Chapter 3
Threshold Selection based on Entropy and Fuzzy Set Models

In this chapter we will take a look at some relevant concepts which are going to be used in this thesis. Also some threshold selection methods based on entropy will be briefly reviewed.

3.1 Entropy

3.1.1 Definition of Entropy

Entropy is a concept first used in the Second Law of Thermodynamics. It measures the spontaneous dispersal of energy - how much energy is spread out in a process, or how widely spread out it becomes – as a function of temperature. It was introduced into communications theory by Shannon [27] following the rapid development of communications. It is used to measure the efficiency of the information transferred through a noisy communication channel.

The mathematical definition of the entropy by Shannon [27] is:

$$H = -\sum_{i=0}^{n} p_i \cdot \lg(p_i), \tag{3.1}$$

in which H is the entropy, p_i is the statistical probability density of the event i. It defines the average amount of information (in r-ary units per symbol) obtained by observing a single source output. As its magnitude increases, more uncertainty and thus more information is associated with the source. If the source symbols are equally probable, the entropy or uncertainty of Equation 3.1 is maximized and the source provides the greatest possible average information per source symbol [6].

3.2 Maximization Entropy Theorem

From the above definition of entropy, the Maximization Entropy Theorem is derived. It is stated below:

For an event with n classes: D_1,D_2,...,D_n, with probability $p_1, p_2, ...,$ and p_n, respectively, the entropy H of the event is defined by Shannon [27] as Equation 3.1.

Proposition: 3.1 *The entropy function Equation 3.1 has a maximum value at, and only at, the point where the probability of each class is equal to* $\frac{1}{n}$.

Proof: It is known that:

$$\sum_{i=1}^{n} p_i = 1. \tag{3.2}$$

To obtain the maximum for Equation 3.1, Equation 3.2 is a constraint condition. Equation 3.2 can be rewritten as:

$$g(p_1, p_2, ..., p_n) = \sum_{i=1}^{n} p_i - 1 = 0. \tag{3.3}$$

Based on the Lagrange multiplier method, the stationary points of the entropy, subject to the constraint condition (Equation 3.2), can be found by constructing another function:

$$w(p_1, p_2, ..., p_n) = H(p_1, p_2, ..., p_n) + \lambda \cdot g(p_1, p_2, ..., p_n),$$

where λ is the Lagrange multiplier.

Let

$$\frac{\partial w}{\partial p_i} = 0,$$

for $i = 1, 2, ..., n$. Therefore

$$\frac{\partial H}{\partial p_i} + \lambda = 0 \Rightarrow \lambda = -\frac{\partial H}{\partial p_i} = 1 + \lg(p_i),$$

for $i = 1, 2, ..., n$. Such that

$$1 + \lg(p_1) = 1 + \lg(p_2) = ... = 1 + \lg(p_n), \tag{3.4}$$

which leads to

$$p_1 = p_2 = ... = p_n. \tag{3.5}$$

Substitution into Equation 3.2, shows that

$$p_1 = p_2 = ... = p_n = \frac{1}{n}. \tag{3.6}$$

We know that the entropy function Equation 3.1 has a stationary point at $p_1 = p_2 = ... = p_n = \frac{1}{n}$. Because

$$\frac{\partial^2 H}{\partial p_i^2} = -\frac{1}{p_i} < 0, \quad (for \quad p_i > 0\ )$$

for $i = 1, 2, ..., n$, the stationary point is a maximum point. Hence, the proposition is proved.

For $n = 2$, the relationship between entropy and probability is shown in Figure 3.1. We can see that at the point where $p_1 = p_2 = 0.5$, the entropy $H(p)$ reaches its peak value.

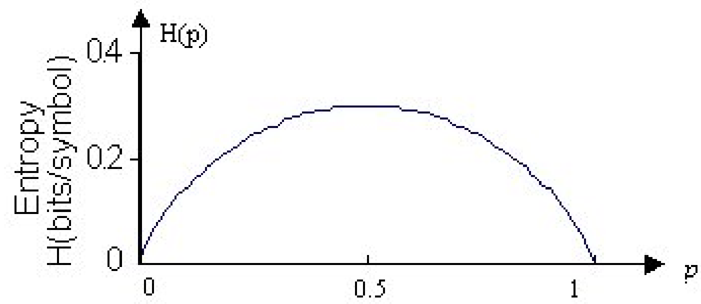

Figure 3.1: Entropy function vs probability

3.3 Fuzzy Sets and Fuzzy Probability

3.3.1 Fuzzy Set

Fuzzy sets were introduced by Zadeh in 1965 [70]. A classical set A is normally defined as a collection of elements or objects $x \in X$ that can be finite, countable, or over-countable. Each single element can either belong to or not belong to a set A, where $A \subset X$, which is denoted by the indicator function $\chi_A(x)$:

$$\chi_A(x) = \begin{cases} 1 & x \in A \\ 0 & otherwise \end{cases}$$

A fuzzy set is an extension of a classical set, in which an element may partially belong to a set. A fuzzy set A is defined as:

$$A = \{(x, \mu_A(x)|x \in X\}$$

where $0 \le \mu_A(x) \le 1$ is called the membership function. The value of $\mu_A(x)$ is the grade of x belonging to A.

Let $X = \{x_1, x_2, ..., x_n\}$ denote a universal, V_{cn} be a set of $c \times n$ matrices, where c is an integer, and $2 \le c \le n$. The elements of V_{cn} at the ith row and the kth column is denoted by μ_{ik} and $0 \le \mu_{ik} \le 1$, for $i = 1, 2, ..., c$, $k = 1, 2, ..., n$. The value of μ_{ik} represents the degree of the x_k belonging to the ith class. Thus the ith row of V_{cn}, and $\mu_i = (\mu_{i1}, \mu_{i2}, ..., \mu_{in})$, is a membership function of the fuzzy set A_i of X. If $\sum_{i=1}^{c} \mu_{ik} = 1$, for $k = 1, 2, ..., n$, then $\{\mu_1, \mu_2, ..., \mu_c\}$ defines a fuzzy c-partition of X which is denoted by the matrix M_{fc}, i.e.,

$$M_{fc} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_c \end{bmatrix}.$$

Note that $M_{fc} \in V_{cn}$.

3.3.2 Fuzzy Probability

Fuzzy probability is defined as:

$$P_i = \mu_i \cdot p_i, \tag{3.7}$$

where P_i represents the fuzzy probability of element i, μ_i represents the fuzzy membership of element i, while p_i is the statistical probability.

3.4 Entropy of a Fuzzy Set

Zadeh suggests a definition about the entropy of a fuzzy set which takes both distribution and the membership into consideration. It is defined as follows:

$$H(A) = -\sum_{i=1}^{n} \mu_A(x_i)p(x_i) \cdot \log(\mu_A(x_i)p(x_i)), \tag{3.8}$$

where A is a fuzzy set; $\mu_A(x_i)$ is the membership function of element x_i which has a probability distribution of $p(x_i)$. Kaufmann [68] defines the entropy of a fuzzy set as

$$H(A) = -\frac{1}{\ln(n)} \sum_{i=1}^{n} \varphi_A(x_i) \ln(\varphi_A(x_i)), \tag{3.9}$$

where

$$\varphi_A(x_i) = \frac{\mu_A(x_i)}{\sum_{i=1}^{n} \mu_A(x_i)}. \tag{3.10}$$

De Luca and Termini[67] proposed a quite different definition about the entropy of the fuzzy set A. This entropy has a nonprobabilistic feature and is defined as

$$H(A) = -\frac{1}{n \ln 2} \sum_{i=1}^{n} S_n(\mu_A(x_i)) \tag{3.11}$$

where $S_A()$ is the Shannon's function as follows:

$$S_n(\mu_A(x_i)) = -\mu_A(x_i) \ln(\mu_A(x_i)) - (1 - \mu_A(x_i)) \ln(1 - \mu_A(x_i)).$$

H. D. Cheng et al.[56] proposed another definition of entropy:

$$H(A) = -\frac{1}{\log n} \sum_{i=1}^{n} P_p(A_i) \log P_p(A_i), \tag{3.12}$$

where $P_p(A_i)$ is the probability summed in the space domain for the x (space domain) mapping into A_i (fuzzy domain) by the membership function $\mu_A(x)$.

$$P_p(A_i) = \sum_{\mu_A(x) \in A_i}^{n} p(x).$$

3.5 Thresholding Methods based on Entropy Criterion

3.5.1 Pun's Method

The entropic thresholding method is developed by Pun [30]. The entropy of the histogram after thresholding is selected as the criterion. According to the definition of entropy (in Shannon's sense [27]),

$$H = -\sum_{i=0}^{n} p_i \cdot \lg(p_i). \tag{3.13}$$

After two levels of thresholding, the image has only two classes: black (D^b) and white (D^w), where $D^b \bigcup D^w = D$ and $D^b \bigcap D^w = \phi$. The entropy of the image is:

$$H = -p_w \cdot \lg(p_w) - p_b \cdot \lg(p_b), \tag{3.14}$$

where $p_w = P(D^w)$ and $p_b = P(D^b)$. To satisfy the condition that the entropy is the maximum, the threshold must be selected such that $p_b = p_w = \frac{1}{2}$. The threshold is not adaptive. Also we can see that no matter how many pixels the object has in the original image, after thresholding there should always be half the total number of pixels remaining. The object pixels in one image may not be exactly half the total number of pixels of the image. This means we have to delete some object pixels, if they are more than half, or, if they are less, set some background pixels to be the object to make them equal to half the total number of pixels. This will lead to a bad segmentation of an image.

Pun [31] introduces a new approach derived from the previous entropic thresholding. The anisotropy coefficient α is introduced and is defined as:

$$\alpha = \left(\sum_{i=0}^{t} p_i \cdot \lg(p_i) \right) / \left(\sum_{i=0}^{n} p_i \cdot \lg(p_i) \right), \tag{3.15}$$

where t is the threshold and α is used to describe the geometric shape of the histogram. The threshold t is selected such that

$$\sum_{i=0}^{t} p_i = \frac{1}{2} + \left| \frac{1}{2} - \alpha \right| = \left\{ \begin{array}{ccc} 1-\alpha & if & \alpha \leq \frac{1}{2}, \\ \alpha & if & \alpha > \frac{1}{2}. \end{array} \right. . \tag{3.16}$$

For an extension to k level thresholding, $k-1$ thresholds $t_1, t_2, ..., t_{k-1}$ need to be selected. We can find $k-1$ levels $s_1, s_2, ..., s_{k-1}$ to divide the histogram into k equal parts and $k-1$ anisotropy coefficients α_1,α_2, ... , α_{k-1} are defined as:

$$\alpha_1 = \left(\sum_{i=0}^{s_1} p_i \cdot \lg(p_i) \right) / \left(\sum_{i=0}^{n} p_i \cdot \lg(p_i) \right);$$

$$\alpha_{k-1} = \left(\sum_{i=s_{k-2}}^{s_{k-1}} p_i \cdot \lg(p_i) \right) / \left(\sum_{i=0}^{n} p_i \cdot \lg(p_i) \right).$$

And the $k-1$ thresholds $t_1, t_2, ..., t_{k-1}$ are selected such that

$$\sum_{i=0}^{t_1} p_i = \sum_{i=0}^{s_1} p_i + \left| \sum_{i=0}^{s_1} p_i - \alpha_1 \right|,$$

$$\sum_{i=0}^{t_{k-1}} p_i = \sum_{i=0}^{s_{k-1}} p_i + \left| \sum_{i=s_{k-2}+1}^{s_{k-1}} p_i - \alpha_{k-1} \right|.$$

This new approach can adjust the threshold according to the shape of the histogram. Compared to the original method which restricts the thresholds selected to segment the histogram into equal parts, this method is adaptive and achieves good results for some images. But because it is derived from the method that segments the image into equal parts, the probability distribution in the thresholded images is nearly equal. But in fact we know that each class in the image may occupy a varying section of the partition. Also the introduction of α has introduced its own problem. For some images with unusual histograms, this method fails to produce acceptable results.

Figure 3.2: Gray scale image of Lena

Both Pun's algorithms are applied to a gray scale image as shown above (Figure 3.2). The original gray scale image is segmented into a two level image. The experiment results are shown below (See Figure 3.3(a), (b)). Also the thresholds generated by both algorithms are plotted in the histogram of the original gray scale image to show their relationship with the histogram (See Figure 3.3(c), (d)).

We can see from the experiment results that there is not much difference between the two segmented images because the two thresholds generated by them are 128 and 129, respectively, only 1 gray level difference.

3.5.2 Kapur's Method

Kapur et al.[32] also uses Shannon's concept of entropy but from a different point of view. Two probability distributions of the entire image instead of one probability distribution are considered in their method. One probability distribution is for the object and the other is for the background. The sum of the individual entropy of the object and the background is then maximized. In other words, this will result in equal-probability gray levels in each region, thus maximizing the sum of homogeneities in gray levels within object and background by making the gray levels equally probable in either region.

The probability distribution of the gray levels over the black part of the image is

$$\frac{p_0}{P_B}, \frac{p_1}{P_B}, ..., \frac{p_s}{P_B}, \tag{3.17}$$

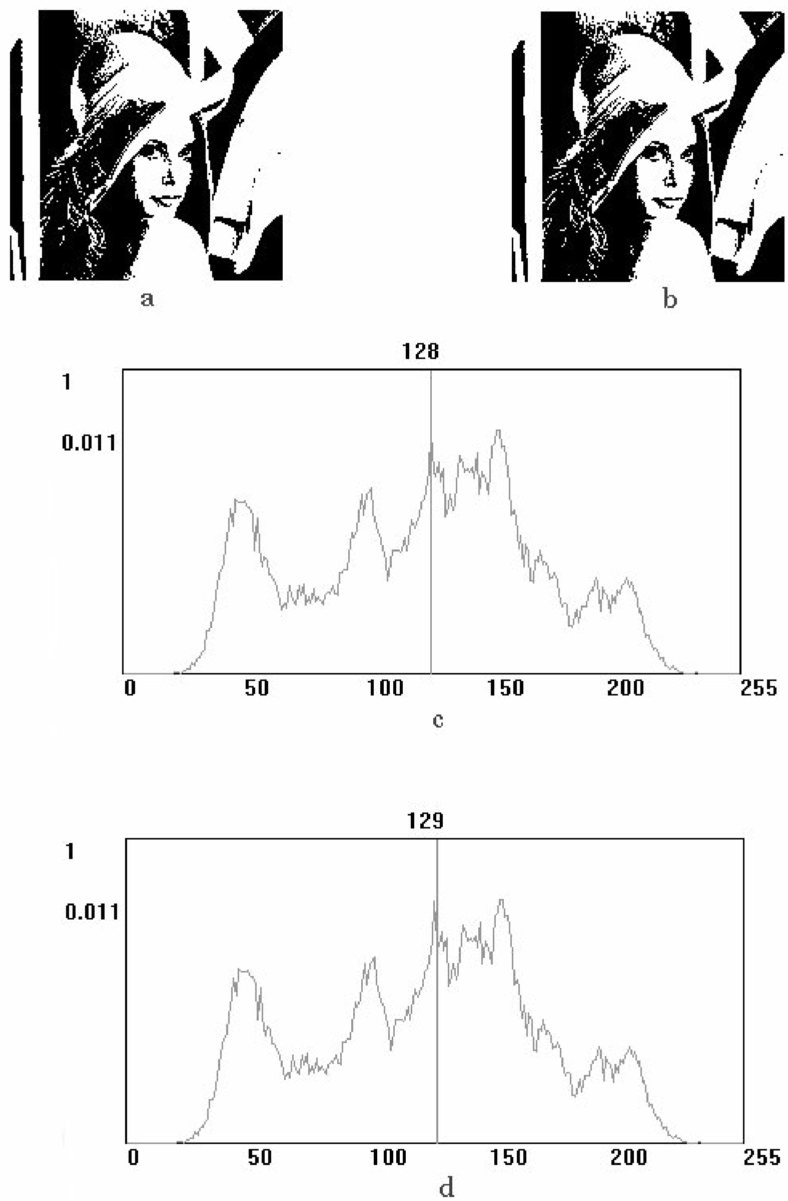

Figure 3.3: Experiment results on the gray scale image of Lena with both Pun's algorithms. (a) Two-Level segmented image with the first Pun's method; (b) Two-Level segmented image with the second Pun's method; (c) Threshold generated by the first method, Threshold=128; (c) Threshold generated by the second method, Threshold=129.

and that of the white part is

$$\frac{p_{s+1}}{1-P_B}, \frac{p_{s+2}}{1-P_B}, ..., \frac{p_{n-1}}{1-P_B}.$$

In which, s is the threshold; $p_i(i = 0, 1, ..., n-1)$ is the statistical probability of pixels with gray level i in the whole image; P_B is the probability of pixels with gray level less than or equal to threshold s.

$$P_B = \sum_{i=0}^{s} p_i. \tag{3.18}$$

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

$$H_B^{(s)} = -\sum_{i=0}^{s} \frac{p_i}{P_B} \log_2(\frac{p_i}{P_B}) \tag{3.19}$$

and that of the white part is

$$H_W^{(s)} = -\sum_{i=s+1}^{n-1} \frac{p_i}{1-P_B} \log_2(\frac{p_i}{1-P_B}). \tag{3.20}$$

The total entropy of the image is then defined as

$$H_T^{(s)} = H_B^{(s)} + H_W^{(s)}. \tag{3.21}$$

The threshold s is selected as the one which maximizes $H_T^{(s)}$.

The experiment result of Kapur's method is shown in Figure 3.4. We can see that the threshold generated by Kapur's method is a couple of levels lower than the ones generated by Pun's methods in this case.

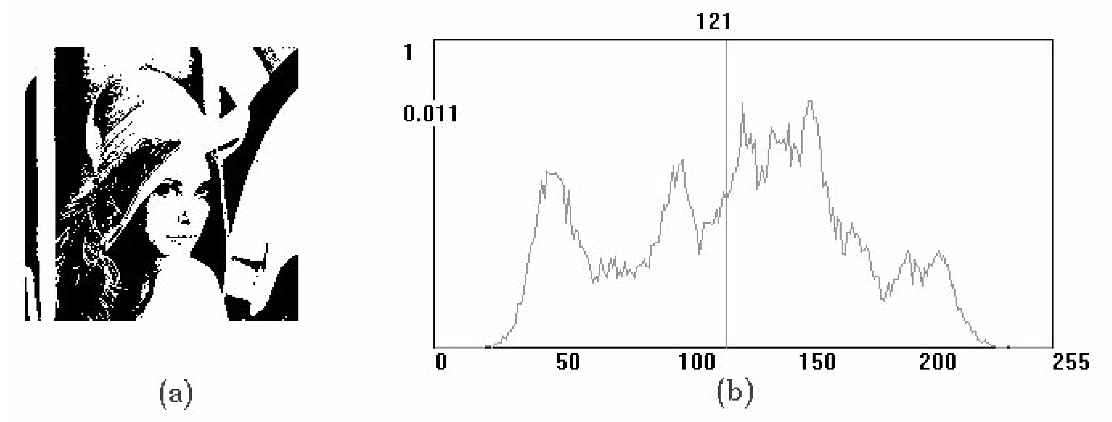

Figure 3.4: Experiment result with Kapur's method. (a) Binarized image of Lena; (b) Threshold generated by Kapur's method is 121.

3.5.3 Entropic Thresholding by N. R. Pal, S. K. Pal

In 1988, Nikhil R. Pal and Sankar K. Pal [35] proposed an entropic thresholding method based on Pun's [30], [31] and Kapur et al.'s [32] method. In their paper, it is proposed that the previous methods were developed without highlighting the adequacy of the concept of Shannon's entropy in the case of an image. This includes the dependency of pixel intensities in an image and the fact that its spatial distribution is not taken into account in the definition of its entropy. Hence, different images with identical histograms will always result in the same entropic value and the same threshold. This is considered by the authors to be not intuitively acceptable. In accordance with these facts, the authors present two other definitions of entropy, namely entropy of order q, $q = 1, 2, ...$, and the conditional entropy of an image. Entropy of order unity ($q = 1$) corresponds to the global entropy as used by Pun [30][31] and Kapur et al. [32]. Higher order ($q > 1$) entropies take into account the information contained in a sequence of gray levels of length q and thus denote the various local entropies of an image. The conditional entropy, on the other hand, gives an average amount of information that may be obtained from a region, after viewing the remaining portion of the image.

Entropy of order q is defined as

$$H^{(q)} = -\frac{1}{q} \sum_i p(s_i) \log_2 p(s_i).$$

For $q = 1$, i.e. sequence of length one, that is

$$H^{(1)} = -\sum_i p_i \log_2 p_i,$$

where p_i is the probability of occurrence of gray level i. We can see that this definition is identical to the one used in Shannon's definition.

For $q = 2$, i.e. sequence of length two, it is

$$H^{(2)} = -\frac{1}{2} \sum_i p(s_i) \log_2 p(s_i) = -\frac{1}{2} \sum_i \sum_j p_{ij} \log_2 p_{ij}, \tag{3.22}$$

where p_{ij} is the probability of the co-occurrence of the gray levels i and j. Expressions for higher order entropies $(q > 2)$ can be deduced in a similar manner. $H^{(q)}$, $q \geq 2$, may be called the "local entropy" of order q of an image.

Conditional entropy is defined as

$$H^{(c)} = (H(X|Y) + H(Y|X))/2,$$

in which

$$H(X|Y) = -\sum_{x_{ij} \in X} \sum_{y_{ji} \in Y} p(x_i|y_j) \log_2 p(x_i|y_j),$$

$$H(Y|X) = -\sum_{y_j \in Y} \sum_{x_i \in X} p(y_j|x_i) \log_2 p(y_j|x_i),$$

where x_i and y_j must be adjacent pixels, and X and Y represent the object and background of the image, respectively.

The co-occurrence matrix of image F with size of $(m \times n)$ is an $L \times L$ dimensional matrix $T = [t_{ij}]_{L \times L}$ which is defined as:

$$t_{ij} = \sum_{l=1}^{m} \sum_{k=1}^{n} \delta,$$

where

$$\delta = 1, if \begin{cases} f(l,k) = i \quad and \quad f(l,k+1) = j, \\ \qquad\qquad or \\ f(l,k) = i \quad and \quad f(l+1,k) = j; \end{cases}$$

$\delta = 0$, otherwise.

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

$$p_{ij} = t_{ij} / \left(\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} t_{ij} \right), \tag{3.24}$$

where $0 \leq p_{ij} \leq 1$.

The co-occurrence matrix is then divided into four quadrants, namely A,B,C and D (See Figure 3.5).

They are defined as:

$$\begin{aligned} P_A &= \sum_{i=0}^{s} \sum_{j=0}^{s} p_{ij}, \qquad P_B = \sum_{i=0}^{s} \sum_{j=s+1}^{L-1} p_{ij}, \\ P_C &= \sum_{i=s+1}^{L-1} \sum_{j=0}^{s} p_{ij}, \quad P_D = \sum_{i=s+1}^{L-1} \sum_{j=s+1}^{L-1} p_{ij}. \end{aligned} \tag{3.25}$$

and

$$p_{ij}^A = \frac{p_{ij}}{P_A} = \frac{t_{ij} \big/ \left(\sum_i \sum_j t_{ij} \right)}{\sum_{i=0}^{s} \sum_{j=0}^{s} \left\{ t_{ij} \big/ \left(\sum_i \sum_j t_{ij} \right) \right\}} = \frac{t_{ij}}{\sum_{i=0}^{s} \sum_{j=0}^{s} t_{ij}},$$

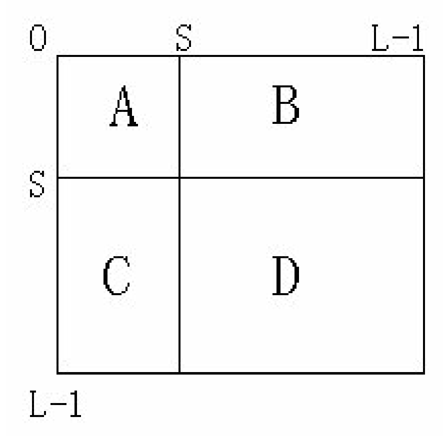

Figure 3.5: Quadrants of co-occurrence matrix.

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

$$p_{ij}^C = \frac{p_{ij}}{P_C} = \frac{t_{ij}}{\sum\limits_{i=s+1}^{L-1}\sum\limits_{j=0}^{s} t_{ij}},$$

$$p_{ij}^D = \frac{p_{ij}}{P_D} = \frac{t_{ij}}{\sum\limits_{i=s+1}^{L-1}\sum\limits_{j=s+1}^{L-1} t_{ij}}.$$

The second order local entropy of the object is defined as

$$H_A^{(2)}(s) = -\frac{1}{2}\sum_{i=0}^{s}\sum_{j=0}^{s} p_{ij}^A \log_2 p_{ij}^A, \tag{3.26}$$

while the second order entropy of background is written as

$$H_D^{(2)}(s) = -\frac{1}{2}\sum_{i=s+1}^{L-1}\sum_{j=s+1}^{L-1} p_{ij}^D \log_2 p_{ij}^D. \tag{3.27}$$

The total second order local entropy of the object and the background is

$$H_T^{(2)}(s) = H_A^{(2)}(s) + H_D^{(2)}(s). \tag{3.28}$$

Threshold s is selected as the one corresponding to the maximum value of $H_T^{(2)}(s)$.

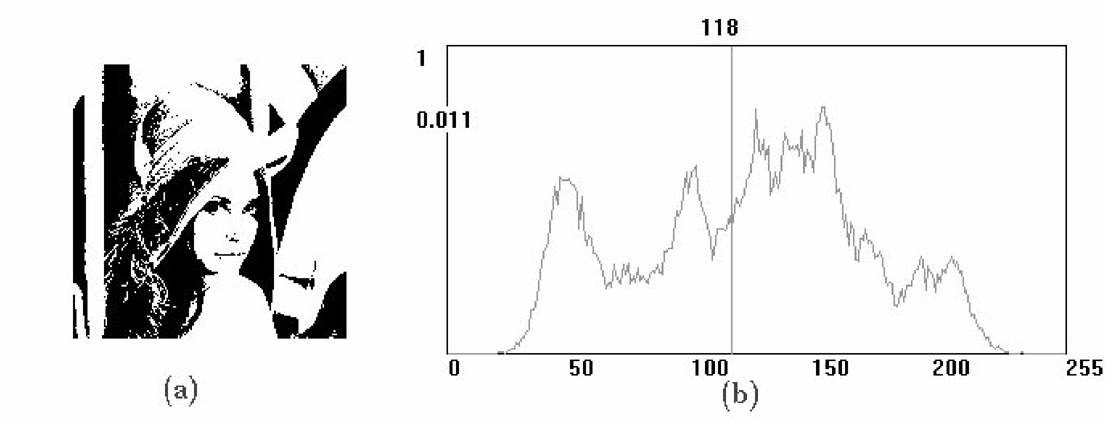

Figure 3.6: Experiment result with Pal and Pal's first method. (a) Binarized image of Lena; (b) Threshold generated by Pal and Pal's method is 118.

Algorithm 2

Based on the concept of conditional entropy, another algorithm is also proposed. The conditional entropy of the image is defined as

$$H_T^{(c)} = \frac{1}{2}\left(H(O|B) + H(B|O)\right), \tag{3.29}$$

where

$$H(O|B) = -\sum_{i=0}^{s}\sum_{j=s+1}^{L-1} p_{ij}^{B}\log_2 p_{ij}^{B}, \tag{3.30}$$

and

$$H(B|O) = -\sum_{i=s+1}^{L-1}\sum_{j=0}^{s} p_{ij}^{C}\log_2 p_{ij}^{C}. \tag{3.31}$$

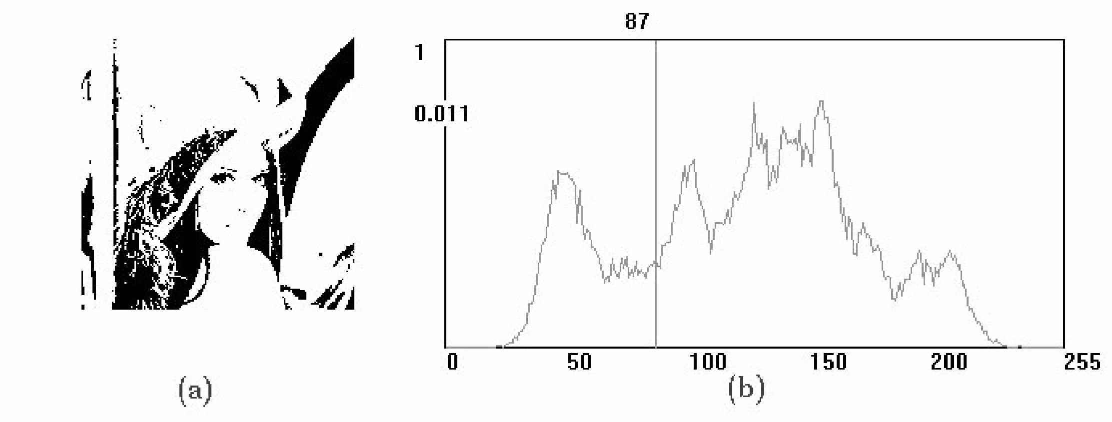

Figure 3.7: Experiment result with Pal and Pal's second method. (a) Binarized image of Lena; (b) Threshold generated by Pal and Pal's second method is 87.

Similarly, the threshold is the one which leads to the maximum $H_T^{(c)}$.

3.5.4 Johannsen And Bille's Method

Johannsen and Bille's method [39] uses the entropy of the gray level histogram of the digital image. The threshold is selected to minimize the interdependence (in information theoretic sense) between the two parts divided by the threshold. Mathematically, it is expressed as:

$$H_b^{(s)} = \log_e\left(\sum_{i=0}^{s} p_i\right) - \frac{1}{\sum\limits_{i=0}^{s} p_i}\left[p_s \log_e p_s + \left(\sum_{i=0}^{s-1} p_i\right)\log_e\left(\sum_{i=0}^{s-1} p_i\right)\right]$$

and

$$H_w^{(s)} = \log_e\left(\sum_{i=s}^{n-1} p_i\right) - \frac{1}{\sum\limits_{i=s}^{n-1} p_i}\left[p_s \log_e p_s + \left(\sum_{i=s+1}^{n-1} p_i\right)\log_e\left(\sum_{i=s+1}^{n-1} p_i\right)\right].$$

.

The interdependence between the black and the white parts is defined as:

$$S = H_b^{(s)} + H_w^{(s)}.$$

The threshold s is selected as the one with the minimum value of S.

The experiment result with Johannsen and Bille method is shown in Figure 3.8.

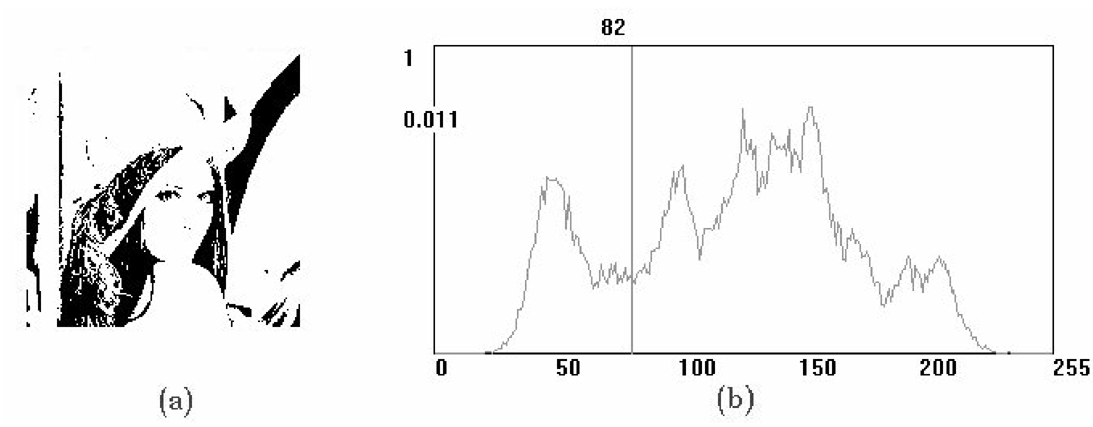

Figure 3.8: Experiment result with Johannsen and Bille method. (a) Binarized image of Lena; (b) Threshold is 82.

3.6 Summary

In this chapter, a brief review of entropy, fuzzy sets and some threshold selection methods based on them are given. We can see that entropy has been widely applied into the image processing field, as well as in the communication area, since it was introduced by Shannon into communications theory in 1948. In this thesis entropy is used as the criterion to select the optimal threshold based on the Maximum Entropy theory. Fuzzy set models are also investigated in order to be applied to image thresholding. Some threshold selection methods based on entropy criterion are also reviewed to see how the different definitions of entropy affect the selection of the threshold.