Journal of Theoretical and Applied Information Technology
31st December 2015. Vol.82. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PERFORMANCE EVALUATION OF QUANTITATIVE METRICS ON ANCIENT TEXT DOCUMENTS USING MIGT

[1]N. VENKATA RAO, [2]Dr. A.S.C.S.SASTRY, [3] Dr. A.S.N.CHAKRAVARTHY, [4]A.V.SRINIVASA RAO

[1]Research Scholar, Department of ECE , K.L.University, Vaddeswaram, Andhra Pradesh, India
[2]Professor, Department of ECE, K.L.University, Vaddeswaram, Andhra Pradesh, India
[3] Assoc. Professor, Department of CSE , University College of Engineering, Kakinadda, Andhra Pradesh, India
[4]Assistant Professor, Department of ECE , Sasi Institute of Technology and Engg, Tadepalligudem, Andhra Pradesh, India
E-mail: [1]venkatnekkanti@rediffmail.com [2] ascssatry @kluniversity.in [3]chakravarthy.cse@jntukucev.ac.in [4]ad_bala@rediffmail.com

ABSTRACT

In the present world scenario Optical Character Recognition (OCR) has wide variety of applications in the text document image analysis for recognizing individual characters of any language. Digitizing the old documents is a tough job for preserving the essence of the documents to the coming eras. In this paper we are summarizing different image quantitative metrics for estimating the loss of information from the image after cleaning the noisy image by using anyone of the local or non-local thresholding techniques. The quality evaluations are made on 40 Telugu and English text documents after cleaning the documents with Modifies Iterative Global Threshold (MIGT) approach.

Keywords: *Thresholding Techniques, Digitization, Evaluation, Optical Character Recognition, Document Image Analysis*

1. INTRODUCTION

In image processing, the gray levels of pixels belonging to the object are substantially different from the gray levels of the pixels belonging to the background. This kind of action we will observe in many applications of Image Processing. Thresholding is a technique, which is a simple and effective tool to separate objects from the background. Some of the examples of thresholding applications are document image analysis where the printed characters are extracted [1,2] extraction of the characters from the ancient documents [3], graphical content analysis, musical scores: map processing, scene processing, etc.

The result of the thresholding operation is a binary image whose one state will indicate the foreground objects, that is, printed text, a legend, a target, defective part of a material, etc., while the complementary state will correspond to the background. Based on the application, the foreground can be represented by gray-level 0, that is, black as for text, and the background by the highest luminance for document paper, that is 255 in 8-bit images, or conversely the foreground by white and the background by black. Various factors, influence the smooth operation of thresholding like correlated noise, high density of pixels in a given area, overlapping of pixel information, blurriness of an image all these will complicate the thresholding operation. Finally, the lack of objective measures to assess the performance of various thresholding algorithms, and the difficulty of extensive testing in a task oriented environment, are the other major handicaps. Mehmet Sezgin and Bulent Sankur [5] performed a survey on image thresholding techniques and quantitative performance evaluation on different images. Yang Gao-bo et.al. proposed [6] an objective performance evaluation on video segmentation algorithms with ground truth images. K.Ntirogiannis, et. al. proposed [7] an evaluator methodology for Document Image Binarization techniques using ground truth images. Chun Che Fung and Rapeeporn Chamchony proposed [8] a review of evaluation on Optical Binarization techniques for character segmentation

Journal of Theoretical and Applied Information Technology
31st December 2015. Vol.82. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

in historical manuscripts. Štepán Šrubar stated in poster presentation [9] that the Segmentation evaluation methods have variable quality. Some of them are focused on a small part of information from the whole segmentation and the quality is therefore poor. Even some recently proposed methods do not assure high quality. The best results were provided by method which uses grouping of segments and distance measuring. This quality measurement is one of the biggest according to the size of test set as well as the number of methods which can ensure high level of objectivity. Ruchika Sharma et.al. present [10] comparision of binarization techniques using several observations, techniques and methods. These approaches aiming at removal of background noise from camera-based images. V. Rabeux et.al. presented [11] 18 features that characterize the quality of a document image. These features are used in step-wise multivariate linear regression to create prediction models for 12 binarization methods. Repeated random subsampling cross-validation shows that the models are accurate (max percentage error equals 11%). Moreover, given the step-wise approach of the linear regression, these models are not over fit. As a result, 10 models out of 12 are validated and show sufficient accuracy to be used in an automated selection method of the optimal binarization method for each image.

In this paper we described different quantitative techniques which are Misclassification error, Region non-uniformity, Relative foreground area error, Hausdorff distance and Jaccard distance for estimating the loss of information during the cleaning process of ancient Telugu and English text samples. Before estimating the quantitative metrics, the documents are cleaned by Iterative Global Threshold algorithm.

This paper is organized in to four sections. In the first section introduction, literature survey and problem definition are discussed. In the second section the quantitative metrics for estimating the performance ofMIGT algorithm for the cleaning of ancient documents is discussed. In the third section, experimental results are discussed. The last section gives the conclusions and future scope of work.

2. METHODOLOGY

Cleaning of Noisy Documents

Binarization is one of the several steps used in most document image analysis systems. It consists of labeling each pixel in an image as foreground and background. It provides a proper distinction between background and foreground. In this paper we used Modified Iterative Global Threshold algorithm for cleaning the documents which is proposed [3] by us.

Thresholding performance criteria

Automated image thresholding encounters difficulties when the foreground object constitutes a disproportionately small large area of the scene, or when the object and background gray levels possess substantially overlapping distributions, even resulting in a unimodal distribution. Furthermore, the histogram can be noisy if its estimate is based on only a small sample size, or it may have a comb-like structure due to histogram stretching/equalization efforts. Consequently, misclassified pixels and shape deformations of the object may adversely affect the quality-testing task in Non-Destructive Testing(NDT) applications. On the other hand, thresholded document images may end up with noise pixels both in the background and foreground, spoiling the original character bitmaps. Thresholding may also cause character deformations such as chipping away of character strokes or conversely adding bumps and merging of characters among themselves and/or with background objects. Spurious pixels as well as shape deformations are known to critically affect the character recognition rate. Therefore, the criteria to assess thresholding algorithms must take into consideration both the noisiness of the segmentation map as well as the shape deformation of the characters, especially in the document processing applications.

To put into evidence the differing performance features of thresholding methods, we have used the following five performance criteria: misclassification error (ME), edge mismatch (EMM), relative foreground area error (RAE), Modified Hausdorff Distance (MHD), and region nonuniformity (NU). Obviously, these five criteria are not all independent: for example, there is a certain amount of correlation between misclassification error and relative foreground area error, and similarly, between edge mismatch and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Hausdorff distance, both of which penalize shape deformation. The region nonuniformity criterion is not based on ground truth data, but judges the intrinsic quality of the segmented areas. We have adjusted these performance measures so that their scores vary from 0 for a totally correct segmentation to 1 for a totally erroneous case.

Misclassification error

Misclassification error (ME) reflects the percentage of background pixels wrongly assigned to foreground, and conversely, foreground pixels wrongly assigned to background. For the two-class segmentation problem, ME can be simply expressed as:

$$ME = 1 - \frac{|B_0 \cap B_T| + |F_0 \cap F_T|}{|B_0| + |F_0|}$$

where B_0 and F_0 denote the background and foreground of the original (ground-truth) image, BT and FT denote the background and foreground area pixels in the test image, and |.| is the cardinality of the set. The ME varies from 0 for a perfectly classified image to 1 for a totally wrongly binarized image.

Region nonuniformity

This measure, does not require ground-truth information, is depends on variance of the whole image is defined as

$$NU = \frac{|F_T|}{|F_T + B_T|} \frac{\sigma_f^2}{\sigma^2}$$

where σ^2 represents the variance of the whole image, and σ_f^2 represents the foreground variance. It is expected that a well-segmented image will have a nonuniformity measure close to 0, while the worst case of NU=1 corresponds to an image for which background and foreground are indistinguishable up to second order moments.

Relative foreground area error

The comparison of object properties such as area and shape, as obtained from the segmented image with respect to the reference image, has been used in Zhang[13] under the name of relative ultimate measurement accuracy (RUMA) to reflect the feature measurement accuracy. We modified this measure for the area feature A as follows:

$$RAE = \frac{A_0 - A_T}{A_0} \quad \text{if } A_T < A_0$$

$$RAE = \frac{A_T - A_0}{A_T} \quad \text{if } A_T \leq A_0$$

where A_0 is the area of reference image, and A_T is the area of thresholded image. Obviously, for a perfect match of the segmented regions, RAE is zero, while if there is zero overlap of the object areas, the penalty is the maximum one.

Hausdorff distance

The Hausdorff distance can be used to assess the shape similarity of the thresholded regions to the ground-truth shapes. Recall that, given two finite sets of points, say ground-truth and thresholded foreground regions, their Hausdorff distance is defined as

H(FO,FT) = max{dH(FO,FT), dH(FT,FO)}

Where dH(FO,FT) = max d(fO, FT)

fO ε FO = max min||fO-fT||, fO ε FOfT ε FT

and ||fO – fT|| denotes the Euclidean distance of two pixels in the ground-truth and thresholded objects. Since the maximum distance is sensitive to outliers, we have measured the shape distortion via the average of the modified Hausdorff distances (MHD)4 over all objects. The modified distance is defined as:

$$MHD(F_0, F_T) = \frac{1}{|F_0|} \sum_{f_0 \varepsilon F_0} d(f_0, F_T)$$

For example, the MHDs are calculated for each 19 x19 pixel character box, and then the MHDs are averaged over all characters in a document. Notice that, since an upper bound for the Hausdorff distance cannot be established, the MHD metric could not be normalized to the interval [0, 1], and hence is treated separately (by dividing each MHD value to the highest MHD value over the test image set NMHD).

Jaccard distance

The Jaccard distance, which measures dissimilarity between sample sets, is complementary to the Jaccard coefficient and is obtained by subtracting the Jaccard coefficient from 1, or, equivalently, by dividing the difference of the sizes of the union and the intersection of two sets by the size of the union:

Journal of Theoretical and Applied Information Technology
31st December 2015. Vol.82. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$J_\delta(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

Where A and B are two sets

3. RESULTS AND DISCUSSIONS

Quantity of the image is very important when cleaning the document with any local or non-local thresholding algorithm because during the process some of the useful information is lost along with noise. So the given quantitative metrics Misclassification Error, Region Nonuniformity, Relative foreground area error, Hausdorff distance, Jaccard distance will provide the amount of information which we lost from the cleaned document by using any local/non-local technique is comparing with original document which is cleaned manually (ground truth image).The above quality metrics are tested on 40 ancient Telugu and English text samples after cleaning these documents with MIGT algorithm. Initially the text samples are cleaned by MIGT algorithm; during the cleaning process some of the useful information along with background noise is eliminated. So the lost information from the image is calculated by using the above said metrics. In this connection the cleaned document is compared by ground truth image which is obtained from the binary image and removing the unnecessary information manually. Original text samples are described in Fig.1 and 4. Resultant images of MIGT algorithm for Telugu and English are given in Figs. 2&5 and their subsequent ground truth images are shown in Figs. 3&6.

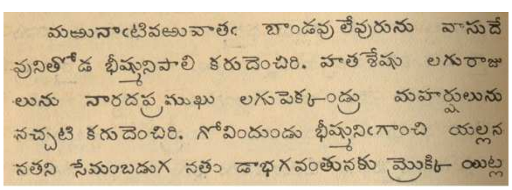

Figure 1 Telugu Text Sample

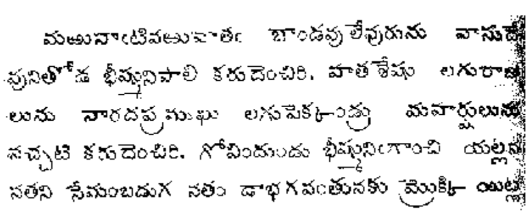

Figure 2 Resultant Image Of MIGT Algorithm For Telugu Text Sample

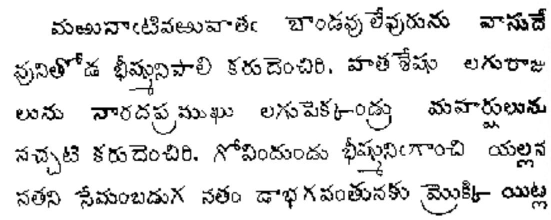

Figure 3 Telugu Ground-Truth Binary Text Image

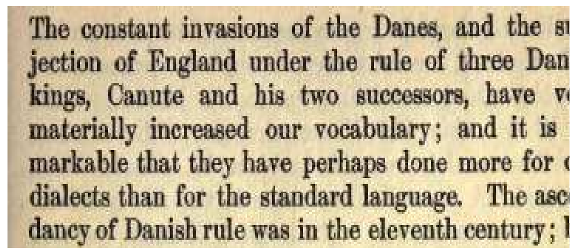
The constant invasions of the Danes, and the su
jection of England under the rule of three Dan
kings, Canute and his two successors, have v
materially increased our vocabulary; and it is
markable that they have perhaps done more for
dialects than for the standard language. The asc
dancy of Danish rule was in the eleventh century;

Figure 4 English Text Sample

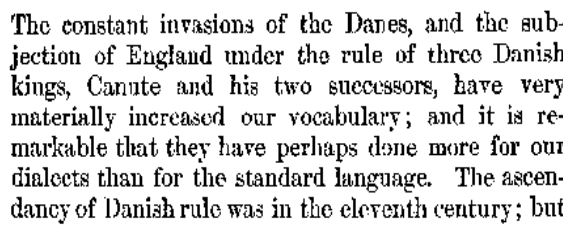
The constant invasions of the Danes, and the sub-
jection of England under the rule of three Danish
kings, Canute and his two successors, have very
materially increased our vocabulary; and it is re-
markable that they have perhaps done more for our
dialects than for the standard language. The ascen-
dancy of Danish rule was in the eleventh century; but

Figure 5 Resultant Image Of MIGT Algorithm For English Text Sample

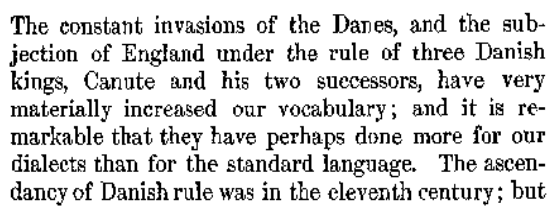
The constant invasions of the Danes, and the sub-
jection of England under the rule of three Danish
kings, Canute and his two successors, have very
materially increased our vocabulary; and it is re-
markable that they have perhaps done more for our
dialects than for the standard language. The ascen-
dancy of Danish rule was in the eleventh century; but

Figure-6 English Ground Truth Binary Text Image

In this regard the tested values of Telugu and English text samples are tabulated in table 1 and table 2. When we observe both the tables there is a correlation between the values of different metrics for both Telugu and English text samples. Especially there is a similarity between the values of Jaccard distance and region nonuniformity for both the samples.

Journal of Theoretical and Applied Information Technology
31st December 2015. Vol.82. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1. Quantitative Metrics For Telugu Text Samples

Misclassification Error (Telugu)	Region non-uniformity	Relative foreground area Error	Hausdorff distance	Jaccard distance
0.00175	0.0135	0.006832	2	0.5734
9	37	4	2.2336	7
0.00323	0.0155	0.017913	2	0.5661
1	02	0.009142	1.7321	9
0.00262	0.0154	5	2	0.5667
44	71	0.013519	2.4495	1
0.00275	0.0096	0.035113	2	0.5711
06	425	0.018371	1.7321	0.5647
0.00562	0.0142	0.002043	2	1
93	86	7	2	0.5564
0.00327	0.0135	0.005692	2.6458	6
76	99	0.003390	2.2361	0.5599
0.00149	0.0186	1	2	3
2	81	0.009611	2.2361	0.5640
0.00104	0.0142	3	1.4142	6
99	54	0.031615	1.414	0.5665
0.00895	0.0124	0.027764	2.2361	0.5585
2	17	0.000579	2	7
0.00267	0.1102	5	2.4495	0.5471
25	9	0.000077	1.4142	9
0.00731	0.0093	25		0.5496
81	227	0.013236		6
0.00319	0.0106	0.014397		0.5505
24	54	0.007035		1
0.00322	0.0120	9		0.5531
0.00046	22	0.003135		4
798	0.0123	0.001677		0.5505
0.00155	74	0.014399		6
4	0.0110			0.5504
0.00096	54			2
388	0.0131			0.5537
0.00127	32			0.5502
92	0.0104			2
0.00079	24			0.5430
35	0.0104			2
0.00072	12			0.5504
267	0.0103			2
0.00153	04			
4	0.0121			
	87			

Table 2. Quantitative Metrics for English Text Samples

Misclassification Error (English)	Region Nonuniformity	Relative Foreground area Error	Hausdorff Distance	Jaccard Distance
0.0042	0.01372	0.029908	1.4142	0.56611
624	2	0.03224	1.4142	0.57041
0.0050	0.01388	0.021919	1	0.55806
167	1	0.025318	1.4142	0.56608
0.0029	0.01179	0.024478	2	0.56895
038	7	0.024278	1.4142	0.56912
0.0037	0.01260	0.037529	2	0.57473
583	2	0.033599	1.7321	0.56893
0.0042	0.01368	0.021414	1.7321	0.56694
186	1	0.02817	2	0.56466
0.0038	0.01386	0.018284	1.4142	0.55112
661	5	0.016093	1.7321	0.57094
0.0064	0.01260	0.003572	2.2361	0.58276
14	5	6	1.7321	0.57094
0.0051	0.01298	0.01543	1.7321	0.5736
177	6	0.022519	3	0.57125
0.0036	0.01319	0.019717	1.7321	0.55373
179	4	0.011262	1.4142	0.56624
0.0042	0.01272	0.022501	1.4142	0.56669
8	9	0.032602	1.7321	0.57176
0.0022	0.01616	0.019951		
734	6			
0.0029	0.01292			
81	6			
0.0012	0.01547			
871	8			
0.0029	0.01292			
033	6			
0.0037	0.01433			
048	6			
0.0045	0.01442			
567	5			
0.0016	0.01148			
497	9			
0.0034	0.01252			
28	9			
0.0049	0.01147			
993	2			
0.0035	0.01445			
84	3			

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. CONCLUSION

We identified some similarity in the values of metrics which are tested on 40 Telugu and English text samples which can be visualized through the table. Based on the available information from the tables 1&2 the loss of information is same in all samples. Here we have not made any comparison with other threshold methods but based on the metric values we came to a conclusion IGT method is better suitable for ancient documents.

5. FUTURE SCOPE OF WORK

There is a possibility to extend the above procedure for noise free samples which are contaminated manually by different noises at different levels like pepper, Gaussian noise and so on. Then clean the noisy documents with IGT as well as other methods and then find the quantitative metrics for identifying the loss of information during the cleaning process.

REFERENCES

[1] M. Kamel and A. Zhao, ‘‘Extraction of binary character/graphics images from grayscale document images,’’ Graph. Models Image Process, vol.55,no3,pp 203–217, 1993.

[2] Abak, U. Baris¸, and B. Sankur, ‘‘The performance of thresholding algorithms for optical character recognition,’’ Intl. Conf. Document Anal. Recog. ICDAR’97, pp. 697–700, 1997.

[3] N Venkata Rao, A V Srinivasa Rao, S Balaji, L Pratap Reddy,”Cleaning of Ancient Document Images using Modified Iterate Global Threshold, ” IJCSI, Vol.8, Issue 6, November 2011

[4] M. Sezgin and B. Sankur, ‘‘Comparison of thresholding methods for non-destructive testing applications,’’ IEEE ICIP’2001, Intl. Conf. Image Process., pp. 764–767, 2001

[5] M. Sezgin, and B. Sankur, “Survey over Image Thresholding Techniques and Quantitative Performance Evaluation”, Journal of Electronic Imaging, January 2004, vol. 13, no. 1, pp. 146-168.

[6] YANG Gao-bo, ZHANG Zhao-yang “Objective Performance Evaluation of Video Segmentation Algorithms with Ground-Truth”. Journal of Shanghai University (English Edition), 2004, 8(1) : 70- 74

[7] K. Ntirogiannis, B. Gatos and I. Pratikakis. “An Objective Evaluation Methodology for Document Image Binarization Techniques”. The Eighth IAPR Workshop on Document Analysis Systems, 2008, 217-224

[8] Chun Che Fung and Rapeeporn Chamchong. “A Review of Evaluation of Optimal Binarization Technique for Character Segmentation in Historical Manuscripts” Third International Conference on Knowledge Discovery and Data Mining. 2010, 236-240

[9] Štepán Šruba “Comparison of segmentation evaluation methods” Conference on Computer Graphics, Visualization and Computer Vision, Poster proceedings,2013.

[10] Ruchika Sharma, Balwinder Singh “A Review on Evaluation of Binarization Techniqueson Camera-Based Images” International Journal of Computer Science And Technology,vol 4,issue2, 2013,pp 606-609.

[11] Vincent Rabeux, “Quality evaluation of degraded document images for binarization result prediction” International Journal ond Dcument Analysis and Recognition, pp 1-13, 2013.

[12] A V Srinivasa Rao “ Segmentation of Ancient Telugu text documents” International Journal of Image Graphics and Signal Processing 6, 2012, 8-14

[13] Y.J.Zhang, “A survey on evaluation methods for image segmentation”, Pattern Recognition, Vol: 29, no.8, pp:1335-1346, 1996.