relative entropy. In this case, the $\boldsymbol{p} = (p_1, \ldots, p_L)$ and $\boldsymbol{h} = (h_1, \ldots, h_L)$ defined in (16) are replaced, respectively, by the grey-level transition probabilities $\{p_{ij}\}_{i=0,j=0}^{L-1,L-1}$ generated by the co-occurrence matrix of the original image and the grey-level transition probabilities, $\{h_{ij}^t\}_{i=0,j=0}^{L-1,L-1}$ generated by the co-occurrence matrix of a thresholded image. The transition probabilities defined by the co-occurrence matrix contain the spatial information that reflects homogeneity of local grey-level transitions in quadrants A and C, and joint grey-level transitions across boundaries in joint quadrants B and D.

Let the second-order relative entropy of the grey-level transition probabilities $\{p_{ij}\}_{i=0,j=0}^{L-1,L-1}$ and $\{h_{ij}^t\}_{i=0,j=0}^{L-1,L-1}$ be defined by

$$J(\{p_{ij}\};\{h_{ij}^t\}) = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} p_{ij} \log \frac{p_{ij}}{h_{ij}^t} \tag{19}$$

where p_{ij} are the transition probabilities from grey level i to grey level j of the original image and h_{ij}^t is the transition probability generated by the thresholded binary image in response to p_{ij}. Despite the fact that the thresholded binary image has only grey level values of 0 (background) and 1 (foreground), it should be noted that the subscript of h_{ij}^t, ij corresponds to the same ij used as the subscript of p_{ij}. Using (19) as a thresholding criterion to minimise $J(\{p_{ij}\};\{h_{ij}^t\})$ over t generally renders a thresholded binary image that best matches the original image.

Suppose that a threshold value t is selected for binarisation. By assigning 1 to all grey levels above t, $G_1 = \{t+1, \ldots, L-1\}$ and 0 to all grey levels equal to or below t, $G_0 = \{0, \ldots, t\}$, we obtain a binary image. Further assume that the grey levels in G_0 and G_1 are uniformly distributed in their respective regions. The resulting h_{ij}^t for each quadrant can be found by

$$h_{ij|A}^t = q_A^t = \frac{P_A^t}{(t+1)(t+1)} \quad \text{for } i,j \in G_0 \tag{20}$$

$$h_{ij|B}^t = q_B^t = \frac{P_B^t}{(t+1)(L-t-1)} \quad \text{for } i \in G_0 \text{ and } j \in G_1 \tag{21}$$

$$h_{ij|C}^t = q_C^t = \frac{P_C^t}{(L-t-1)(L-t-1)} \quad \text{for } i \in G_1 \text{ and } j \in G_1 \tag{22}$$

$$h_{ij|D}^t = q_D^t = \frac{P_D^t}{(L-t-1)(t+1)} \quad \text{for } i \in G_1 \text{ and } j \in G_0 \tag{23}$$

where P_A^t, P_B^t, P_C^t and P_D^t were defined by (4). For each selected t, $h_{ij|A}^t$, $h_{ij|B}^t$, $h_{ij|C}^t$ and $h_{ij|D}^t$ are constants in each individual quadrant and they only depend upon which quadrants they belong to. Therefore they can be simplified by q_A^t, q_B^t, q_C^t and q_D^t, respectively, which represent conditional probabilities of each of four quadrants produced by h_{ij}^t.

3.3 Three relative entropy-based methods

Expanding (19) yields

$$\begin{aligned} J(\{p_{ij}\};\{h_{ij}^t\}) &= \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} p_{ij} \log \frac{p_{ij}}{h_{ij}^t} \\ &= -H(\{p_{ij}\}) - \sum_{i,j} p_{ij} \log h_{ij}^t \end{aligned} \tag{24}$$

where $H(\{p_{ij}\})$ is the entropy of the probability vector specified by $\{p_{ij}\}_{i=0,j=0}^{L-1,L-1}$ and is independent of t. As the relative entropy, $J(\{p_{ij}\};\{h_{ij}^t\})$ in (29) measures the discrepancy between two probability vectors specified by $\{p_{ij}\}_{i=0,j=0}^{L-1,L-1}$ and $\{h_{ij}^t\}_{i=0,j=0}^{L-1,L-1}$, which describe the original image, and thresholded image, respectively. So, the smaller the $J(\{p_{ij}\};\{h_{ij}^t\})$, the better the approximation of $\{p_{ij}\}_{i=0,j=0}^{L-1,L-1}$ to $\{h_{ij}^t\}_{i=0,j=0}^{L-1,L-1}$. Therefore the best threshold will be the one that yields the smallest value of $J(\{p_{ij}\};\{h_{ij}^t\})$. However, minimising $J(\{p_{ij}\};\{h_{ij}^t\})$ in the left-hand-side of (24) is equivalent to maximising the second term of the right-hand-side of (24), $\sum_{i,j} p_{ij} \log h_{ij}^t$ which can be further reduced to

$$P_A^t \log q_A^t + P_B^t \log q_B^t + P_C^t \log q_C^t + P_D^t \log q_D^t \tag{25}$$

So, in order to minimise (24) over t, we only have to maximise (25) over t. In analogy with Section 2.3, three different relative entropies can be defined via (25).

3.3.1 GRE thresholding: Equation (25) is identical to the one proposed by Chang *et al.* [24] and is referred to as GRE, $H_{\text{GRE}}(t)$ here and is expressed as follows

$$\begin{aligned} H_{\text{GRE}}(t) = -(&P_A^t \log q_A^t + P_B^t \log q_B^t + P_C^t \log q_C^t \\ &+ P_D^t \log q_D^t) \end{aligned} \tag{26}$$

It describes the global feature of grey-level transitions in the image. So, the GE defined by (14) in entropy thresholding can be viewed as its counterpart. Finding a threshold value t_{GRE} that minimises (26) is called GRE thresholding method, that is

$$t_{\text{GRE}} = \arg\left\{\min_{t \in G} H_{\text{GRE}}(t)\right\} \tag{27}$$

3.3.2 LRE thresholding: Analogous to Pal and Pal's LE, we can also define its counterpart in relative entropy, called LRE via (26). It was originally proposed by Lee *et al.* [27] in which, $\{P_A^t, P_C^t\}$ did not constitute a probability distribution. In order to make it a probability distribution, extra care must be taken by normalising the probabilities in the local quadrants A and C. If we define $p_{ij|AC}^t = p_{ij}/(P_A^t + P_C^t)$, then the correct version of LRE is given by

$$\begin{aligned} J_{\text{LRE}}(\{p_{ij|AC}\};h_{ij}^t) &= \sum_{(i,j)\in \text{BB}\cup\text{FF}} p_{ij|AC} \log \frac{p_{ij|AC}}{h_{ij}^t} \\ &= -H_{\text{BB+FF}}(t) - \sum_{(i,j)\in \text{BB}\cup\text{FF}} p_{ij|AC} \log h_{ij}^t \end{aligned} \tag{28}$$

where

$$H_{\mathrm{BB+FF}}(t) = -\sum_{(i,j)\in \mathrm{BB}\cup\mathrm{FF}} p_{ij|AC} \log p_{ij|AC} \tag{29}$$

is the entropy of local quadrants A and C in the co-occurrence matrix $\boldsymbol{W}$. The second term in (28) can be further reduced to

$$\begin{aligned}
&\sum_{(i,j)\in \mathrm{BB}\cup\mathrm{FF}} p_{ij|AC} \log h_{ij}^t \\
&= \sum_{(i,j)\in BB} p_{ij|AC} \log\left(\frac{q_A^t}{P_A^t + P_C^t}\right) \\
&\quad + \sum_{(i,j)\in FF} p_{ij|AC} \log\left(\frac{q_C^t}{P_A^t + P_C^t}\right) \\
&= \frac{P_A^t}{P_A^t + P_C^t} \log\left(\frac{q_A^t}{P_A^t + P_C^t}\right) + \frac{P_C^t}{P_A^t + P_C^t} \log\left(\frac{q_C^t}{P_A^t + P_C^t}\right)
\end{aligned} \tag{30}$$

Substituting (30) into (28) results in

$$\begin{aligned}
J_{\mathrm{LRE}}(\{p_{ij|AC}\}; h_{ij}^t) = -H_{\mathrm{BB+FF}}(t) - \Bigg[&\frac{P_A^t}{P_A^t + P_C^t} \log\left(\frac{q_A^t}{P_A^t + P_C^t}\right) \\
&+ \frac{P_C^t}{P_A^t + P_C^t} \log\left(\frac{q_C^t}{P_A^t + P_C^t}\right)\Bigg]
\end{aligned} \tag{31}$$

It should be noted that $H_{\mathrm{BB+FF}}(t)$ given by (29) is different from $H_{\mathrm{LE}}(t)$ given by (8), in the sense that the former considers quadrants A and C as an entity and normalises probabilities to unity, whereas the latter considers quadrants A and C as separate individual entities and normalises their probabilities in two different quadrants A and C to unity separately. Interestingly, the $J_{\mathrm{LRE}}(\{p_{ij|AC}\}; h_{ij}^t)$ in (31) captures the local features of grey-level transitions within background and foreground that can be expressed by $-H_{\mathrm{BB+FF}}(t)$ minus an extra term given by (30). The LRE thresholding method is to find a threshold value t_{LRE} that minimises $J_{\mathrm{LRE}}(\{p_{ij|AC}\};h_{ij}^t)$, that is

$$t_{\mathrm{LRE}} = \arg\left\{\min_{t\in G} J_{\mathrm{LRE}}(\{p_{ij|AC}\}; h_{ij}^t)\right\} \tag{32}$$

3.3.3 JRE thresholding: The JE has also its counterpart, JRE in relative entropy, which measures the information of joint features of grey-level transitions from background to foreground and foreground to background. Like the LRE, the JRE defined by Lee *et al.* [27] was not correct in terms of probability distribution. Analogous to (32), a normalisation factor $p_{ij|BD} = p_{ij}/(P_B^t + P_D^t)$ must be included to normalise the probabilities in the joint quadrants B and D. The correct JRE is given by

$$\begin{aligned}
J_{\mathrm{JRE}}(\{p_{ij|BD}\}; h_{ij}^t) &= \sum_{(i,j)\in \mathrm{BF}\cup\mathrm{FB}} p_{ij|BD} \log\frac{p_{ij|BD}}{h_{ij}^t} \\
&= -H_{\mathrm{BF+FB}}(t) - \sum_{(i,j)\in \mathrm{BF}\cup\mathrm{FB}} p_{ij|BD} \log h_{ij}^t
\end{aligned} \tag{33}$$

$$\begin{aligned}
J_{\mathrm{JRE}}(\{p_{ij|BD}\}; h_{ij}^t) = -H_{\mathrm{BF+FB}}(t) - \Bigg[&\frac{P_B^t}{P_B^t + P_D^t} \log\left(\frac{q_B^t}{P_B^t + P_D^t}\right) \\
&+ \frac{P_D^t}{P_B^t + P_D^t} \log\left(\frac{q_D^t}{P_B^t + P_D^t}\right)\Bigg]
\end{aligned} \tag{34}$$

where

$$H_{\mathrm{BF+FB}}(t) = -\sum_{(i,j)\in \mathrm{BF}\cup\mathrm{FB}} p_{ij|BD} \log p_{ij|BD} \tag{35}$$

is the entropy of the joint quadrants B and D in the co-occurrence matrix $\boldsymbol{W}$. So, finding a threshold t_{JRE} that minimises $J_{\mathrm{JRE}}(\{p_{ij|BD}\}; h_{ij}^t)$ is called JRE thresholding method, that is

$$t_{\mathrm{JRE}} = \arg\left\{\min_{t\in G} J_{\mathrm{JRE}}(\{p_{ij|BD}\}; h_{ij}^t)\right\} \tag{36}$$

One comment is noteworthy. It should be noted that the LRE and JRE originally defined by Lee *et al.* [27] are not conditional probability distributions as they are not normalised by probabilities of the two quadrants that constitute LRE and JRE. Because of this reason, technically, they cannot be called relative entropy, even these two seemed to work well as threshold criteria [27].

4 Histogram compression and translation

It was reported by Ramac and Varshney [26] that Chang *et al.*'s relative entropy method, GRE did not perform well for some images. This was mainly due to fact that their image histograms are distributed sparsely with large gaps between two consecutive grey levels. Unlike entropy-based methods, relative entropy-based methods are generally sensitive to such sparse image histograms. In this case, in order for relative entropy-based methods to work effectively, a sparse image histogram must be compressed to a more compact histogram. This idea is called histogram compression and translation (HCT), which is very similar to the commonly used histogram equalisation. However, instead of stretching a 1-D image histogram to cover the entire grey-level range as the histogram equalisation does, the HCT, does inversely by compressing the 2-D histogram due to relationship of one grey level relative to another. This is a major difference between the histogram equalisation and the HCT, because the former deals with a 1-D image histogram, whereas the latter has to take into account the relative spatial relationship characterised by a 2-D histogram resulting from a co-occurrence matrix, in which case the image histogram must be compressed rather than being stretched. In what follows, we develop a method for this purpose.

Suppose that the total number of distinct grey levels in an image is N. Without loss of generality, we assume that g_1, $g_2, \ldots, g_N$ are these N distinct grey levels that can be arranged in accordance with $g_1 < g_2 < \cdots < g_N$, where $g_1 = g_{\min}$ is the smallest grey level and $g_N = g_{\max}$ is the largest grey level. Let $n(g_k)$ be the total number of pixels in the image whose grey level is g_k. Two parameters will be used to measure the sparseness of a 1-D image histogram. One parameter is the N. Another is the width of a histogram defined by $w = g_N - g_1$. In general, $w \geq N$. If a 1-D image histogram whose width w is very close to N, then its histogram will be dense and distributed compactly. On the contrast, if w is much greater than N, the histogram will be distributed sparsely. In this case, a histogram compression and translation is generally needed for relative entropy-based thresholding methods. The process is referred to as HCT, defined by mapping $g_k \rightarrow k$ with

$$\mathrm{HCT}(g_k) = k \text{ and } n_k = n(g_k) \quad \text{for } 1 \leq k \leq N \tag{37}$$

Using (37), a new HCT-compressed and translated 1-D image histogram can be created for the original image, which is a plot of n_k against k with $1 \leq k \leq N$.

5 Performance measures

In order to avoid human interpretation, two objective measures, uniformity and shape [1, 2], will be used for performance evaluation.

5.1 Uniformity measure

The uniformity measure is generally used to describe region homogeneity in an image. For a given threshold t, it is defined by

$$U(t) = 1 - \frac{\sigma_B^2(t) + \sigma_F^2(t)}{C} \tag{38}$$

where B and F represent background and foreground regions, $f(x, y)$ is the grey level of the pixel (x, y)

$$C = \frac{1}{2}(g_{\max} - g_{\min})^2, \quad \mu_B^t = \frac{\sum_{(x,y)\in B} f(x,y)}{n_B^t},$$

$$\mu_F^t = \frac{\sum_{(x,y)\in F} f(x,y)}{n_F^t},$$

$$\sigma_B^2(t) = \frac{1}{n_B^t}\sum_{(x,y)\in B}(f(x,y) - \mu_B^t)^2,$$

$$\sigma_F^2(t) = \frac{1}{n_F^t}\sum_{(x,y)\in F}(f(x,y) - \mu_F)^2,$$

n_B^t is the number of pixels in background region and n_F^t is the number of pixels in foreground region.

5.2 Shape measure

The shape measure is generally used to measure geometric features of objects present in an image. It is calculated as follows.

(a) We first define a generalised gradient function $\Delta(x, y)$ by

$$\Delta(x,y) = \begin{bmatrix} \sum_{k=1}^{4} D_k^2 + \sqrt{2}D_1(D_3 + D_4) \\ -\sqrt{2}D_2(D_3 - D_4) \end{bmatrix}^{1/2} \tag{39}$$

where $D_1 = f(x+1, y) - f(x-1, y)$, $D_2 = f(x, y-1) - f(x, y+1)$, $D_3 = f(x+1, y+1) - f(x-1, y-1)$ and $D_4 = f(x+1, y-1) - f(x-1, y+1)$, and assign its value to every pixel (x, y). It should be noted that the gradient D_1 dictates the grey-level changes along x-axis (i.e. 0°–180° horizontal line), whereas the gradient D_2 dictates the grey-level changes along the y-axis (i.e. 90°–270° vertical line). Additionally, the gradient D_3 dictates the grey-level changes diagonally (i.e. 45°–225° diagonal line) compared with the gradient D_4 that dictates the grey-level changes anti-diagonally (i.e., 135°–315° second diagonal line). Basically, these four gradients cover all the eight orientations, 0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°, which can be used to capture image shape features.

(b) Second, if the pixel (x, y) has a grey value higher than the average of its neighbours, then assign '+' sign to $\Delta(x, y)$ and assign '−' sign, otherwise.

(c) Third, compute the shape measure using the following formula

$$S_F(t) = \frac{\sum_{(x,y)\in F} \text{sign}(f(x,y) - \bar{f}_B)\Delta(x,y)\text{sign}(f(x,y) - t)}{C_F} \tag{40}$$

where

$$\text{sign}(x) = \begin{cases} +1; & \text{if } x \geq 0 \\ -1; & \text{if } x < 0 \end{cases}$$

is the sign function, C_F is a normalisation constant given by

$$C_F = \max_t \left\{ \sum_{(x,y)\in F} \text{sign}(f(x,y) - \bar{f}_B)\Delta(x,y) \times \text{sign}(f(x,y) - t) \right\} \text{ and}$$

$$\bar{f}_B = \frac{1}{8}\left[\sum_{i=x-1}^{x+1}\sum_{j=y-1}^{y+1} f(i,j) - f(x,y) \right].$$

6 Experiments

In this section, seven entropy-based thresholding methods (LE, JE, GE, Kittler and Illingworth's MET, LRE, JRE, GRE) will be implemented and compared via a series of experiments. They can be categorised into three groups. The first group contains one first-order thresholding method, Kittler and Illingworth's MET, which relies on 1-D image histograms without taking into account inter-pixel spatial correlation. The second and third groups are made up of second-order thresholding methods, which utilise a co-occurrence matrix to account for spatial correlation among pixels. The second group comprises of three entropy thresholding methods, LE, JE and GE and the third group consists of three relative entropy thresholding methods, LRE, JRE and GRE, which are considered to be counterparts of the methods in the second group.

In order to make our comparative study more complete, we also include Otsu's [28] method as a benchmark comparison. Otsu's method is a widely used thresholding method and has been shown to perform well in general. It is based on a criterion that maximises ratio of between-class variance to within-class variance and can be described briefly as follows.

6.1 Otsu's method

Otsu's method is a special case of two-class Fisher's linear discriminant analysis (LDA) in pattern classification [31], where the optimal criterion is the ratio of between-class variance to within-class variance. Let the 1-D histogram of an image be described by a probability vector, $(p_0, p_1, \ldots, p_{L-1})$, where $p_i = n_i/n$, n_i is the number pixels with grey-level value i and n is the number of image pixels. Suppose that t is a selected threshold value. Then probabilities of background and foreground of the t-thresholded binary image can be defined by

$$P_B^t = \sum_{i=0}^{t} p_i \quad \text{and} \quad P_F^t = 1 - P_B^t = \sum_{i=t+1}^{L-1} p_i \tag{41}$$

Using (41), the means and variances associated with background and foreground can be further defined, respectively,

as follows

$$\mu_B^t = \frac{1}{P_B^t}\sum_{i=0}^{t} ip_i \quad \text{and} \quad \mu_F^t = \frac{1}{P_F^t}\sum_{i=t+1}^{L-1} ip_i \tag{42}$$

$$\text{var}_B^t = \frac{1}{P_B^t}\sum_{i=0}^{t} (i-\mu_B^t)^2 p_i \quad \text{and}$$

$$\text{var}_F^t = \frac{1}{P_F^t}\sum_{i=t+1}^{L-1} (i-\mu_F^t)^2 p_i \tag{43}$$

It further considers the between-class variance and within-class variance defined similarly in Fisher's LDA by

$$\begin{aligned}\text{var}_{\text{between-class}}^t &= P_B^t(\mu_B^t-\mu)^2 + P_F^t(\mu_F^t-\mu)^2 \\ &= P_B^t P_F^t(\mu_B^t-\mu_F^t)^2\end{aligned} \tag{44}$$

where $\mu = \sum_{i=0}^{L-1} ip_i$ is the global mean of the image and

$$\text{var}_{\text{within-class}}^t = P_B^t \text{var}_B^t + P_F^t \text{var}_F^t \tag{45}$$

It then finds a threshold value, t_{Otsu} that maximises $\text{var}_{\text{between-class}}^t$, or equivalently minimises $\text{var}_{\text{within-class}}^t$, that is

$$\begin{aligned}t_{\text{Otsu}} &= \arg\left\{\max_{1\le t\le L}\{\text{var}_{\text{between-class}}^t\}\right\} \\ &= \arg\left\{\min_{1\le t\le L}\{\text{var}_{\text{within-class}}^t\}\right\}\end{aligned} \tag{46}$$

Conceptually, Otsu's idea is very similar to Pal and Pal's JE method. If we interpret the local quadrants in Fig. 1 as within-class quadrants and the joint quadrants in Fig. 1 as between-class quadrants, Otsu's method is essentially similar to the JE and JRE methods, despite the fact that they are technically different methods. Otsu's method is a first-order method, which uses the 1-D image histogram to form within-class and between-class variances and maximises the between-class variance, whereas JE and JRE are second-order methods, which maximise the entropy and relative entropy of joint quadrants of a co-occurrence matrix, respectively. Additionally, the measure used in Otsu's method is variance compared with the self-information (i.e. $-\log p_i$) in the joint quadrant used in the JE and the discrepancy of self-information between two joint quadrants (i.e. $\log(p_{ij}|BD/h_{ij}^t) = -\log h_{ij}^t - \log p_{ij|BD}$) used in JRE methods.

More interestingly, $\sigma_B^2(t)$, $\sigma_F^2(t)$ in (38) can be re-expressed as

$$\begin{aligned}\sigma_B^2(t) &= \sum_{(x,y)\in B}(f(x,y)-\mu_B^t)^2 = \sum_{i=0}^{t}(i-\mu_B^t)^2 n_i \\ &= n\sum_{i=0}^{t}(i-\mu_B^t)^2\left(\frac{n_i}{n}\right) = n\sum_{i=0}^{t}(i-\mu_B^t)^2 p_i = n\,\text{var}_B^t\end{aligned} \tag{47}$$

$$\begin{aligned}\sigma_F^2(t) &= \sum_{(x,y)\in F}(f(x,y)-\mu_F)^2 = \sum_{i=t+1}^{L-1}(i-\mu_F^t)^2 n_i \\ &= n\sum_{i=t+1}^{L-1}(i-\mu_F^t)^2\left(\frac{n_i}{n}\right) \\ &= n\sum_{i=t+1}^{L-1}(i-\mu_F^t)^2 p_i = n\,\text{var}_F^t\end{aligned} \tag{48}$$

From (43), we obtain

$$\text{var}_{\text{within-class}}^t = P_B^t \text{var}_B^t + P_F^t \text{var}_F^t = \frac{1}{n}[\sigma_B^2(t)+\sigma_F^2(t)] \tag{49}$$

By virtue of (49), maximising $U(t)$ in (38) is equivalent to minimising $\text{var}_{\text{within-class}}^t$, which is also equivalent to maximising $\text{var}_{\text{between-class}}^t$ according to (46). As a result, the threshold value produced by Otsu's method, t_{Otsu} is identical to the t that maximises $U(t)$ in (38). It should be noted that the values of $U(t)$ vary with images. However, the normalisation constant C in $U(t)$ is independent of the threshold value t. In this case, C can be chosen to normalise the values of $U(t)$ to the range of [0, 1] such that the minimum and maximum of $U(t)$ for each image were always set to 0 and 1 respectively for comparison. Using this process, the uniformity values calculated from $U(t)$ in the following experiments are always in between 0 and 1.

6.2 Experiments

The following experiments are conducted to demonstrate the performance of nine thresholding methods: a classification-based thresholding method, Otsu's method, a first-order entropic thresholding method; Pun and Kapur *et al.*'s ME, a first-order relative entropic thresholding methods; Kittler and Illingworth's MET, three second-order entropy thresholding methods; Pal and Pal's JE and LE, GE; three second-order relative entropy thresholding methods: LRE, JRE and GRE with/without HCT where the uniformity and shape measures were also used for objective performance criteria. Additionally, the two parameters w and N were also studied to evaluate the need of HCT. Four different images were selected for experiments.

Experiment 1: Watch: The image studied in this experiment is a watch shown in Fig. 2*a*. Its 1-D histograms before and after HCT are nearly the same. They are plotted in Figs. 2*b* and *c* with w and N shown in Fig. 2*b*. The plots of the co-occurrence matrices of Figs. 2*b* and *c* are shown in Figs. 2*d* and *e*. The values of uniformity and shape were calculated and also plotted in Figs. 2*f* and *g*. Figs. 3*a*–*l* show the binary images resulting from ME, MET, Otsu, JE, LE, GE, LRE with HCT, JRE with HCT, GRE with HCT, LRE, JRE and GRE, respectively. As we can see from the thresholded images in Fig. 3, the best results were produced by the MET and Otsu's method in Figs. 3*b* and *c*, which outperformed all the second-order entropy and relative entropy thresholding methods. Table 1 tabulates the uniformity and shape values of their threshold values where the MET and the Otsu's method yielded largest values. As noticed, most thresholding methods generated higher uniformity values than shape values. This implies that the uniformity of the watch image had more influence than shape does on the thresholded images.

Experiment 2. House: The w and N of the watch image studied in Experiment 1 were approximately the same where HCT did not have impact on the thresholded results. This experiment shows another extreme as opposed to Experiment 1. The image is shown in Fig. 4*a* with its 1-D histograms before and after HCT plotted in Figs. 4*b* and *c* where w and N are also shown in Fig. 4*b*, $g_1 = g_{\min} = 78$ and $g_N = g_{\max} = 255$ with $N = 68$ and $w = 238$. In this case, the width, w, is much greater than N with $w/N = 3.5$. The original histogram in Fig. 4*b* looks very sparse with grey-level values spread from 78 to 255. In contrast, the HCT-compressed and translated histogram in Fig. 4*c* was compacted with grey-level values in a compressed and translated range from 1 to 69. The plots of

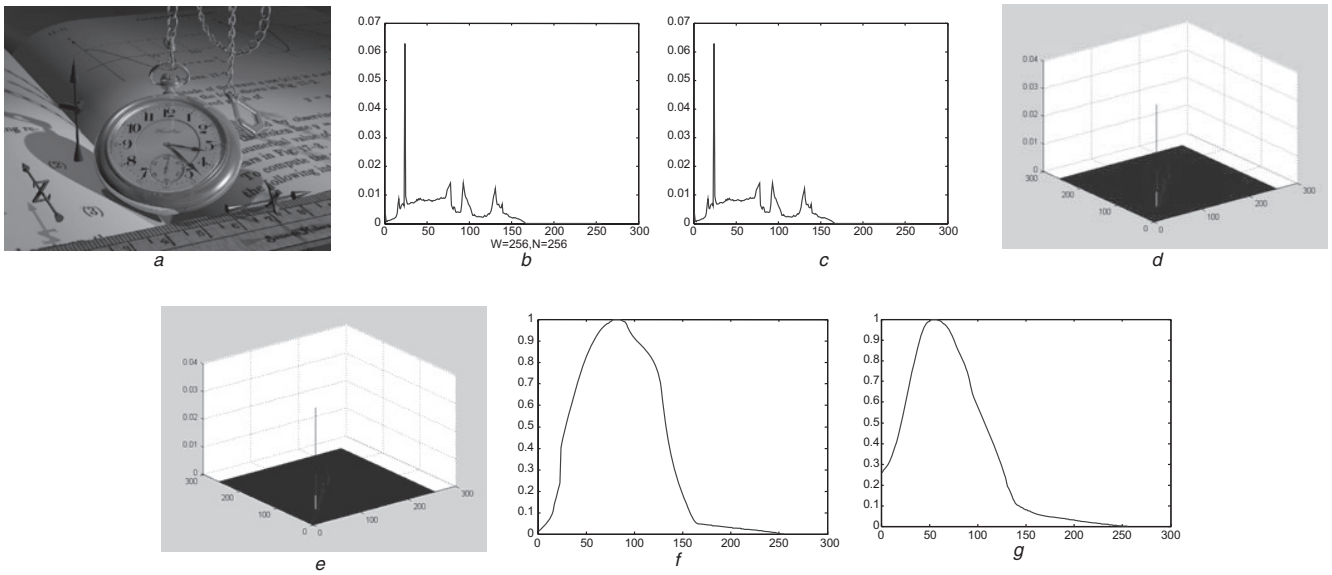

Fig. 2 *Watch image*

a Watch *b* 1-D histogram *c* 1-D histogram with HCT *d* 2-D histogram *e* 2-D histogram after HCT *f* Uniformity *g* Shape

the co-occurrence matrices of Figs. 4*b* and *c* are shown in Figs. 4*d* and *e*. As we can see, the inter-pixel spatial correlation between grey-level values in Fig. 4*e* was much denser than that in Fig. 4*d*. The values of uniformity and shape were also calculated and plotted in Figs. 4*f* and *g*. Figs. 5*a–l* show the thresholded binary images resulting from the methods of ME, MET, Otsu, JE, LE, GE, LRE with HCT, JRE with HCT, GRE, with HCT, LRE, JRE

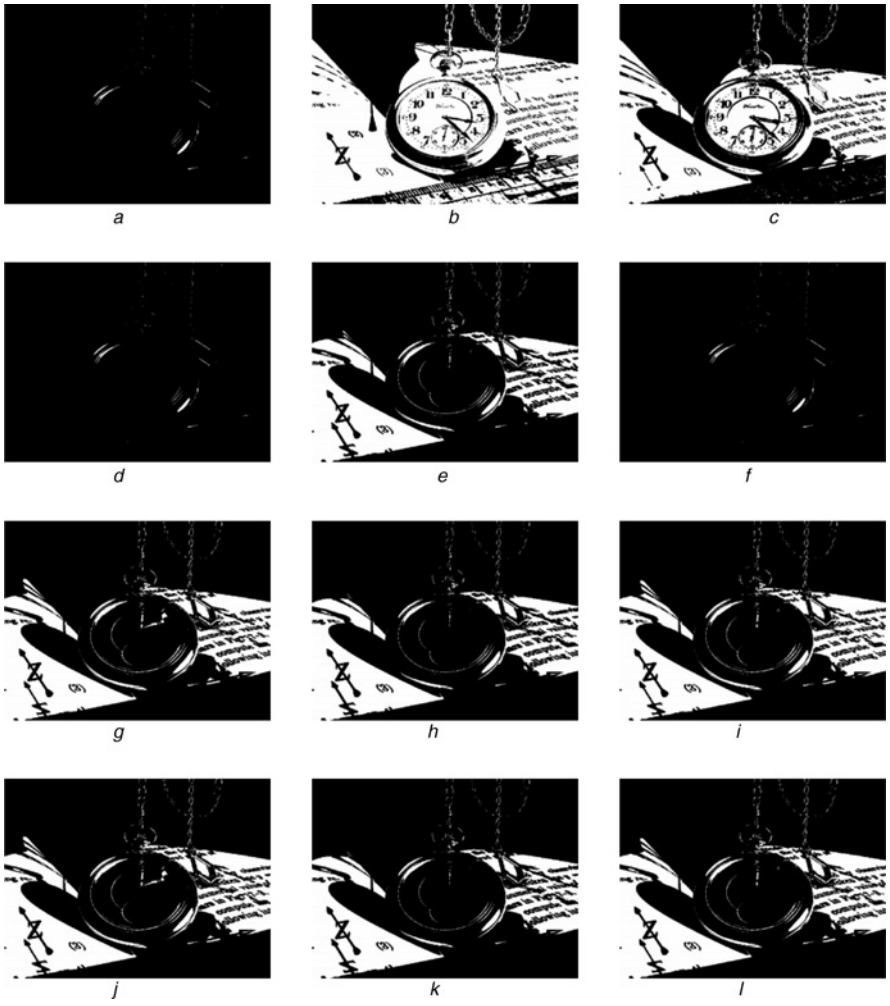

Fig. 3 *Binary thresholded images resulting from various methods*

a ME ($t = 166$) *b* MET ($t = 63$) *c* OTSU ($t = 81$) *d* LE ($t = 165$) *e* JE ($t = 107$) *f* GE ($t = 165$) *g* LRE with HCT ($t = 101$) *h* JRE with HCT ($t = 109$) *i* GRE with HCT ($t = 102$) *j* LRE ($t = 101$) *k* JRE ($t = 109$) *l* GRE ($t = 102$)

Table 1: Uniformity and shape values resulting from nine thresholding methods in this paper

	Uniformity				Shape			
	Watch	House	Tank	Text	Watch	House	Tank	Text
ME	0.0473	0.9149	0.9043	0.4564	0.0526	0.9449	0.8486	0.7307
MET	0.9439	0.9952	0.8967	0.9212	0.9818	0.6052	0.9957	0.5927
Otsu	1	1	1	1	0.8290	0.6687	0.9258	0.7472
LE	0.0480	0.9098	0.9043	0.5906	0.0535	0.9712	0.8486	0.9124
JE	0.8745	0.9874	0.6746	0.4229	0.4907	0.5849	0.8028	0.6771
GE	0.0480	0.9874	0.8329	0.4638	0.0535	0.5849	0.8182	0.7427
LRE with HCT	0.9007	0.9098	0.9977	0.9106	0.5539	0.9712	0.9331	0.5775
JRE with HCT	0.8653	0.9874	0.6746	0.4229	0.4680	0.5849	0.8028	0.6771
GRE with HCT	0.8959	0.9874	0.9991	0.9866	0.5427	0.5849	0.9108	0.6831
LRE	0.9007	0.9149	0.0656	0.9106	0.5539	0.9712	0.1052	0.5775
JRE	0.0185	0.0032	0.3414	0.2876	0.2666	0.5849	0.7277	0.3735
GRE	0.8959	0.9868	0.0208	0.9866	0.5427	0.5849	0.0289	0.6837

and GRE respectively. Table 1 also tabulates their respective uniformity and shape values. Apparently, the best thresholded images were those produced by the LE, the LRE with HCT and the LRE which yielded very high values of uniformity and shape measures, where the shape values were higher than uniformity values. In contrast to the watch image in Fig. 2*a* where the uniformity was more important than the shape, this observation suggested that the shape of the house image was more crucial than its uniformity. This was also verified by Otsu's method where it generated the highest uniformity value 1, but a low shape value of 0.6687.

Experiment 3. Tank: This experiment was conducted to show the need of HCT for relative entropy-based entropy thresholding to be effective. The image is a tank parked on the grass field shown in Fig. 6*a*. with $N = 138$ and $w = 212$. In this case, the width, w, is much greater than N with $w/N \simeq 1.5$. The original 1-D histogram in Fig. 6*b* was compressed and translated by HCT in Fig. 6*c*. The plots of the co-occurrence matrices of Figs. 6*b* and *c* are shown in Figs. 6*d* and *e*. As we can see, the inter-pixel spatial correlation among grey-level values in Fig. 4*e* is much more denser than that in Fig. 6*d*. The values of uniformity and shape were calculated and plotted in Figs. 6*f* and *g*. Fig. 7*a*–*l* shows the binary images resulting from the methods of ME, MET, Otsu, JE, LE, GE, LRE with HCT, JRE with HCT, GRE with HCT, LRE, JRE and GRE, respectively, with their respective uniformity and shape values tabulated in Table 1. Obviously, the relative entropy thresholding methods with HCT performed better than their counterparts without HCT as shown in Figs. 7*a*–*i* and Figs. 7*j*, and *k*. According to visual inspection, the best results came from the Otsu method, LRE with HCT, and GRE with HCT which also produced the highest values of uniformity and shape. Unlike Experiments 1 and 2, the uniformity and shape measures of the tank image were equally important. For example, the MET method produced the highest shape value, 0.9957, but the fifth highest uniformity value, 0.8967. The thresholded image shown in Fig. 7*b* was not good as

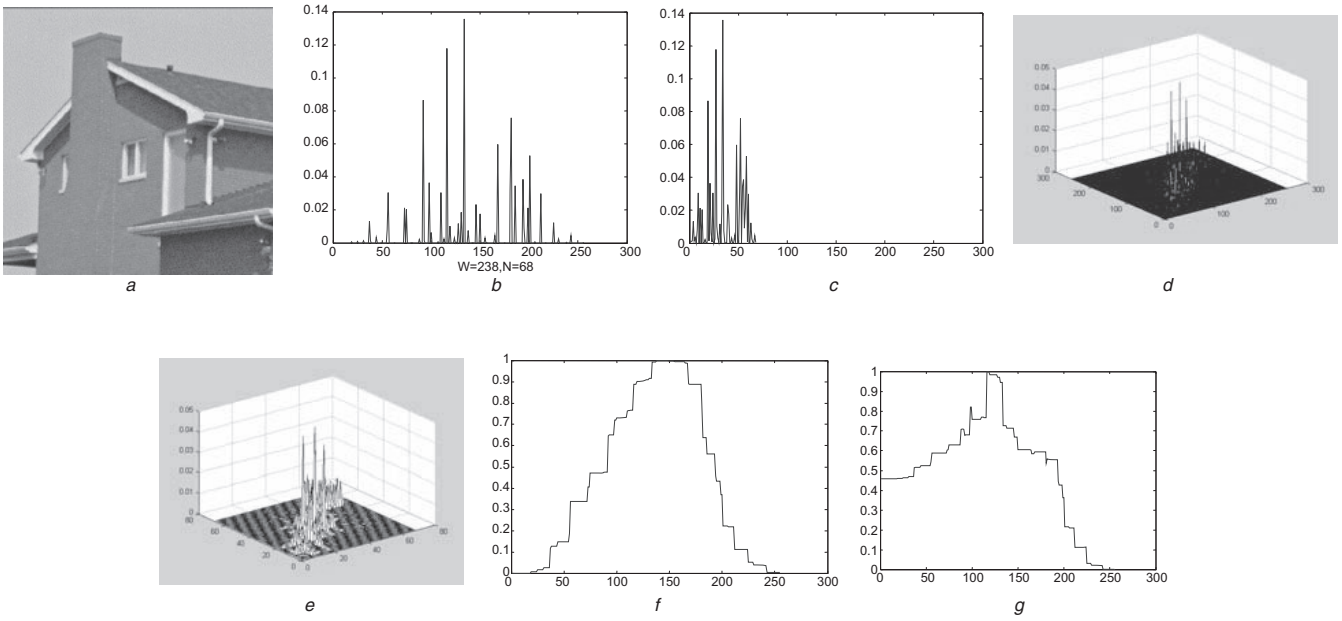

Fig. 4 *House image*

a House *b* 1-D histogram *c* 1-D histogram with HCT *d* 2-D histogram *e* 2-D histogram after HCT *f* Uniformity *g* Shape

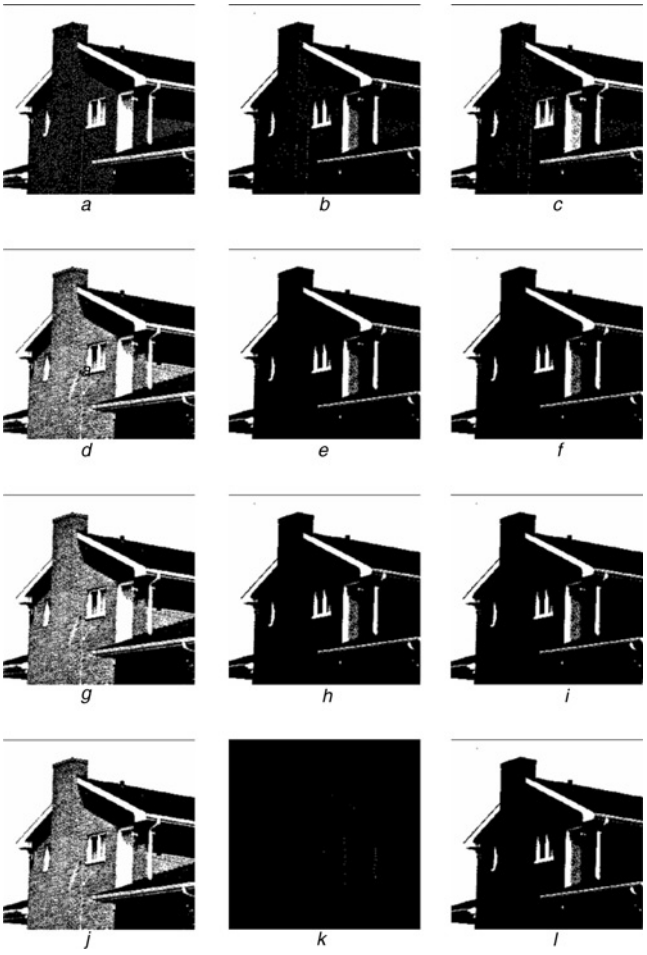

Fig. 5 *Binary thresholded images resulting from various methods*

a ME ($t = 134$) *b* MET ($t = 156$) *c* OTSU ($t = 145$)
d LE ($t = 130$) *e* JE ($t = 166$) *f* GE ($t = 166$)
g LRE with HCT ($t = 130$) *h* JRE with HCT ($t = 166$)
i GRE with HCT ($t = 166$) *j* LRE $t = 132$
k JRE $t = 254$ *l* GRE $t = 166$

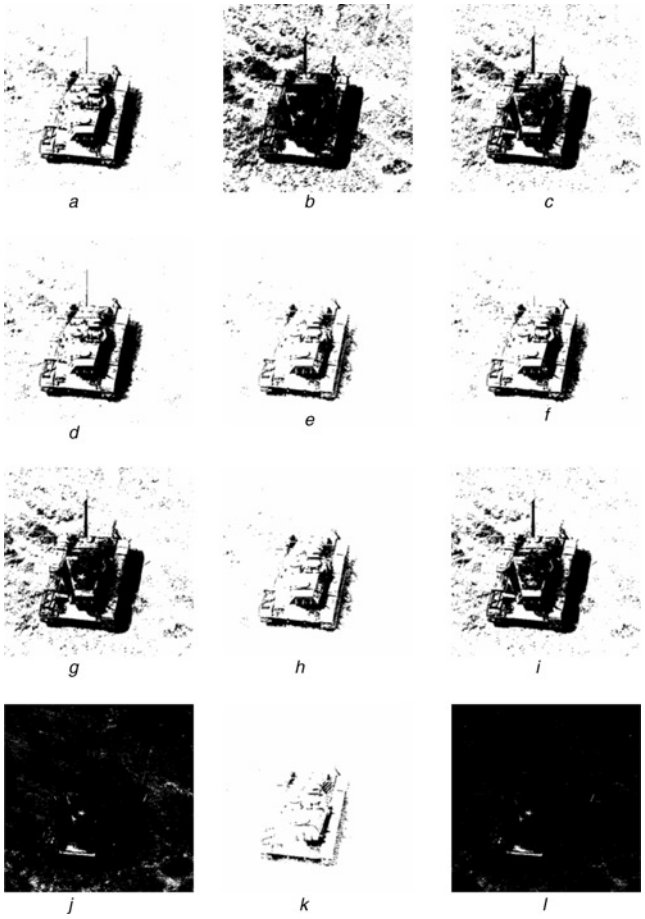

Fig. 7 *Binary thresholded images resulting from various methods*

a ME ($t = 97$) *b* MET ($t = 131$) *c* OTSU ($t = 116$)
d LE ($t = 96$) *e* JE ($t = 77$) *f* GE ($t = 88$)
g LRE with HCT ($t = 118$) *h* JRE with HCT ($t = 77$)
i GRE with HCT ($t = 113$) *j* LRE ($t = 113$)
k JRE ($t = 53$) *l* GRE ($t = 166$)

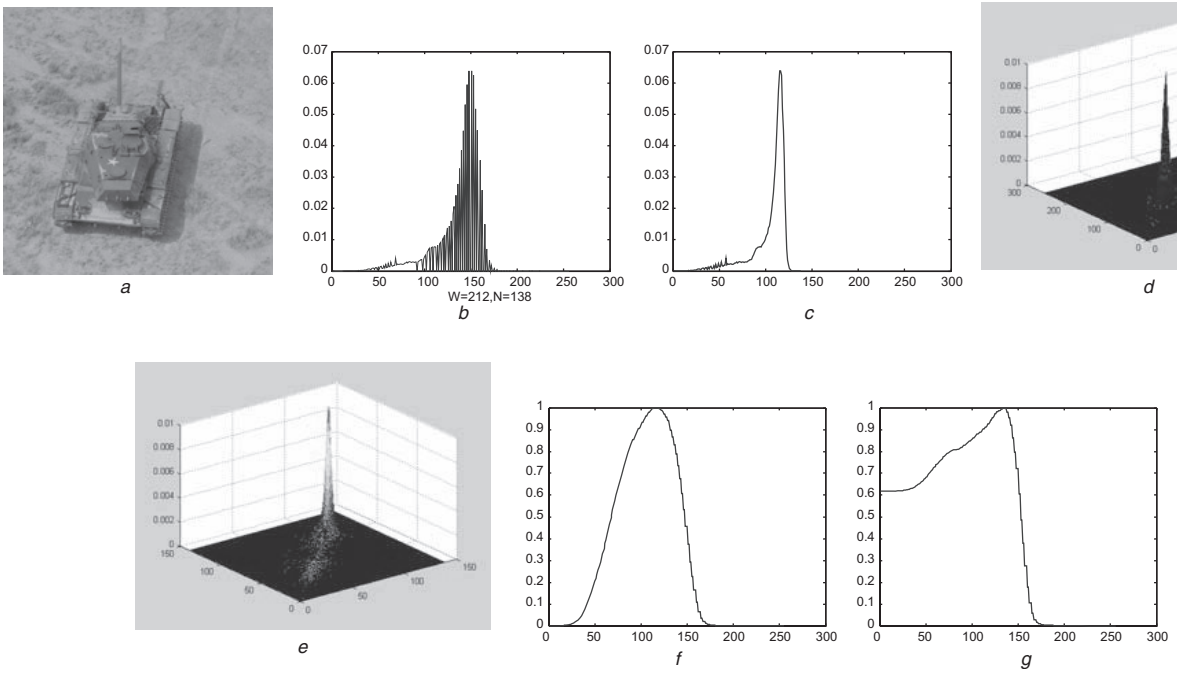

Fig. 6 *Tank image*

a Tank *b* 1-D histogram *c* 1-D histogram with HCT *d* 2-D histogram *e* 2-D histogram after HCT *f* Uniformity *g* Shape

those in Figs. 7*c*, *g* and *i*, all of which produced the uniformity values ≥ 0.99 and shape values ≥ 0.91.

Experiment 4. Text video image: In Experiments 1–3 we have shown that the uniformity and shape provided good objective measures of thresholded results as expected in the literature. On the basis of the results of the previous experiments, we may promptly jump into a conclusion that a good threshold value should result in high uniformity or shape values. Unfortunately, such a conclusion is misleading and is generally not true. The following experiment offers a counterexample. The image studied in this experiment was a video image shown in Fig. 8*a* with its 1-D histograms without/with HCT and their corresponding co-occurrence matrices plotted in Figs. 8*b* and *c* and Figs. 8*d* and *e*, respectively. Because the w and N shown underneath Fig. 8*b* are the same, there was no need to perform HCT. However, $w = N = 256$ suggested that the video image used up all grey-level values to describe the complicated image background where the main scene was simple text shown in the centre of the image. The values of uniformity and shape were calculated and plotted in Figs. 8*f* and *g*. Figs. 9*a*–*l* show the binary thresholded images resulting from the methods of ME, MET, Otsu, JE, LE, GE, LRE with HCT, JRE with HCT, GRE with HCT, LRE, JRE and GRE, respectively, where their respective uniformity and shape values are also tabulated in Table 1. From an application of information retrieval and index, the best thresholded image is the one produced by the JRE where the text in the video image was clearly extracted. However, if we compare the uniformity and shape values in Table 4, the JRE yielded the lowest values in both uniformity and shape. This is because the video image in Fig. 8*a* has very complicated image background where the effectiveness of shape and uniformity were substantially impaired by low resolution and distorted image background.

In addition to the previous experiments, an extensive set of experiments was also conducted for performance evaluation of the nine methods described in this paper. Unfortunately, including all of these experiments in this paper is not possible. Instead, we have chosen to include only four representatives of these experiments in this paper for illustration. Table 2 summarises these experiments where a 'yes' of HCT implies that the thresholded image can be improved by relative entropy methods; a 'yes' of uniformity means that uniformity plays a more crucial role in thresholding than does shape, and similarly for a 'yes' of shape. Nevertheless, several observations resulting from our experiments are noteworthy and can be briefly described as follows.

1. No single thresholding technique could claim the best method among all the experiments. However, second-order entropy thresholding methods generally performed better than first-order entropy thresholding methods. This is also true for relative entropy thresholding methods.
2. Interestingly, Otsu's method generally performed reasonably well in most of our experiments due to its classification-based thresholding criterion, which results in the highest uniformity value of 1. Nonetheless, in our conducted experiments, there always existed at least one or more from entropy and relative entropy methods that could perform comparably or better than Otsu's method. This suggested that entropy-based thresholding methods are generally a better approach than traditional thresholding methods.
3. A good threshold value generally produced high uniformity and shape values.
4. In most of our experiments, relative entropy thresholding methods with HCT performed better than their counterparts without HCT. However, on some occasions, relative entropy thresholding methods without HCT could perform better than their counterparts with HCT. More experiments for such comparison can be found in the work of Wang *et al.* [32].
5. Owing to the complicated image background shown in Experiment 4, first-order thresholding methods generally performed poorly compared with second-order thresholding methods, because it requires second-order statistical information to better capture background variations. More interestingly, Experiment 4 also demonstrated that for images with complicated background the commonly used objective measures, uniformity and shape might not be good criteria to be used for performance evaluation after all.

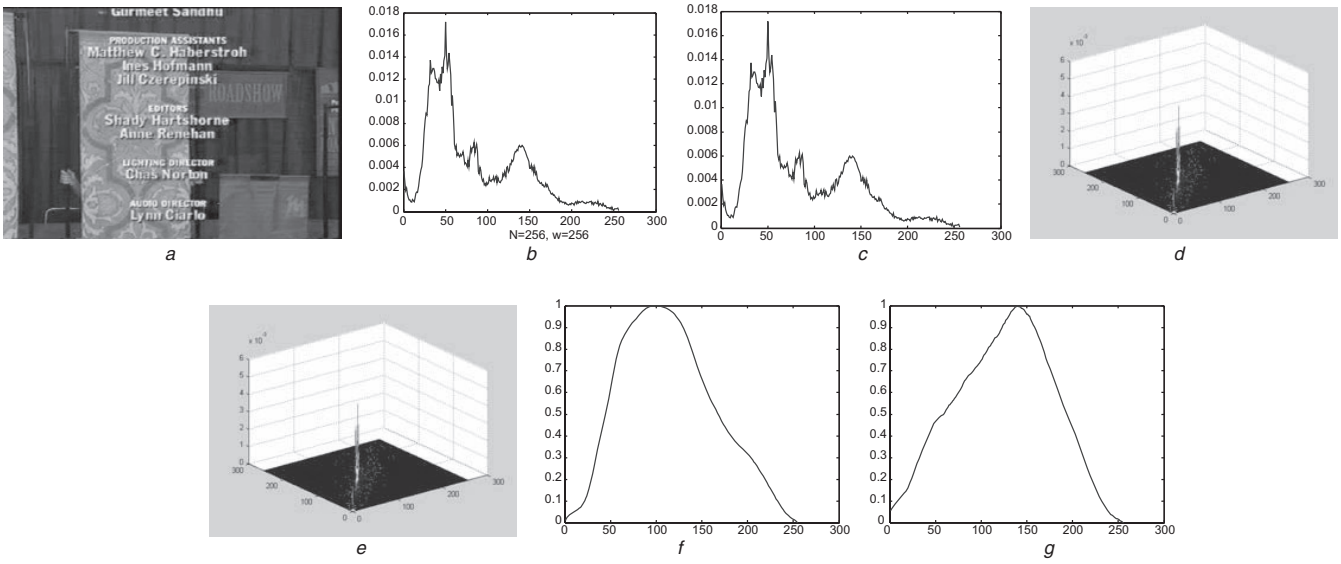

Fig. 8 *Video image with text*

a Video image with text *b* 1-D histogram *c* 1-D histogram after HCT *d* 2-D histogram *e* 2-D histogram after HCT *f* Uniformity *g* Shape

Fig. 9 *Binary thresholded images resulting from various methods (text)*
a ME ($t = 172$) *b* MET ($t = 73$) *c* OTSU ($t = 99$) *d* LE ($t = 156$) *e* JE ($t = 177$) *f* GE ($t = 171$) *g* LRE with HCT ($t = 71$) *h* JRE with HCT ($t = 205$) *i* GRE with HCT ($t = 87$) *j* LRE ($t = 71$) *k* JRE ($t = 205$) *l* GRE ($t = 87$)

Table 2: Summary of experiments resulting from nine thresholding methods in this paper

	(*w*, *N*)	HCT	Uniformity	Shape	Best thresholding methods
Watch	(256,256)	Yes	Yes	No	MET
House	(238,68)	Yes	No	Yes	LE, LRE w/o HCT
Tank	(212,138)	Yes	Yes	Yes	LRE with HCT, GRE with HCT, Otsu
Text	(256,256)	No	No	No	JRE

Table 3: One-to-one correspondence between entropic thresholding and relative entropic thresholding methods

Entropic thresholding methods	Relative entropic thresholding methods
Pun/Kapur *et al.*'s ME [8, 9]	Kittler and Illingworth's MET [20]
GE	GRE [24]
LE [19]	LRE
JE [19]	JRE

Table 4: Relationship among entropic thresholding and relative entropic thresholding

	Entropic thresholding	Relative entropic thresholding
Criterion (information theoretic measures)	Shannon's entropy	Kullback–Leibler information measure (also known as directed divergence, cross entropy, relative entropy)
First-order methods (histogram-based)	Pun/Kapur *et al.*'s ME	Kittler and Illingworth's MET
Second-order methods (co-occurrence matrix-based)	GE	GRE
	LE	LRE
	JE	JRE

7 Conclusion

In this paper, a comprehensive and comparative study of entropy thresholding and relative entropy thresholding techniques is presented. A total of eight different entropy-based information theoretic methods, ME, MET, LE, JE, GE, LRE, JRE, GRE, along with Otsu's method are considered and evaluated by two objective measures, uniformity and shape. There are several contributions made in this paper. One major contribution is to provide a detailed treatment on entropy thresholding and relative entropy thresholding with their counterparts tabulated in Table 3 and corresponding relationship summarised in Table 4. Another contribution is three new thresholding methods, an entropy thresholding method, GE; and two relative entropy thresholding methods, LRE and JRE. A third contribution is an introduction of the HCT to improve relative entropy thresholding methods. A fourth contribution is to show that uniformity and shape are generally good thresholding measures for grey-scale images, but not necessarily true for video images.

8 Acknowledgment

The first two authors would like to thank the US Department of Defense for supporting their work through contract number MDA-904-00-C2120.

9 References

1 Sahoo, P.K., Soltani, S., Wong, A.K.C., and Chen, Y.C.: 'A survey of thresholding techniques', *Comput. Vis. Graph. Image Process.*, 1988, **41**, pp. 233–260
2 Lee, S.U., Chung, S.Y., and Park, R.H.: 'A comparative performance study of several global thresholding techniques for segmentation', *Comput. Vis. Graph. Image Process.*, 1990, **52**, pp. 171–190
3 Haralick, R.M., and Shapiro, L.G.: 'Computer and robot vision, vol. 1' (Addison-Wesley, 1992)
4 Pal, N.R., and Pal, S.K.: 'A new definition and its application', *IEEE Trans. Syst. Man Cybern.*, 1991, **21**, (5), pp. 1260–1270
5 Sahoo, P.K., Slaaf, D.W., and Albert, T.A.: 'Threshold selection using a minimal histogram entropy difference', *Opt. Eng.*, 1997, **36**, pp. 1976–1981
6 Cover, T., and Thomas, J.: 'Elements of information theory' (John Wiley & Sons, Inc., 1991)
7 Pun, T.: 'A new method for grey-level picture thresholding using the entropy of the histogram', *Signal Process.*, 1980, **2**, pp. 223–237
8 Pun, T.: 'Entropic thresholding: a new approach', *Comput. Graph. Image Process.*, 1981, **16**, pp. 210–239
9 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. Vis. Graph. Image Process.*, 1985, **29**, pp. 273–285
10 Sahoo, P., Wilkins, C., and Yeager, J.: 'Threshold selection using Renyi's entropy', *Pattern Recognit.*, 1997, **30**, (1), pp. 71–84
11 Abutaleb, A.S.: 'Automatic thresholding of grey-level pictures using two-dimensional entropy', *Comput. Vis. Graph. Image Process.*, 1989, **47**, pp. 22–32
12 Brink, A.D.: 'Thresholding of digital images using two-dimensional entropies', *Pattern Recognit.*, 1992, **25**, (8), pp. 803–808
13 Chen, W.-T., Wen, C.-H., and Yang, C.-W.: 'A fast two-dimensional entropic thresholding algorithm', *Pattern Recognit.*, 1994, **27**, (7), pp. 885–893
14 Yang, C.-W., Chung, P.-C., and Chang, C.-I: 'A hierarchical fast two-dimensional entropic thresholding algorithm using a histogram pyramid', *Opt. Eng.*, 1996, **35**, (11), pp. 3227–3241
15 Gong, J., Li, L., and Chen, W.: 'Fast recursive algorithms for two-dimensional thresholding', *Pattern Recognit.*, 1998, **31**, (3), pp. 295–300
16 Li, L., Gong, J., and Chen, W.: 'Gray-level image thresholding based on Fisher linear projection of two-dimensional histogram', *Pattern Recognit.*, 1997, **30**, (5), pp. 743–749
17 Gonzalez, R., and Woods, J.: 'Digital image processing' (Addison-Wesley, 1992)
18 Haralick, R.M., Shanmugam, K., and Dinstein, I.: 'Textural features for image segmentation', *IEEE Trans. Syst. Man Cybern.*, 1973, **SMC-3**, (6), pp. 610–621
19 Pal, N.R., and Pal, S.K.: 'Entropic thresholding', *Signal Process.*, 1989, **16**, pp. 97–108
20 Kittler, J., and Illingworth, J.: 'Minimum error thresholding', *Pattern Recognit.*, 1986, **19**, (1), pp. 41–47
21 Li, C.H., and Lee, C.K.: 'Minimum cross entropy thresholding', *Pattern Recognit.*, 1993, **26**, (4), pp. 617–625
22 Pal, N.R., and Pal, S.K.: 'Image model, Poisson distribution and object extraction', *Int. J. Pattern Recognit. Artif. Intell.*, 1991, **5**, (3), pp. 459–483
23 Pal, N.R.: 'On minimum cross-entropy thresholding', *Pattern Recognit.*, 1996, **29**, (4), pp. 575–580
24 Chang, C.-I, Chen, K., Wang, J., and Althouse, M.L.G.: 'A relative entropy-based approach to image thresholding', *Pattern Recognit.*, 1994, **27**, (9), pp. 1275–1289
25 Lee, S.S., Horng, S.-J., and Tsai, H.-R.: 'Entropy thresholding and its parallel algorithm on the reconfigurable array of processors with wide bus networks', *IEEE Trans. Image Process.*, 1999, **8**, (9), pp. 1229–1242
26 Ramac, L.C., and Varshney, P.K.: 'Image thresholding based on Ali-Silvey distance measures', *Pattern Recognit.*, 1997, **30**, (7), pp. 1161–1174
27 Lee, S.-K., Lo, C.-S., Wang, C.-M., Chung, P.-C., Chang, C.-I, Yang, C.-W., and Hsu, P.-C.: 'A computer-aided design mammography screening system for detection and classification of microcalcifications', *J. Med. Inform.*, 2000, **60**, (1), pp. 29–57
28 Otsu, N.: 'A threshold selection method from grey-level histograms', *IEEE Trans. Syst. Man Cybern.*, 1979, **SMC-9**, (1), pp. 62–66
29 Kullback, S.: 'Information theory and statistics' (Dover, 1968)
30 Chang, C.-I: 'Hyperspectral imaging: techniques for spectral detection and classification' (Kluwer Academic Publishers, New York, 2003), Chap. 2
31 Duda, P.E., and Hart, R.E.: 'Pattern classification and scene analysis' (John Wiley & Sons, 1973)
32 Wang, J., Du, Y., Chang, C.-I, and Thouin, P.: 'Relative entropy-based methods for image thresholding'. Int. Symp. Circuit and Systems (ISCAS) 2002, Scottsdale, AZ, May 2002