Comparison of Binarization Algorithm in Indian Language OCR

Tushar Patnaik, Shalu Gupta, Deepak Arya

Abstract- Image binarization is an important step for document image analysis and recognition pipeline. It converts an image (up to 256 gray levels) to black and white images (0 or 1). A threshold value needs to be defined for this. Various algorithms have been proposed for image binarization. In this paper, the comparison of mentioned Binarization algorithms with respect to OCR performance is presented. A method to select a optimal Binarization algorithm is proposed.

Keywords: Binarization, Ground Truth Data, SNR, OCR

1. Introduction

Document Binarization has been an active research area for the past many years. The choice of most optimum binarization algorithm has proved to be difficult, as different binarization algorithm gives different performance on different data sets. This is especially true in the case of historical documents with variation in contrast and illumination, smearing and smudging of text. The quality of the image has a significant impact on the OCR performance. Noise present in images after binarization would reduce the performance of subsequent processing steps and in many cases could even cause their failure. So, selection of appropriate binarization algorithm is very important for OCR performance.

In this paper, various binarization algorithms are discussed and compared with respect to OCR performance. The algorithms considered are Otsu, Adaptive and Sauvola. To compare these, a method having two phases has been proposed.

This paper is organized in following sections: Section 2 describes the different binarization methods. Section 3 describes the comparison of binarization method. Section 4 presents the analysis of results. Section 5 gives the conclusion of results.

2. Binarization Methods

Image binarization converts an image (up to 256 gray levels) to a black and white image (0 or 1). The simplest way to use image binarization is to choose a threshold value, and classify all pixels with values above this threshold as white, and all other pixels as black. The binarization algorithms considered are Otsu, Adaptive and Sauvola.

2.1 Otsu Global Algorithm

This method is both simple and effective. The algorithm assumes that the image to be threshold contains two classes of pixels (e.g. foreground and background) and calculates the optimum threshold separating those two classes so that their combined spread (intra-class variation) is minimal. In Otsu's method we exhaustively search for the threshold that minimizes the intra-class variance, defined as a weighted sum of variances of the two classes:

$$\sigma_w^2(t) = \omega_1(t)\sigma_1^2(t) + \omega_2(t)\sigma_2^2(t)$$

Weights ω_i are the probabilities of the two classes separated by a threshold *t* and σ_i^2 variances of these classes.
Otsu shows that minimizing the intra-class variance is the same as maximizing inter-class variance

$$\sigma_b^2(t) = \sigma^2 - \sigma_w^2(t) = \omega_1(t)\omega_2(t)\left[\mu_1(t) - \mu_2(t)\right]^2$$

which is expressed in terms of class probabilities ω_i and class means μ_i which in turn can be updated iteratively.

2.2 Adaptive Binarization

Adaptive binarization method extends Otsu's method to a novel adaptive binarization scheme. The first step of our method is to divide images into NxN blocks, and then Otsu's method is applied straightaway in each of the blocks. Then each and every pixel is applied with a nonlinear quadratic filter to fine tune all the pixels according to the local information available. Adaptive Binarization technique combines global thresholding with quadratic filter using local information to fine tune the pixel.

2.3 Sauvola binarization

Sauvola binarization technique aims to convert a gray tone document image into two tone image. For bad quality image global thresholding cannot work well. For this, we like to apply a technique, which is window-based local one. Sauvola binarization technique is window-based, which calculates a local threshold for each image pixel at (x, y) by using the intensity of pixels within a small window W (x, y). Here we have taken the window of size 19x19 pixels with (x, y) as centre except at the edge pixels of the image frame. So, we start computation from x =10, y =10. The threshold T (x, y) is computed using the following formula-

T (x, y) = Int [X. (1 + k. (/R - 1))]

Where X is the mean of gray values in the considered window W (x, y), is the standard deviation of the gray levels and R is the dynamic range of the variance, k is a constant (usually 0.5 but may be in the range 0 to 1).

3. Comparison of the Binarization Methods

This section proposes a two phase method for comparison of Binarization Methods. In the first phase, SNR calculation is done on every image. In the second phase, number of OCR errors is calculated. The optimal binarization algorithm is one, which has the highest SNR and least number of errors.

3.1 Calculate SNR

The ideal way of evaluating binarization algorithm is to check number of pixels in output image, which have changed their values (i.e. black pixel in original image changed to white in output and vice versa). The evaluation of the binarization methods is made on noisy images. Starting from a clean document image, this is considered as the ground truth image, noise of different types is added (noisy images). After adding noises, binarization is applied to the noisy images. During the evaluation, every single pixel value of binarization output is compared with the corresponding pixel in the original image for which we use statistical measures of image quality description called MSE [7] and SNR [7].

Let x (i, j) represent the value of the i-th row and j-th column pixel in the original document and let y (i, j) represent the value of the corresponding pixel in the output image y:MxN. Since we deal with black and white images, both values will be either 0 (black) or 255 (white). The local error is e (i, j)=x (i, j)-y (i, j) and the total square error rate:

$$MSE = \sum_i \sum_j e(i, j)\, 2/M*N$$

SNR [7] is defined as the ratio of average signal power to average noise power. SNR is calculated by the following formula, for a MxN image is

$$SNR\ (DB) = 10 \log 10 \sum_i \sum_j x(i, j) / \sum_i \sum_j (x(i, j) - y(i, j))\, 2$$

Where x (i, j) represent the value of the i-th row and j-th column pixel in the original image (Ground Truth Image) and y (i, j) represent the value of the corresponding pixel in the output image (Binarized Image). The local error is e (i, j) = x (i, j)-y (i, j). Higher the SNR better is the efficiency of binarization algorithm. We applied all the binarization methods described in above section to 100 sets of images. The pixels that changed value (white-to-black or vice versa) were counted by comparing the output image with the original document image and then SNR is calculated.

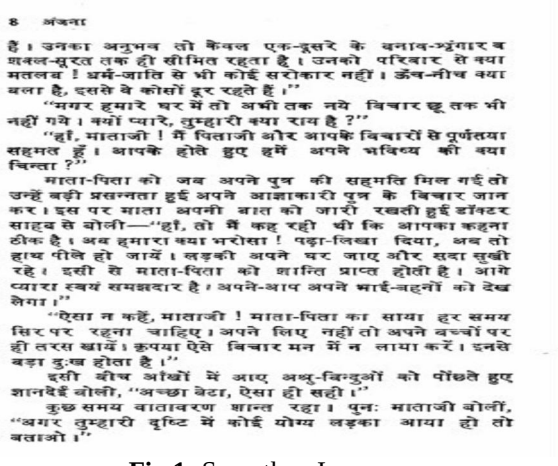
3.2 OCR testing with binarized image

Through SNR only, optimality of a binarization algorithm cannot be predicted, as OCR output also depends on other preprocessing routines like skew correction and recognition engine etc. To accurately predict the accuracy of binarization algorithm, OCR output is taken of all the binarized images. The OCR output is compared with ground truth data and numbers of errors are calculated. The total error calculation is based on the sum of Insertion, Deletion and Substitution using Levenshtein algorithm.

4. Analysis and Results

4.1 Choose smoothen images

The smoothen images are those in which there is no skew, noise and two tone image (0 or 1). We have taken hundred smoothen images to test the binarization algorithm. All of the images are taken from OCR Project corpus. One of the smoothen image is shown below: -



8 अंजना

हैं। उनका अनुभव तो केवल एक-दूसरे के बनाव-शृंगार व शक्ल-सूरत तक ही सीमित रहता है। उनको परिवार से क्या मतलब ! धर्म-जाति से भी कोई सरोकार नहीं। ऊँच-नीच क्या बला है, इससे वे कोसों दूर रहते हैं।"

"मगर हमारे घर में तो अभी तक नये विचार छू तक भी नहीं गये। क्यों प्यारे, तुम्हारी क्या राय है ?"

"हाँ, माताजी ! मैं पिताजी और आपके विचारों से पूर्णतया सहमत हूँ। आपके होते हुए हमें अपने भविष्य की क्या चिन्ता ?"

माता-पिता को जब अपने पुत्र की सहमति मिल गई तो उन्हें बड़ी प्रसन्नता हुई अपने आज्ञाकारी पुत्र के विचार जान कर। इस पर माता अपनी बात को जारी रखती हुई डॉक्टर साहब से बोली—"हाँ, तो मैं कह रही थी कि आपका कहना ठीक है। अब हमारा क्या भरोसा ! पढ़ा-लिखा दिया, अब तो हाथ पीले हो जायें। लड़की अपने घर जाए और सदा सुखी रहे। इसी से माता-पिता को शान्ति प्राप्त होती है। आगे प्यारा स्वयं समझदार है। अपने-आप अपने भाई-बहनों को देख लेगा।"

"ऐसा न कहें, माताजी ! माता-पिता का साया हर समय सिर पर रहना चाहिए। अपने लिए नहीं तो अपने बच्चों पर ही तरस खायें। कृपया ऐसे विचार मन में न लाया करें। इनसे बड़ा दुःख होता है।"

इसी बीच आँखों में आए अश्रु-बिन्दुओं को पोंछते हुए शानदेई बोली, "अच्छा बेटा, ऐसा ही सही।"

कुछ समय वातावरण शान्त रहा। पुनः माताजी बोलीं, "अगर तुम्हारी दृष्टि में कोई योग्य लड़का आया हो तो बताओ।"

Fig.1: Smoothen Image

4.2 Add noise to images

We have added different type of noises to the smoothen images. Following types of noises are added in the smoothen images:

4.2.1 Gaussian
4.2.2 Poisson
4.2.3 Speckle
4.2.4 Localvar

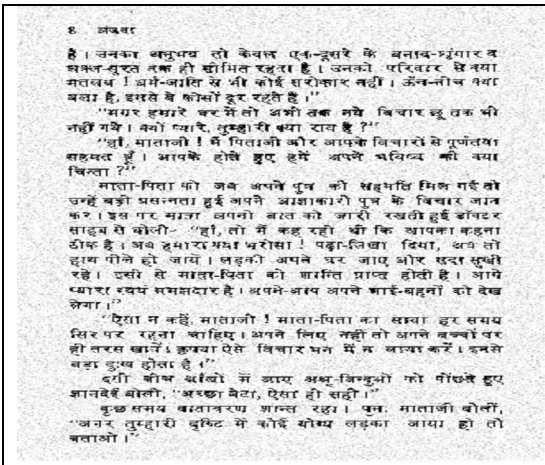


Fig.2: Gaussion Noisy Image

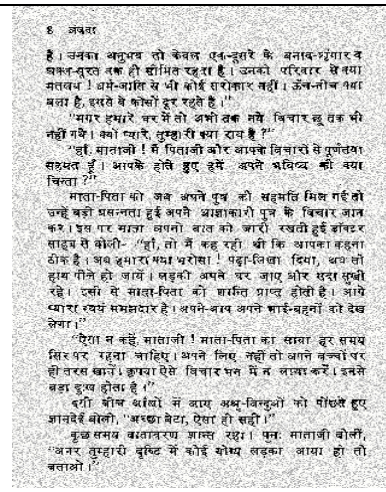


Fig.4: Speckle Noisy Image

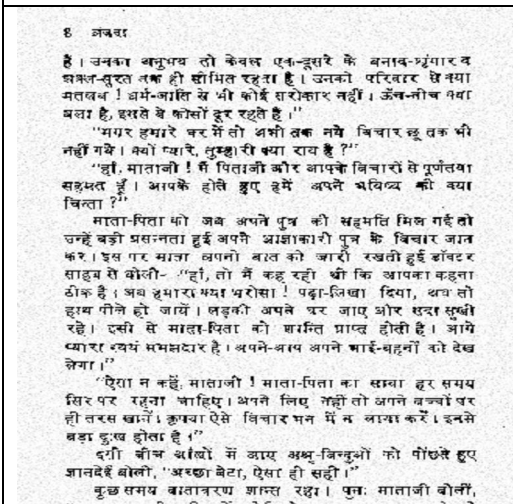


Fig.3: Poisson Noisy Image

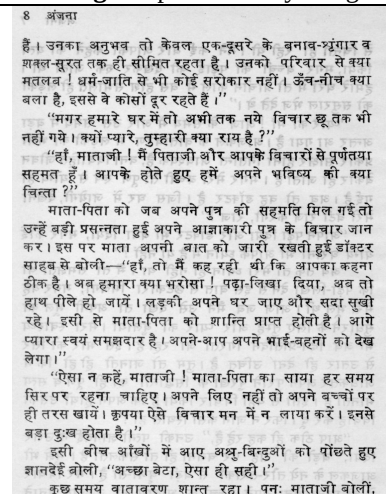
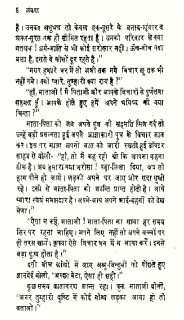


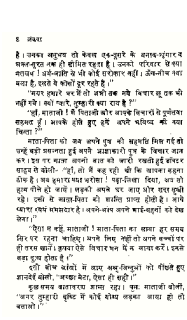
Fig.5: Localvar Noisy Image

4.3 Output of Binarization Algorithms

Gaussian Noise:-



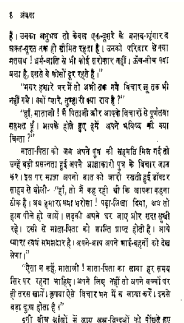
Otsu



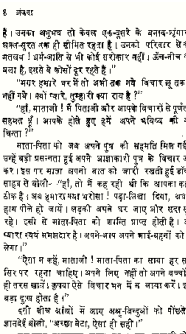
Adaptive

Sauvola

Poisson Noise:-



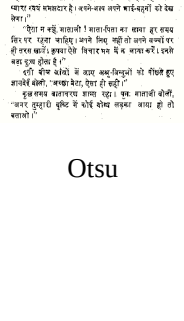
Otsu



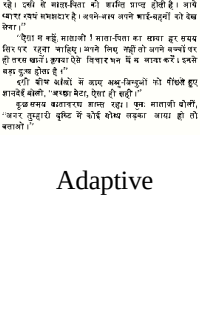
Adaptive

Sauvola

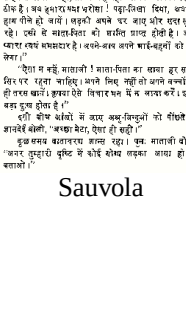
Speckle Noise:-



Otsu



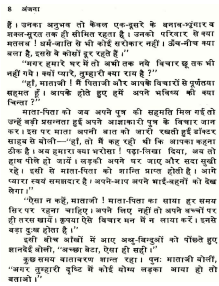
Adaptive



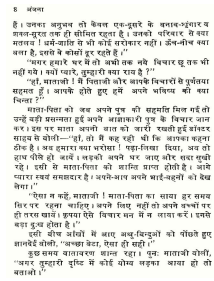
Sauvola

Localvar Noise:-

Otsu



Adaptive



Sauvola

4.4 SNR comparison with different binarization algorithm

Table1 gives the SNR comparison with different binarization algorithms

SNR				
	Gaussion	Poisson	Speckle	Localvar
Otsu	17.41	14.72	11.56	16.5
Adaptive	17.66	15.34	12.25	17.48
Sauvola	15.71	14.98	13.22	16.03

Table1: SNR

By looking carefully at the table where the algorithm performance in terms of SNR is given for the image sets, some remarks are made:

1) If we accept that the mean SNR gives a good estimation of the final image, the variation of the SNR gives a good indication of the algorithm stability. Thus, Sauvola results are looking more stable as compared to other algorithms.
2) In case of speckle noise, mostly algorithms are giving less SNR as compared to other noises. This is because speckle noise is random, deterministic, interference pattern in an image. So a filter has to be applied to remove this type of noise from the Image.

For all types of noises, except speckle noise, Adaptive binarization works better because it is giving the highest SNR for these types of images as compared to other binarization algorithms. After Adaptive binarization, Otsu is working better and Sauvola is working least of all the algorithms. Table 2 gives the best algorithms for each noisy image on our experimental sets

Algorithm	**Noise**
Adaptive	**Gaussian**
Adaptive	**Poisson**
Adaptive	**localvar**
Sauvola	**Speckle**

Table2: Optimal Algorithm

4.5 Compare the OCR Results with respect to ground truth data

The OCR output is compared with ground truth data and numbers of errors are calculated.
Otsu:- Table 3 gives Number of OCR errors produced using otsu algorithm.

Number of Errors(Otsu)

Otsu				
	Gaussian	poisson	speckle	localvar
Added Characters	2114	3243	4512	1983
Substituted Characters	4229	5728	6880	4468
Deleted Characters	1687	1983	2737	1934
Total Errors	8030	10954	14129	8385

Mean Error= (8030+10954+14129+8385)/4 = 10374
Adaptive:- Table 4 gives Number of OCR errors produced using adaptive algorithm.

Number of Errors(Adaptive)

Adaptive				
	Gaussian	poisson	speckle	localvar
Added Characters	1922	3123	4328	1921
Substituted Characters	4264	5537	6724	4332
Deleted Characters	1754	1975	2643	1879
Total Errors	7940	10635	13695	8132

Mean Error= (7940+10635+13695+8132)/4 = 10100

Sauvola:- Table 5 gives Number of OCR errors produced using sauvola algorithm.

Number of Errors(Sauvola)

Sauvola				
	Gaussian	poisson	speckle	localvar
Added Characters	2462	3197	4217	1981
Substituted Characters	4389	5703	6533	4532
Deleted Characters	1895	1874	2638	2004
Total Errors	8746	10774	13388	8517

Mean Error= (8746+10774+13388+8517)/4= 10356

4.6 Find the Optimal Binarization Algorithms based on SNR and OCR Error

OCR testing of binarized image and SNR calculation of binarized image following remarks can be mentioned:

1) As inferred from the table, Adaptive is the optimal algorithm of all the binarization algorithms because in case of Adaptive SNR is maximum and mean errors are least So the results which are obtained by calculating no of errors are inline with SNR.

2) Adaptive is working better on Gaussian, Poisson and Localvar noisy document images because errors are less as compared to Otsu and Sauvola.
3) Sauvola algorithm is working best on speckle noise document images as inferred from Tables 1 and 5.

7. Conclusion

A technique is proposed for the evaluation of binarization algorithms. This technique is appropriate for document images that are difficult to be evaluated by techniques based on only segmentation or recognition of the text. In order to demonstrate the proposed method we have tested three existing binarization algorithms. We performed experiments on 100 document images. Although there is a better performance of the Adaptive binarization algorithms compared to other, the other ones have produced almost similar results.

References

|1| Jiang Duan, Mengyang Zhang, Qing Li, "A Multi-stage Adaptive Binarization Scheme for Document Images, " International Joint Conference on Computational Sciences and Optimization, vol. 1, pp.867-869, 2009.
|2| Liju Dong, Ge Yu, "An Optimization-Based Approach to Image Binarization,", Fourth International Conference on Computer and Information Technology (CIT'04), pp.165-170, 2004
|3| J. He, Q. D. M. Do, A. C. Downton, J. H. Kim, "A Comparison of Binarization Methods for Historical Archive Documents," Eighth International Conference on Document Analysis and Recognition (ICDAR'05), pp.538542, 2005
|4| Ergina Kavallieratou, Stamatatos Stathis, "Adaptive Binarization of Historical Document Images," 18th International Conference on Pattern Recognition (ICPR'06) vol. 3, pp.742-745, 2006
|5| Carlos A. B. Mello, Adriano L.I.Oliveira, Ángel Sánchez: Historical Document Image Binarization.VISAPP (1) 2008: 108-113
|6| B. Gatos, I. Pratikakis, S.J. Perantonis, "Efficient Binarization of Historical and Degraded Document Images," The Eighth IAPR International Workshop on Document Analysis Systems, pp.447454, 2008
|7| Pavlos Stathis, Ergina Kavallieratou and Nikos Papamarkos " An Evaluation Survey of Binarization Algorithms on Historical Documents " 19th International Conference on Pattern Recognition, pp. 1-4, Dec.2008

About Authors

Mr. Tushar Patnaik (Sr. Lecturer/Sr. Project Engineer) joined CDAC in 1998. He has ten years of teaching experience. His interest areas are Computer Graphics, Multimedia and Database Management System. At present he is leading the consortium based project "Development of Robust Document Analysis and Recognition System for Printed Indian Scripts".

Ms. Shalu Gupta (Scientist-'C') joined C-DAC in 2008. She has seven years of experience in software development. She has worked in the field of NMS, SNMP, Optical comm., DSLAM and OCR. She has worked in various companies like C-DoT, Wipro Technology and Flextronics Software Systems. Currently she is associated with the project "Development of Robust Document Analysis and Recognition System for Printed Indian Scripts".

Mr. Deepak Kumar Arya passed B. Tech. (CSE) From Govind Ballabh Pant University Of Agriculture and Technology in 2006. He is involved in project Development of robust document analysis and Recognition system for printed Indian scripts since last two year. He is working in CDAC Noida as Contract Engineer (II).