International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Evaluation of Thresholding Algorithms for Document Images

M. S. Sonawane[1], Dr.C.A.Dhawale[2]

[1]Research Scholar, SGBAU, Amravati (M.S.), *manojkumar.sonawane@rediffmail.com*

[2]P.R. Pote College of Engineering andManagement, Amravati (M.S.), *cadhawale@rediffmail.com*

Abstract:-*Document image processing consists of several phases. One important phase is preprocessing on which accuracy of remaining phases relies. Binarization is one of sub phases that belong to preprocessing. Binarization is separation of foreground text from background of document image. Various thresholding algorithms are exists for binarization of document images. Thresholding algorithms are divided into three categories which are global thresholding, local thresholding and hybrid thresholding. In this paper 5 global thresholding, 5 local thresholding and 1 hybrid thresholding algorithms are evaluated by using different assessment parameters. No single thresholding algorithm can resolve every complication; however some algorithms are superior to others for particular circumstances.*

Keywords:-Thresholding, binarization, evaluation.

1. Introduction

Optical Character Recognition, frequently shortened to OCR, is renovation of scanned images of printed text, handwritten text into computer readable text. OCR is widespread technique of computerizing printed data such that they can be automatically searched, trimly saved, shown online, and exercised in computer. There are number of phases while doing recognition such as image acquisition, preprocessing, segmentation, feature extraction, classification etc. Preprocessing consist of noise removal, binarization, skew correction, size normalization, boundary detection, thinning etc. This is depicted in following figure-1. Correct and speedy binarization method is vital for OCR [1]. Binarization is introductory method and the subsequent binary images typically affect accuracy of the later procedures like document image segmentation, recognition.

Aim of binarization is to separate out foreground text from background of document image. Pixels within characters, curves, lines are foreground pixels and must be binarized as black pixels whereas other background pixels must be binarized as white pixels [2].Historical documents undergo through numerous degradations because of ageing, prolonged utilization, unclear characters shadows, odd enlightenment, bleed-through, smudge, strain etc. These things are tricky for document image analysis techniques [3]. In order to carry out binarization there are numerous thresholding algorithms. Thresholding algorithms are mainly classified into three categories; Global thresholding algorithms, Local thresholding algorithms and hybrid thresholding algorithms.

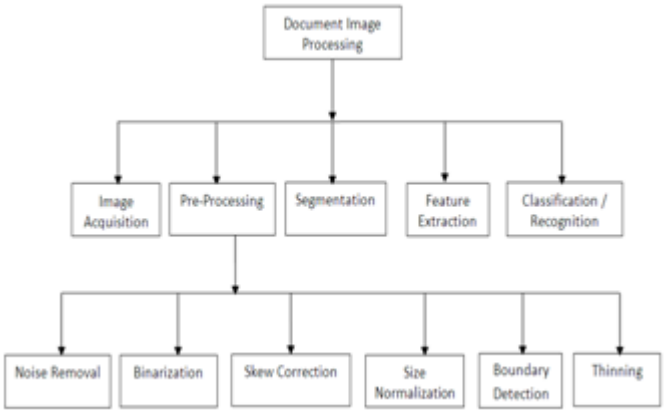

Figure 1: Image Processing Phases Tree

2. Global Thresholding Algorithms

Global thresholding algorithms employ techniques founded on classification procedures, histogram, clustering, entropy, Gaussian distribution etc. [4]. There exists number of global thresholding algorithms. Algorithms are categorized into

2.1. Classification Procedures

One of the popular global thresholding algorithms is of Otsu's. Otsu's algorithm belongs to first category that is classification procedures. Otsu [Otsu 1979] has considered a global threshold by admitting the presence of 2 categories foreground and background and selecting the threshold that decreases the interclass variance of the thresholded black, white pixels. Reddi et al. [Reddi et al 1984] method can be deliberated as an extension of Otsu method for multithresholding purpose. In this effort they have used Otsu's method as a global thresholding method. Its aim is the maximization of the interclass variance. IIFA that is Improved Integrated Function Algorithm [Trier and Taxt 1995] relates a gradient like operator defined as the activity A(x, y), which the absolute addition of approximated derivatives for raster and scan directions is taken over a

slight area of image. 3-level label image with pixel levels '+', '-', '0' is formed. All '+' indicated regions are labeled as print, '-' indicated regions are labeled as background, '0'indicated regions are labeled print if a most of pixels with 4-connected are '+' indicated else it is labeled as background [5].

2.2. Histogram based

Global thresholding algorithm that belongs to second category is Histogram peaks [Prewitt and Mendelsohn 1966]. Histogram peak is ordinarily used global thresholding method. This method depends on analysis of histogram. It considers a bimodal histogram. The histogram is leveled by using three point mean filter repeatedly till it has merely 2 local maxima. Black percentage [Doyle 1962] is a parametric procedure which considers that the percentage of black pixels is known (p). The histogram is utilized and the threshold is fixed to the highest gray-level which maps at least (100 − p) % of the pixels into the background class. Here p=5 is set. Ramesh et al. [Ramesh et al 1995] utilize an easy functional approximation to the PMF comprising of a 2 step function. Hence addition of squares between histogram and bi-level function is minimized and the solution is achieved by iterative search. Rosenfeld and Kak [Rosenfeld and Kak 1982] choose global threshold from histogram of 2 dimensional images. They consider that gray values of every entity are probable to cluster around a peak of the histogramof 2 dimensional images and attempt to calculate the position of peaks or valley straightly from histogram [6].

2.3. Clustering based

Clustering based global thresholding algorithm is K-means [Jain and Dubes 1988]. In k-means algorithm gray level samples are bundled into 2 parts that are foreground and background using corresponding clustering algorithm. Likewise fuzzy c-means [Duda and Hart 1973] is a fuzzy clustering methodology where gray scale values are bundled into 2 fuzzy classes corresponding to foreground and background pixels [7].

2.4. Entropy based

Global thresholding algorithm that belongs to forth category is of Pun's. Pun [Pun 1980] assumes gray level histogram as a G-symbol source; in which whole symbols are statistically not dependent. He assumes the proportion of the posteriori entropy as a function of the threshold to that of the source entropy. Yen et al. [Yen 1995] explain the entropic correlation and gain the threshold that maximizes it [8].

2.5. Gaussian Distributions

Kittler and Illingworth [Kittler and Illingworth 1985] demonstrate a procedure that depends on fitting of mixture of Gaussian distributions and it converts problem of binarization into a minimum error Gaussian density fitting problem. Likewise Lloyd´s [Lloyd 1985] method assumes equal variance Gaussian density functions and minimizes the total misclassification error through an iterative examination. Lastly, Riddler and Calvard [Ridler and Calvard 1978] by repeated thresholding progresses one of the firstiterative techniques based on 2 class Gaussian mixture models. At iteration n, a different threshold Tnis recognized using the average of the background and foreground class means. In general, iterations end whenever the variations|Tn-Tn+1| become sufficiently lesser [9].

3. Local Thresholding Algorithms

There are so many local binarization algorithms. Local binarization algorithms are categorized into

3.1. Clustering Procedures

Algorithm that belongs to clustering procedures is Kohonen SOM [Papamarkos and Atsalakis 2000], which suggest that neural network could be utilized for common gray scale reduction. Precisely, gray level nourishes neural network classifier of Kohonen SOM. When training is done, the neurons of the output competition layer describe the gray level classes. If output layer takes only 2 neurons then bi-level clustering is carried out. It means after completion of training phase, the output neurons identify 2 classes found. Afterwards by using a mapping procedure, these classes are considered as classes of the background, foreground pixels [10].

3.2. Local Variation

For local variation category, Niblack's algorithm exists. Niblack [Niblack 1986] determines value of local threshold for every pixel that relies on local mean value, local standard deviation in the pixel neighborhood. A constant decides how much of entire print object edge is considered as a portion of the assumed object. The neighborhood size must be small adequate to keep local, large adequate to defeat noise. It is observed that a neighborhood *15x15* is a decent selection. Likewise Sauvola [Sauvola and Pietikainen 2000] determines value of local threshold that relies on local mean value, local standard deviation in the pixel neighborhood, but has utilized difficult formula. Bernsen [Bernsen 1986] too get local thresholding, computed by mean value of the minimum, maximum values within a window surrounding pixel. If difference of 2 values is greater than a threshold then pixel is considered as a part of foreground else of background and takes a default value [11].

3.3. Entropy

Abutaleb [Abutaleb 1989] algorithm belongs to third category. It uses a local method that assumes joint entropy of 2 associated random variables; image gray value at a pixel, and average gray value of a neighborhood centered at that pixel. By use of 2D histogram, for every threshold twosomes, one could compute cumulative distribution and then explain foreground entropy. Brink and Pendock [Brink and Pendock 1996] advise an amendment of Abutaleb's method by redefinition of class entropies and discovering threshold as value that exploits the minimum of background, foreground entropies. A local method is analogous to earlier ones is also considered by Kapur et al. [Kapur et al 1985]. The maximization of the entropy of the thresholded image is taken as signal of maximum statistics transfer. The background, foregrounds are assumed as 2 diverse signal sources, such that if addition of 2 class entropies touches its

maximum then an image is assumed to be optimally thresholded. Johannsen and Bille [Johannsen and Bille 1982] recommend an entropy dependent algorithm which attempt to minimize function *Sb (t) + Sw (t)*, with:

Here *T*=threshold value Whereas E*(x) =-xlog(x)*. [12]

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

$$S_b(T)=\log\left(\sum_{i=0}^{T} p_i\right)+\left(1/\sum_{i=0}^{T} p_i\right)\left[E(p_T)+E\left(\sum_{i=0}^{T-1} p_i\right)\right]$$

3.4. Neighborhood Information

Neighborhood Information dependent algorithm is of Palumbo at al. [Palumbo et al 1986]. Local algorithm comprises measurement of local contrast of five 3x3 neighborhoods structured in a center surround pattern. Parker's [Parker 1991] local technique initially identifies edges and afterwards an area between edges is to be filled. Initially for 8 connected neighborhood of every pixel negative of darkest neighbor D is searched. Afterwards it is fragmented up to regions r x r and for every region sample mean, standard deviations are computed. These 2 values are smooth out and then bilinearly interpolated to produce 2 new images; M and S, devising from mean values, standard deviations. After that for every pixels (x, y), if M(x, y) ≥ m0 or S(x, y) < s0, then pixel is considered as portion of a flat region and keeps unlabeled else if D(x, y) < M(x, y) + kS(x, y), then (x, y) is considered as foreground pixel else (x, y) remains unlabeled. The subsequent binary image shows edges. One more method Adaptive Local Level Thresholding (ALLT) [Yang and Yan 2000] is a local thresholding method. Initially, they investigate connection features of character stroke from run-length histogram for chosen image areas and different heterogeneous gray scale backgrounds. Afterwards they recommend modified logical thresholding technique to mine binary image adaptively from despoiled gray scale document image having complex, heterogeneous background. Gatos et al. [Gatos et al 2006] local technique is designed to consider degradations which happen because of un-uniform enlightenment, smear, strain, low contrast, shadows etc. They considered various discrete phases: a pre-processing phase by utilizing rough estimation of foreground areas, low pass Wiener filter, and background surface computation by interpolating neighboring background intensities etc. [13].

3.5. Otsu's Method

Liu and Li [Liu and Li 1993] developed 2D Otsu thresholding technique. This technique privilege to do better than 1D Otsu technique does, if images are degraded due to noise. This technique computes local average gray level in limited window. They made 2D histogram where x-axis, y-axis are gray value, local average gray level correspondingly. The optimal threshold is chosen at maximum between class variance. Mardia and Hainsworth [Mardia and Hainsworth 1988] provided local technique that accomplishes early binarization by utilizing Otsu's [Otsu 1979] method. Then various phases are iterated till conjunction is reached. Vonikakis et al. [Vonikakis et al 2008] shows a local technique whose key interest is to accept features of OFF ganglion cells of the Human Visual System, use them in binarization process of text [14].

4. Hybrid Thresholding Algorithms

A combination of global and local binarization will give us a hybrid binarization technique. Improved IGT [Kavallieratou 2005] is one of hybrid methodology. Here initially global algorithm is applied on whole document image, followed by application of local thresholding only on needed regions. It's depend on global IGT technique contains of following phases; (a) Perform IGT on entire document image to calculate global threshold value. (b) Fine regions having noise. (c) Reevaluate IGT on every noticed region to compute local threshold value for every region. The IGT involves 2 trials that are smeared interchangeably numerous times. Initially average color value of an image is computed, subtracted from an image. In next part of an algorithm histogram stretching is carried out. So that remaining pixels will enlarge and gain all of gray scale natures. This process is repeated until difference between consecutive thresholds is small adequate [15].

5. Assessment Parameters

For performance evaluation of thresholding algorithms there are numerous assessment parameters. In this paper four assessment parameters are used [16], which are as follows.

5.1. Pixel Error Rate (PERR)

Pixel Error is total number of pixels in output image those have wrong color that is white if black in original image or black if white in original image. Therefore Pixel Error Rate (PERR) will be

PERR=Pixel Error/ (M*N)

5.2. Mean Square Error (MSE)

Suppose x (i, j) denote the value of the i^{th} row and j^{th} column pixel in the original image x and suppose y (i, j) denote the value of the equivalent pixel in the resultant image y. As it is all about black and white images, both values may be either 0 for black or 255for white. The local error is e (i, j) =x (i, j)- y (i, j) and the total square error rate will be

$$MSE=\frac{\sum_i\sum_j e(i,j)^2}{MxN}$$

Remember that if a pixel is accurate color then value of e (i, j) ^{2} will be 0, whereas if the pixel is not accurate color then it will be $(255)^2$. Hence, by considering PERR explanation, it will be PERR=MSE / $(255)^2$

5.3. Signal to Noise Ratio (SNR)

SNR is calculated as the ratio of average signal power to average noise power and for an MxN image is

$$SNR(DB) = 10\log_{10}\frac{\sum_i\sum_j x(i,j)}{\sum_i\sum_j (x(i,j)-y(i,j))^2}$$

$$= 10\log_{10}\frac{\sum_i\sum_j x(i,j)}{MSE} = 10\log_{10}\frac{\sum_i\sum_j x(i,j)}{PERR\cdot 255^2}$$

5.4. Peak Signal to Noise Ratio (PSNR)

The peak measure PSNR relies on word length of an image pixels and it is calculated as the ratio of peak signal power to average noise power. In case of 8 bit images it is

$$PSNR(DB) = 10\log_{10}\frac{255^2\cdot MN}{\sum_i\sum_j (x(i,j)-y(i,j))^2}$$

$$= 10\log_{10}\frac{255^2\cdot MN}{MSE} = 10\log_{10}\frac{MN}{PERR}$$

6. Result Analysis

In section 2 five global thresholding algorithms are discussed. Depends on their work they were classified into categories like classification procedures, histogram based, clustering based, entropy based, Gaussian distributions etc. In section 3 five local thresholding algorithms were described. They were categorized into Clustering Procedures, Local Variation, Entropy, Neighborhood Information, and Otsu's Method etc. Hybrid algorithm is described in section 4. Afterwards in section 5 four assessment parameters are discussed those are used to evaluate the algorithms. For the work fifteen degraded images of 18th century with average intensity are considered. Figure-2 is one of those images. According to [17] as shown in table-1 for PERR measure Pun-1980 algorithm gives highest result (42.82607) whereas Reddi-1984 algorithm gives less result (1.641987). If evaluation is done using MSE then K-means technique of Jain and Dubes-1988 gives excellent output, for the same lower result is 1067.702. While considering SNR parameter best effect is given by Reddi-1984 that is 18.1055. Whenever execution of algorithms is done by using PSNR then finest outcome is 18.31057 which are provided by Reddi-1984. Results of local thresholding algorithms are summarized in table-2. In case of hybrid algorithm, for PERR, MSE, SNR, PSNR measures results are 2.448303, 1592.008, 16.22354, and 16.31476 respectively.

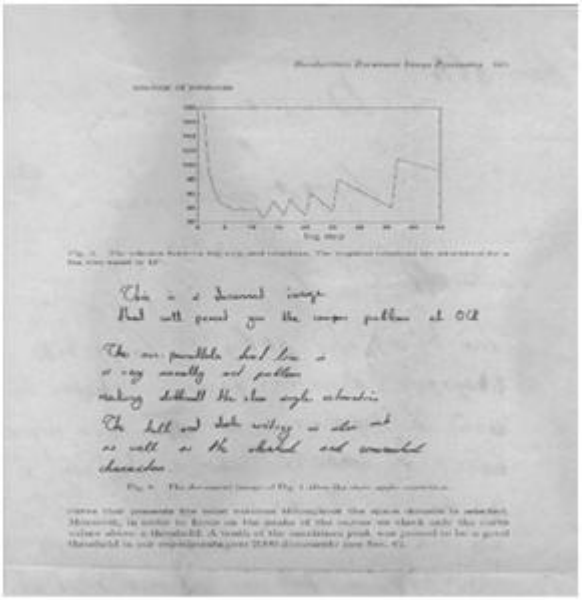

Figure 2: Degraded average intensity image.

Table 1: Global Algorithms versus Assessment Parameters

Algorithm/ Parameter	PERR	MSE	SNR	PSNR
[Otsu 1979]	1.747414	1136.256	17.82513	18.03767
[Reddi 1984]	1.641987	1067.702	18.1055	18.31057
[Trier,Taxt 1995]	3.142154	2043.185	15.25285	15.58478
[Prewitt, Mendelsohn 1966]	3.359808	2184.715	15.7486	15.91618
[Doyle 1962]	2.501591	1626.66	15.93992	16.15941
[Ramesh et al 1995]	2.026501	1317.732	17.48979	17.65106
[Rosenfeld,Kak 1982]	28.10834	18277.45	3.958347	5.613259
[Jain,Dubes 1988]	13.94842	9069.963	14.99826	15.19859
[Duda,Hart 1973]	1.758393	1143.395	17.85714	18.04058
[Pun 1980]	42.82607	27847.65	0.950004	3.697339
[Yen 1995]	3.199158	2080.253	15.2717	15.55781
[Kittler, Illingworth 1985]	22.22692	14453.05	8.047554	9.957478
[Lloyd 1985]	30.18251	19626.18	3.494714	5.314108
[Ridler,Calvard 1978]	24.56092	15970.74	6.585206	8.091815

Table 2: Local Algorithms versus Assessment Parameters

Algorithm/Parameter	PERR	MSE	SNR	PSNR
[Kohon. SOM 2000]	2.275293	1479.509	17.33808	17.57842
[Sauvola et.al.2000]	2.297247	1493.785	16.5649	16.80586
[Brink et.al.1996]	3.009194	1956.728	16.05018	16.15053
[Gatos et.al.2006]	1.664706	1082.475	18.13241	18.39107
[Vonikakis et.al.2008]	1.717771	1116.98	17.86367	18.08074

7. Conclusion

This analysis work has focused on evaluation of some thresholding algorithms for document images. The outcomes of threshold algorithms are extremely rely on the quality and degradation of document or image. It is not mandatory that an algorithm that gives better result for particular measure has to give better outcome for another measure. No particular thresholding method might be appealed as the

superlative technique for all assessment parameters or input document images. Expansion of thresholding techniques for document images is vital and needed. Alternatively one may use multi-stage thresholding or hybrid thresholding algorithm.

References

[1] Bolan Su, Shijian Lu and Chew Lim Tan, "Robust Document Image Binarization Technique for Degraded Document Images", IEEE Transactions on Image Processing, Vol. 22, No. 4, April 2013.
[2] Yung-Hsiang Chiu , Kuo-Liang Chung , Wei-Ning Yang, Yong-Huai Huang and Chi-Huang Liao, "Parameter-free based two-stage method for binarizing degraded document images", Y.-H. Chiu et al. / Pattern Recognition 45 (2012) 4250–4262, Elsevier 2012.
[3] Konstantinos Ntirogiannis, Basilis Gatos and Ioannis Pratikakis, "A Performance Evaluation Methodology for Historical Document Image Binarization", IEEE 2011.
[4] Vavilis Sokratis, Ergina Kavallieratou, Roberto Paredes and Kostas Sotiropoulos, "A Hybrid Binarization Technique for Document Images", Springer 2011.
[5] Trier and Taxt, "Improvement of 'integrated function algorithm' for binarisation of document images", Pattern Recognition Letters, 16, pp. 277–283, 1995.
[6] Ramesh, N., Yoo, J.H., Sethi, I.K., "Thresholding based on histogram approximation", IEE Proc.-Vis.Image Signal Process., Vol.142, No.5 pp: 4147, 1995.
[7] Jain, A. K., and Dubes, R. C.: "Algorithms for Clustering Data", Prentice Hall, 1988.
[8] Yen, J.C., Chang, F.J., and Chang, S., "A New Criterion for Automatic Multilevel Thresholding", IP (4), No. 3, March pp. 370-378, 1995.
[9] Kittler, J., Illingworth, J., "On threshold selection using clustering criteria", IEEE Trans. Systems Man Cybernet.15, 652–655, 1985.
[10] Papamarkos N., and Atsalakis, A.: "Gray-level reductionusing local spatial features"; Computer Vision and Image Understanding, pp. 336-350, 2000.
[11] Sauvola, J., Pietikainen, M.: "Adaptive document image binarization "; Pattern Recognition 33, 225–236, 2000.
[12] Brink, A.D., Pendock, N.E.: "Minimum Cross-Entropy Threshold Selection"; PR (29), pp. 179-188, 1996.
[13] Gatos, B., Pratikakis, I.E., Perantonis, S.J.: "Adaptive degraded document image binarization"; Pattern Recognition (39), No. 3, pp. 317-327, 2006.
[14] Vonikakis, V., Andreadis, I., and Papamarkos, N.: "Robust Document Binarization with OFF Center-surround Cells"; Pattern Analysis & Applications, to appear.
[15] Kavallieratou, E.: "A Binarization Algorithm Specialized on DocumentImages and Photos"; 8th Int. Conf. on Document Analysis and Recognition, 2005, pp.463-467.
[16] Kite, T.D., Evans, B.L., Daamera-Venkata, N., and Bovil, A.C., "Image Quality Assessment Based on a Degradation Model"; in IEEE Trans. Image Processing, vol.9, pp.909-922, 2000.
[17] Stathis, Kavallieratou, Papamarkos, "An Evaluation Technique for Binarization Algorithms", Journal of Universal Computer Science, vol. 14, no. 18 (2008), 3011-3030.