Image Thresholding of Historical Documents Using Entropy and ROC Curves

Carlos A.B. Mello and Antonio H.M. Costa

Department of Computing Systems, Polytechnic School of Pernambuco,
Rua Benfica, 455, Madalena, Recife, PE, Brazil
cabm@dsc.upe.br

Abstract. It is presented herein a new thresholding algorithm for images of historical documents. The algorithm provides high quality binary images using entropy information of the images to define a primary threshold value which is adjusted with the use of ROC curves.

1 Introduction

Thresholding or binarization is a conversion from a color image to a bi-level one. This is the first step in several image processing applications. This process can be understood as a classification between objects and background in an image. It does not identify objects; just separate them from the background. This separation is not so easily done in images with low contrast. For these cases, image enhancement techniques must be used first to improve the visual appearance of the image. Another major problem is the definition of the features that are going to be analyzed in the search of the correct threshold value which will classify a pixel as object or background. The final bi-level image presents pixels whose gray level of 0 (black) indicates an object (or the signal) and a gray level of 1 (white) indicates the background. With document images, the background can be seen as the paper of the document and the object is the ink.

When the images are from historical documents this problem is quite singular. In these cases, the paper presents several types of noise. In some documents, the ink has faded; some of the others were written on both sides of the paper presenting ink-bleeding interference. A conversion into a bi-level image of this kind of documents using a nearest color threshold algorithm does not achieve high quality results. Thus ink and paper separation is not always a simple task.

In this work, we analyze the application of the thresholding process to generate high quality bi-level images from grey-scale images of documents. The images are of letters, documents and post cards from the end of the 19th century and beginning of the 20th century. The Image Processing of Historical Documents Project (DocHist) aims at the preservation of and easy access to the content of a file of thousands of documents.

In the complete file, there are documents written on one side or on both sides of the sheet of paper. In the latter case, two classes are identified: documents with or without back-to-front interference.

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 905 – 916, 2005.

The second class is the most common and it is easy to reduce the color palette suitably. The bi-level image can be generated from the grayscale one through the application of a threshold filter. A neighborhood filter [15] can also be used to reduce the "salt-and-pepper" noise in the image.

Palette reduction of documents with ink-bleeding interference is far more difficult to address. A straightforward threshold algorithm does not eliminate all the influence of the ink transposition from one side to the other in all cases.

It is presented herein a variation on a previous entropy-based algorithm [12]. It is used to define a primary threshold value which is adjusted using Receiver Operating Characteristic (ROC) curves [13].

2 Materials and Methods

This research takes place in the scope of the DocHist Project for preservation and broadcasting of a file of thousand of historical documents. The bequest is composed of more than 6,500 letters, documents and post cards which amounts more than 30,000 pages.

To preserve the file, the documents are digitized in 200 dpi resolution in true color and stored in JPEG file format with 1% loss for better quality/space storage rate. Even in this format each image of a document reaches, in average, 400 Kb. Although broadband Internet access is a common practice nowadays, the visualization of a bequest of thousand of files is not a simple task. Even in JPEG file format all the bequest must consume Giga bytes of space. There are new mobile devices which are not suitable to access large files as palm tops or PDA´s (Personal Digital Assistants).

A possible solution to this problem is to convert the images to bi-level which is not a simple task. As said before, some documents are written on both sides of the paper creating back-to-front interference; in others the ink has faded. Thus, the binarization by commercial softwares with standard settings is not appropriate. Figure 1 presents a sample document and its bi-level version produced by straightforward threshold algorithms.

Besides compression rates, high quality bi-level images yield better response from OCR tools. This allows the use of text files to make available the contents of the documents instead of its full digitized image.

The problem remains in the generation of these bi-level images from the original ones. For this, an entropy-based segmentation algorithm was proposed and extended with variations in the logarithmic basis [12].

2.1 Thresholding Algorithms

There are several algorithms for thresholding purposes. The first ones were based on simple features of the images or their histograms. The mean of the grayscale histogram is used as cut-off value in the thresholding by mean gray level [15]. Another algorithm is based on the percentage of black pixels desired [15] (10% is the value suggested in [15]). In the two peaks algorithm, the threshold occurs at the low point between two peaks in the histogram [15]. In adaptive algorithms, the iterative selection [17] makes an initial guess at a threshold value which is refined improving this value. The initial guess is the mean gray level which separates two areas and the mean

values of these areas are evaluated (Tb and To). A new estimative of the threshold is evaluated as (Tb + To)/2. The process repeats with this new value of threshold until no change is found in the value in two consecutives steps.

It is presented herein some of the most well-known thresholding algorithms, which are classified based on the type of information used. The taxonomy used herein defines three categories of thresholding algorithms based on *histogram entropy*, *maximization or minimization functions* and *fuzzy theory*.

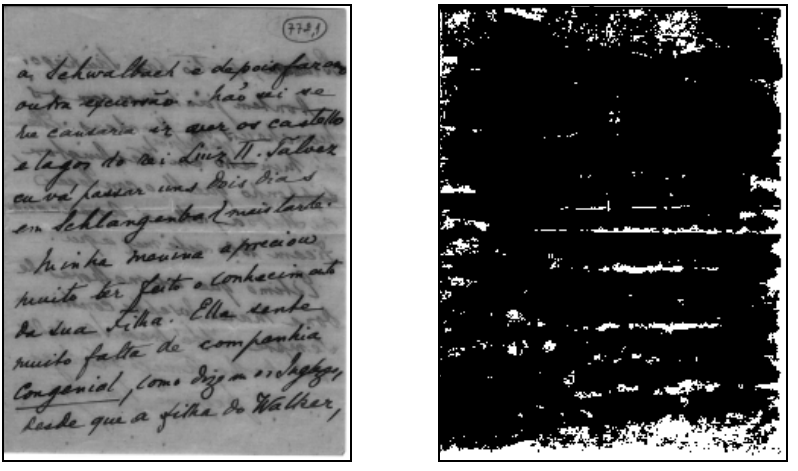

Fig. 1. (left) Grayscale sample document written on both sides of the paper and (right) its bi-level version by a threshold algorithm

Entropy [19] is a measure of information content. In Information Theory, it is assumed that there are n possible symbols $\boldsymbol{s}$ which occur with probability $p(s)$. The entropy associated with the source S of symbols is:

$$H(S) = -\sum_{i=0}^{n} p[s_i]\log(p[s_i])$$

where the entropy can be measured in bits/symbols. Although the logarithmic base is not defined, [7] and [10] analyze that changes in the base do not affect the concept of entropy as it was explored in [12].

Six entropy-based segmentation algorithms are briefly described herein: Pun [16], Kapur *et al* [6], Johannsen [5], Li-Lee [11], Wu-Lu [20] and Renyi [18].

Pun's algorithm [16] analyses the entropy of black pixels, *Hb*, and the entropy of the white pixels, *Hw*, bounded by the threshold value *t*. The algorithm suggests that *t* is such that maximizes the function $H = Hb + Hw$, where *Hb* and *Hw* are defined by:

$$Hb = -\sum_{i=0}^{t} p[i]\log(p[i]) \qquad \text{(Eq. 1)}$$

$$Hw = -\sum_{i=t+1}^{255} p[i]\log(p[i]) \qquad \text{(Eq. 2)}$$

where *p[i]* is the probability of pixel *i* with color *color[i]* is in the image.

In [6], Kapur *et al* defined a probability distribution *A* for an object and a distribution *B* to the background of the document image, such that:

A: p0/Pt, p1/Pt, ..., pt/Pt
B: (pt+1)/(1 – Pt), (pt + 2)/(1 - Pt),..., p255/(1 – Pt)

The entropy values *Hw* and *Hb* are evaluated using Equations 1 and 2 with *p[i]* defined with these new distributions. The maximization of the function *Hw* + *Hb* is analyzed to define the threshold value *t*.

Another variation of an entropy-based algorithm is proposed by Johannsen and Bille [5] trying to minimize the function *Sb(t)* + *Sw(t)*, with:

$$S_w(t) = \log(\sum_{i=t+1}^{255} p_i) + (1/\sum_{i=t+1}^{255} p_i)[E(p_t) + E(\sum_{i=t+1}^{255} p_i)]$$

and

$$S_b(t) = \log(\sum_{i=0}^{t} p_i) + (1/\sum_{i=0}^{t} p_i)[E(p_t) + E(\sum_{i=0}^{t-1} p_i)]$$

where *E(x)=-xlog(x)* and *t* is the threshold value.

The Li-Lee algorithm [11] uses the minimum cross entropy thresholding, where the threshold selection is solved by minimizing the cross entropy between the image and its segmented version.

The basic idea of the Wu-Lu algorithm is the use of the lower difference between the minimum entropy of the objects and the entropy of the background [20]. The method is very useful in ultra-sound images which have few different contrast values.

The Renyi method [18] uses two probability distribution function (one for the object and the other for the background), the derivatives of the distributions and the methods of Maximum Sum Entropy and Entropic Correlation.

Other algorithms are based on the maximization or minimization of functions. Although Kapur and Johannsen algorithms, presented previously, work in the same way, they were classified as Entropy algorithms because of the major importance of this feature in them. For this category of algorithms, five techniques are selected.

The Brink method [8] identify two threshold values (T1 and T2), using the Brink´s maximization algorithm. The colors below T1 are turned to black and the colors above T2 are turned to white. The values between T1 and T2 are colorized analyzing the neighbors of the pixel. A 25x25 area is analyzed and, if there is a pixel in this area which color is greater than T2, then the pixel is converted to white.

In the Minimum Thresholding algorithm, Kittler and Illingworth [9] use the histogram as a measured probability density function of two distributions (object and background pixels). The minimization of a criterion function defines the threshold.

Fisher method [1] consists in the localization of the threshold values between the gray levels classes. These threshold values are found using a minimization of the sum of the inertia associated to the two different classes.

In the Kittler and Illingworth Algorithm based on Yan´s Unified algorithm [22] the foreground and background class conditional probability density functions are assumed to be Gaussian, but in contrast to the previous method the equal variance assumption is removed. The error expression can be interpreted also as a fitting expression to be minimized.

Otsu [14] suggested minimizing the weighted sum of within-class variances of the foreground and background pixels to establish an optimum threshold. The algorithm has its basics in the discriminant analysis. The segmentation is done using the mean values of the foreground and background classes (μ_b and μ_w, for the pixels classified as ink or paper, respectively), of the between-classes variances σ_b^2, within-classes variances σ_w^2 and total variance σ_T^2. Otsu demonstrated that the optimal value of the threshold *t** can be reached by the maximizing the function $\eta(t) = \frac{\sigma_b^2(t)}{\sigma_T^2}$, *i.e.*, the ratio between the variance between-classes and the total variance.

In a fuzzy set, an element x belongs to a set S with probability p_x. This definition of fuzzy sets can be easily applied to the segmentation problem. Most of the algorithms use a measure of fuzziness which is a distance between the original gray level image and the segmented one. The minimization of the fuzziness produces the most accurate binarized version of the image. We can cite three binarization algorithms that use fuzzy theory: C Means [4], Huang [3] and Yager [21].

In addition, there is also the Ye-Danielsson [2] algorithm which is implemented as an iterative thresholding.

Fig. 2 presents the application of these algorithms in the sample document of Fig. 1. It can be observed that some algorithms performance was very poor as some images are completely black or white.

2.2 Entropy-Based Segmentation Algorithm

At first, the algorithm scans the image in search for the most frequent color, t. As we are working with images of letters and documents, it is correct to suppose that this color belongs to the paper. This color is used as an initial threshold value for the evaluation of Hb and Hw as defined in Eq. 1 and 2 before.

As defined in [7], the use of different logarithmic bases does not change the concept of entropy. This base is taken as the area of the image: width by height.

With *Hw* and *Hb*, the entropy, *H*, of the image is evaluated as their sum:

$$H = Hw + Hb\,. \qquad \text{(Eq. 3)}$$

Based on the value of *H*, three classes of documents were identified, which define two multiplicative factors, as follows:

- $H \leq 0.25$ (documents with few parts of text or very faded ink), then $mw = 2$ and $mb = 3$;
- $0.25 < H < 0.30$ (the most common cases), then $mw = 1$ and $mb = 2.6$;
- $H \geq 0.30$ (documents with many black areas), then $mw = mb = 1$.

These values of *mw* and *mb* were found empirically after several experiments where the hit rate of OCR tools in typed documents (as the one of Fig. 3-left) defined the correct values. With the values of *Hw*, *Hb*, *mw* and *mb* the threshold value, *th*, is defined as:

$$th = mw.Hw + mb.Hb\,. \qquad \text{(Eq. 4)}$$

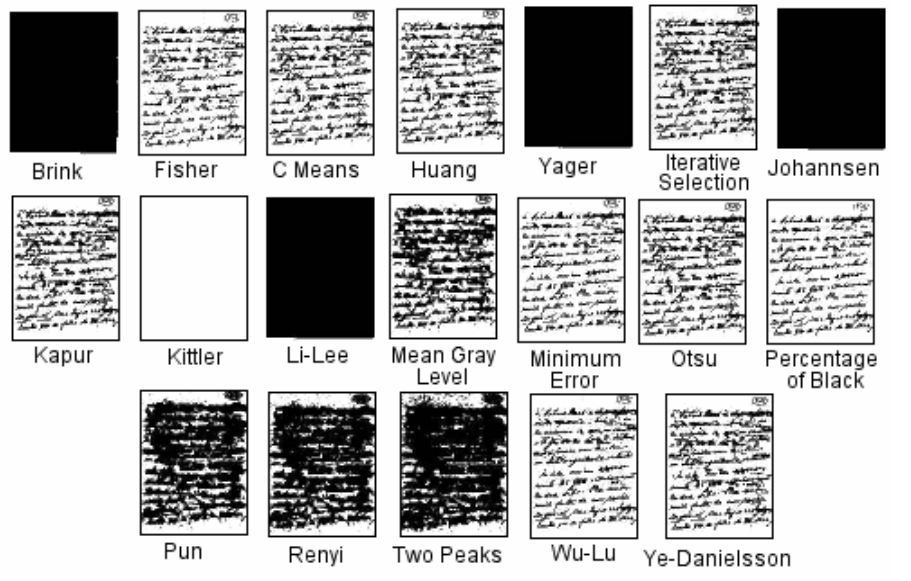

Fig. 2. Application of several thresholding algorithms in document presented in Fig. 1 with back-to-front interference

The grayscale image is scanned again and each pixel *i* with *graylevel[i]* is turned to white if:

$$(graylevel[i]/256) \geq th\,. \qquad \text{(Eq. 5)}$$

Otherwise, its color remains the same (to generate a new grayscale image but with a white background) or it is turned to black (generating a bi-level image). This is called the *segmentation condition*.

Fig. 3 presents a zooming into a document and its binarized version generated by the entropy-based algorithm.

The problem comes when the images have back-to-front interference. As it can be seen in Fig. 4, the results of the algorithm are not the best, even though it is far better

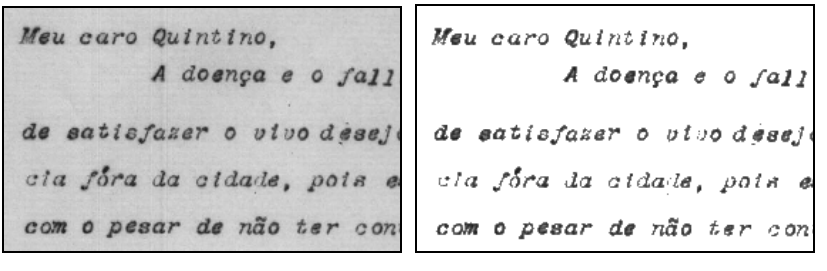

Fig. 3. (left) Sample document and (right) its bi-level version by entropy algorithm

than other ones. It can be noticed in Fig. 4-left that the bi-level image presents some elements of the opposite side of the paper, although its quality is much better than the one created by a straightforward thresholding algorithm (Fig. 4-center). The correction of this threshold value is proposed in the next Section with the use of ROC curves.

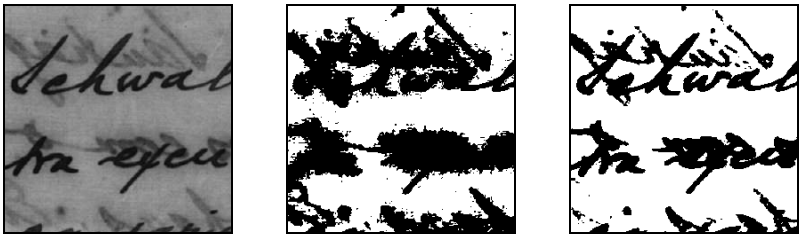

Fig. 4. (left) Sample document with back-to-front interference, (center) binarized image using a nearest color thresholding algorithm with default values and (right) bi-level image generated by new entropy-based algorithm

2.3 Thresholding by ROC Curves

The threshold value defined by the entropy-based algorithm is not always the best value. So, to adjust this value, it used a receiver operating characteristic (ROC) curve from Detection Theory [13]. This is usually used in medical analysis where some tests can generate *true positives* (TP), *false positives* (FP), *true negatives* (TN) and *false negatives* (FN) answers. TP represents the number of patients who have some disease, and have this corroborated by having a "high" test (above some chosen cutoff level). FP represents false positives - the test was wrong, and resulted that non-diseased patients are really ill. Similarly, true negatives are represented by TN, and false negatives by FN.

In elementary statistical texts, some will encounter other terms:

- The sensitivity is how accurate the test is at picking out patients with the disease. It is simply the True Positive. In other words, sensitivity gives us the proportion of cases picked out by the test, relative to all cases that actually have the disease.
- Specificity is the ability of the test to pick out patients who do not have the disease. This is synonymous with the True Negative.

A receiver operating characteristic (ROC) curve shows the relationship between probability of detection (PD) and probability of false alarm (PFA) for different threshold values. The two numbers of interests are the probability of detection (TP) and the probability of false alarms (FP). The probability of detection (PD) is the probability of correctly detecting a Threat user. The probability of false alarm (PFA) is the probability of declaring a user to be a Threat when s/he is Normal. The detection threshold is varied systematically to examine the performance of the model for different thresholds. Varying the threshold produces different classifiers with different (PD)

and probability of false alarm (PFA). By plotting PD and PFA for different thresholds values, one can get a ROC curve.

For thresholding applications, this theory can be easily adapted as one can see the TP as the ink pixels correctly classified as object; FP represents background elements classified as object, and so on.

The new proposed algorithm starts with the application of the previous entropy-based algorithm. This initial threshold value (th) is used to define a binary matrix (M) with the same size of the input image. Each cell of this matrix is set to *true* if the corresponding pixel in the input image (IM) is equal to th. This leads to the building of the PD *versus* PFA curve (the ROC curve) according to algorithm 1.

Algorithm 1

n1 ← the number of true elements in M (elements equal to th in IM)
n0 ← the number of false elements in M (elements different to th in IM)
for t = 0 to 255
    pd(t) ← $\sum$ (IM > t **AND** M)/n1
    pfa(t) ← $\sum$ (IM > t **AND** ¬M)/n0
end

For our kind of images, the ROC curve defined by this algorithm is a step like function which has its maximum values equal to 1 for both axes. Different initial threshold values define different ROC curves.

Fig. 5 presents the PD *versus* PFA curve for the sample image of Fig. 4-left. For this document, th = 104 and PFA is equal to 1 when PD is 0.758.

One can see in the bi-level image (Fig. 4-right) that there are still many elements of the ink that is in the other side of the paper. So this cut-off value is not the best one.

It was observed in the handwritten documents that the percentage of ink is about 10% of the complete image. So, the correct ROC curve must grow to 1 when PD values about 0.9. For this, different values of th must be used. This creates different M matrixes leading to new PDxPFA curves. If the curve grows to 1 with PD less than 0.9, then the initial th must decrease; otherwise, it must increase. Fig. 6 presents some resulting images for different th and the PD value which turns PFA equals to 1, starting from the initial th = 104, and PD = 0.758 (present in Fig. 5).

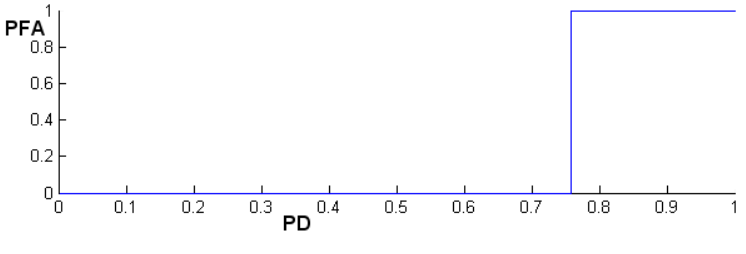

Fig. 5. (top-left) Original document with back-to-front interference. (top-right) Binarized version generated with th = 104. (bottom) PD *versus* PFA graphic; PFA = 1 for PD = 0.756.

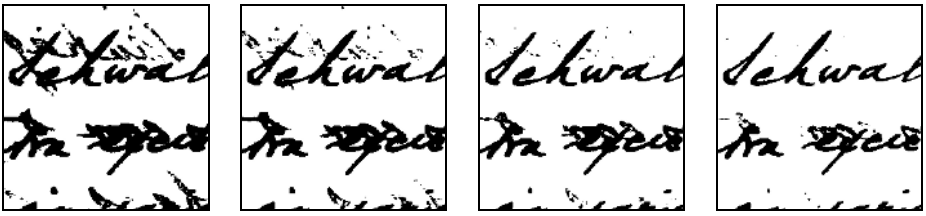

Fig. 6. Bi-level images generated by different threshold values (th) and the corresponding PD value for which PFA turns equal to 1: a) th = 100, PD = 0.771, b) th = 90, PD = 0.8244, c) th = 80, PD = 0.8534 and d) th = 70, PD = 0.8749

3 Results

For the sample document of Fig. 4, the initial threshold value is 104 and, as it could be seen, it did not result a good quality image. For this th, PD is 0.756 (Fig. 5). So, the th value must be decreased until PD equals to 0.9. In fact, a small variation of this PD value is accepted. Changing the th value, PD reaches the value of 0.8983 (when PFA turns from 0 to 1) with th = 57. The final PD *versus* PFA graphic just as the final bi-level image of the sample document of Fig. 4 are shown in Fig. 7.

Fig. 8 presents others sample documents, their bi-level images generated by the entropy-based algorithm with and without the ROC correction and the threshold values defined (initial and final).

As can be seen in Fig. 8, the correction achieved better quality images for all cases. The same happened with images without back-to-front interference. But, in these cases, the difference between the initial threshold value and the final one is smaller. Thus, the correction can be applied to every case.

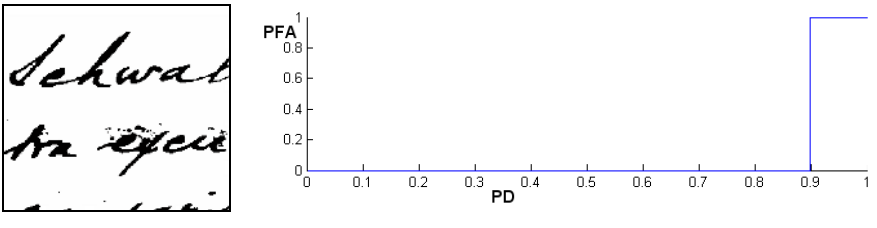

Fig. 7. (left) Final bi-level version of document presented in Fig. 4-top-left after correction by ROC curve. (right) PD *versus* PFA graphic. The threshold value is now 57, with PD = 0.8983.

4 Conclusions

This paper presents a variation of an entropy-based thresholding algorithm for images of historical documents. The algorithm defines an initial threshold value which is adjusted by the use of ROC curves. These adjustments define new cut-off values and they generate better quality bi-level images. The method is quite suitable when

Fig. 8. (left) Sample original documents and bi-level images generated by entropy-based threshold algorithm (center) without and (right) with ROC correction

applied to documents written on both sides of the paper, presenting back-to-front interference. By visual inspection, the binary images are far better than the ones produced by others well-known algorithm.

The monochromatic images can be used to make files of thousand of historical documents more easily accessible by the Internet even through mobile devices which have slower connections.

A MatLab implementation of the proposed algorithm just as a sample image of a document is avaliable at: http://www.upe.poli.br/dsc/recpad/site_hist/throc.htm

Acknowledgments

This research is partially sponsored by CNPq (PDPG-TI 55.0017/2003-8), FACEPE and University of Pernambuco.

References

1. M.S.Chang, S.M.Kang, W.S.Rho, H.G.Kim and D.J.Kim. "Improved binarization algorithm for document image by histogram and edge detection", *Proc. 3rd Intern. Conf .on Document Analysis and Recognition*, pp.636-639, Canada, 1995.
2. C.A.Glasbye, "An Analysis of Histogram-Based Thresholding Algorithm", CVGIP: Graphical Models and Image Processing, nov 1993.
3. L.K.Huang L.K and M.J.Wang, "Image Thresholding by Minimizing the Measures of Fuzziness", Pattern Recognition, 1995.
4. C.V.Jawahar, P.K.Biswas and K.Ray, "Investigations On Fuzzy Thresholding Based On Fuzzy Clustering", Pattern Recognition, 1997.
5. G.Johannsen and J.Bille, "A Threshold Selection Method using Information Measures", *Proceedings, 6th Int. Conf. Pattern Recognition*, Munich, Germany, pp.140-143, 1982.
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", Computer Vision, Graphics and Image Processing, 29(3), 1985.
7. J. N. Kapur, *Measures of Information and their Applications*, John Wiley and Sons, 1994.
8. S.W.Katz and A.D.Brink, "Segmentation of Chromosome Images", IEEE, 1993, pp 85-90.
9. J.Kittler and J.Illingworth, "Minimum Error Thresholding", Pattern Recognition, Volume 19, Issue 1, pp 41-47, 1986.
10. S.Kullback. *Information Theory and Statistics*.Dover Publications, Inc.1997.
11. C.H.Li and C.K.Lee, "Minimum Cross Entropy Thresholding", Pattern Recognition, v.26, no 4, pp 616-626, 1993.
12. C.A.B.Mello. "A New Entropy and Logarithmic Based Binarization Algorithm for Grayscale Images". IASTED VIIP 2004, Hawaii, USA, 2004.
13. N.A.McMilan, C.D.Creelman. *Detection Theory*. LEA Pub., 2005.
14. N.Otsu. "A threshold selection method from gray-level histogram". *IEEE Trans .on Systems, Man, and Cybernetics*, vol 8: 62-66, 1978.
15. J.R.Parker, *Algorithms for Image Processing and Computer Vision*, John Wiley and Sons, 1997.
16. T.Pun, "Entropic Thresholding, A New Approach", C.Graphics and Image Proc., 1981.
17. T.W.Ridler and S.Calvard. "Picture Thresholding Using an Iterative Selection Method", *IEEE Trans. on Systems, Man and Cybernetics*, Vol.SMC-8, 8:630-632, 1978

18. P.Sahoo, C.Wilkins and J.Yeager, “Threshold Selection using Renyi’s Entropy”, Pattern recognition Vol 30, No 1, pp 71-84, 1997
19. C.Shannon. “A Mathematical Theory of Communication”. Bell System Technology Journal, vol. 27, pp. 370-423, 623-656, 1948.
20. Lu Wu, Songde Ma, Hanqing Lu, “An Effective Entropic thresholding for Ultrasonic Images”, IEEE, pp.1552-1554, 1998.
21. R.R.Yager, “On the Measures of Fuzziness and Negation.Part.1: Membership in the Unit Interval”, Int Journal of Gen. Sys, 1979.
22. H.Yan, “Unified Formulation of a Class of Image Thresholding Techniques”, Pattern Recognition, Vol. 29, No 12, pp 2025-2032, 1996.