### 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 This is especially true in the case of historical documents with variation in contrast and illumination, smearing and smudging of text.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:

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

Tushar Patnaik, Shalu Gupta, Deepak Arya

Weights  $\omega_i$  are the probabilities of the two classes separated by a threshold t and  $\sigma_i^2$  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  $\omega_i$  and class means  $\mu_i$  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:

$$MSE = \sum_{i j} e(i, j) 2/M*N$$

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) = 
$$10 \log 10 \sum_{i = j} x(i, j) / \sum_{i = j} (x(i, j) - y(i, j)) 2$$

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: -

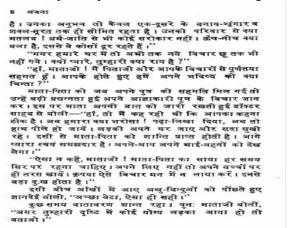


Fig.1: Smoothen Image

#### 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

है। जनका बलुपय सो केवल एक-दूसरे के जनक-पूर्यार व भक्त-पूरत तक ही श्रीमित स्तुरा है। जनको परिवार से तथा मतलब ! अमे-जाति से भी कोई शरोकार नहीं। ऊँन-नीच तथा जड़ा है, इससे वे कोसो दूर रहते हैं। "मगर हगरे कर में तो जी तिवार कुछ की नहीं गये। वशो प्यारे, पूर्वश्री क्या राय हैं!" "श्री, माताली! मैं पिताली और जापने जिला से सूर्णत्या सहस्त हैं। आपके होते हुए हमें अपने स्तित्य की तथा करा। ''

किता ??!

माता-पिरा भी जब अपने पुन की संहमति मिल गई तो

माता-पिरा भी जब अपने आशाकाशी पुन के विचार जान

कर। इस गर माता अगनी बात को जारी रखती हुई बॉक्टर
साहुत से औली- 'ही, तो में कह रही खी कि आपना कहा दीक है। अब हुआ रस माता अरसी दूर ही खी कि आपना कहा दीक है। अब हुआ रस माता अरसी ! पहा-दिनखा दिया, अर्थ सी

हाय पीने हो जाये। सहनी अपने यर जाए और सदस सुची

स्वीर स्वयं समसदार है। अपने-अर्थ अपने कार्र-बहनों दो देख

नेवा।

प्काराध्यक्ष समावाकाः है लेगा। पूराम कहीं, मालाली ! माला-पिता का सावा हर समय सिरपर रहता चाहिए। अपने लिए नहीं तो आने तब्बों पर ही तरस खार्ने। हुप्या ऐसे विवार घने में त लाजा करें। इनसे

हा नदस बारा : हुण्या (से विचार घने में स बाजा करें। इनसे नहा दुख दोता है।" दर्गी कीच कार्यों म जाय अध्या कार्युओं को गोंधती हुए प्रावदेश कीजों, 'अष्टक बेटा, देसा हो सही।" पुछ समय कारावरण जान्य रहा। पुन, माताजी बोली, 'अवर सुदहारी दुक्टि से बोदें बोक्य लड़का आया हो तो बताओं।"

Fig.2: Gaussion Noisy Image

#### 2 तंत्रतर

है। उनका अनुभय तो केवल एक-दूसरे के बनाव-प्रांगार व शक्त-बुरत तक ही सीमित रहता है। उनकी परिवार से न्या मतबब ! धर्म-जाति से भी कोई सरोकार नहीं। ऊँच-नीच व्या

बलाहै, इससे वे कीसों दूर रहते हैं।" "मगर हमारे कर में तो अभी तक नमें विचार छूतक भी

नहीं गये। क्यों प्यारे, सुम्हारी क्या राय है ?"
"हाँ, माताजी! मैं पिताजी और आपको जिलारों से पूर्णतया

सङ्क्षत हूँ। आपके होते हुए हमें अपने भविष्य की क्या चिन्ता?"

माता-पिता भी जब अपने पुत्र की सहमति मिल गई तो जन के पुत्र कर सहिता है कि स्वाप्त कर के किया जाता कर के किया जाता कर । इस गर माता अपनी मात्र की कारी रखती हुई झंदर साहुम से किया जाता के कारी रखती हुई झंदर साहुम से किया जाता के किया जाता कर हो। खी कि खायका कहारा ठीक है। अब हुआर स्वाप्त स्वाप्त से साहित से किया है। हाय पीले हो जायें। लड़की अपने घर जाए और खदा सुखी रहे। इसी से माहा-पिता को सानित प्राप्त होती है। आगे प्यारा स्थयं समझदार है। अपने-अस्य अपने माई-बहनों को देख लेगा।"

"ऐसा न कहें, माताजी ! माता-पिता का साबा हर समय सिर पर रहना चाहिए। अपने लिए नहीं तो अपने बच्चों पर ही तरस खानें। कृपया ऐसे विचार भन में न आया करें। इनसे बड़ा दुःख होता है।" दगी चीच आंखों में आए अश्र-जिन्दुओं को पींछते हुए

जानदेई बोली, "अरछा बेटा, ऐसा ही सही।

बूछ समय बातात्ररण शास्त रहा। पुनः माताजी बोलीं, "अगर तुम्हारी दृष्टि में कोई शोध्य लढ़का आया हो तो बताओं।"

Fig.3: Poisson Noisy Image

#### g संख्यात

है। उनका अनुभव तो केवल एव-दूसरे के बनाव-धूंगार व शक्त-तुरुत तक ही सीमित रहता है। उनकी परिवार सेन्या मतवर्ष ! अमे-जाति से भी कोई शरोकार नहीं। ऊँच-नीच नगा

मता है, इसते वे कोसी दूर रहते हैं।"
"मगर हमारे वर में तो अभी तक नये विचार छ तक भी नहीं गर्थे। क्यों प्यारे, तुम्हारी क्या दाय है ?"
"हाँ, माताजों! में पिताजी और आपको विचारों से पूर्णतया

सहमत हूँ। आपके होते हुए हमें अपने भविष्य की क्या चिन्ता?"

माता-पिता को जब अपने पुत्र को सहमति भिक्ष गई हो उन्हें बड़ो अस-नता हुई अपने आज्ञाकारी पुत्र के विचार जान करे। इस गर माता अपनो बात को जारी रखती हुई डॉवटर साहम से कोली- ''हाँ, तो मैं कह रही थी कि आपका कहना ठीक है। अब हमारा स्था भरोसा! पढ़ा-लिखा दिया, अब लो हाय गीने हो जायें। लड़की अपने घर जाए और सदा सुबी रहे। इसी से मादा-पिता को शास्ति प्राप्त होती है। आये ष्यारा स्थयं समझदार है। अपने-बाय अपने भाई-बहुनों को देख

'ऐशा न कहें, माताजी ! माता-पिता का साबा हर समय द्वार पहुंचा नाहिए। अपने लिए नहीं तो अपने बच्चों पर सिरपर रहेना नाहिए। अपने लिए नहीं तो अपने बच्चों पर ही तरस खारों। क्रमया ऐसे विचार भने में कलाया करें। इनसे

बड़ा दुःख होता है।" दगी भीच आंबों में जाए अध्-जिन्दुओं को पीछते हुए

रा नाम व्यव्हा न काम अवस्थानमुक्ता को पेक्सि हुए भावदे बेली , 'अपका नेरा, ऐशा हो सही' मुळ बनाव वातानरण ज्ञास्त रहा। पुनः माताजी बोली, 'अनर तुम्हारी बृष्टि में कोई कोम्य बढ़का आया हो तो बताओं!'

## Fig.4: Speckle Noisy Image

हैं। उनका अनुभव तो केवल एक-दूसरे के बनाव-शृंगार व शक्ल-सूरत तक ही सीमित रहता है। उनको परिवार से क्या मतलव ! 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 छूतक 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 सखी 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 दु:ख होता है।"

इसी बीच आँखों में आए अश्रु-बिन्दुओं को पोंछते हुए ज्ञानदेई 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.5: Localvar Noisy Image

#### 4.3 Output of Binarization Algorithms

#### Gaussian Noise:-

है। जनक बहुत्या तो बेनल हर-नुष्टों के जनक-प्रंतार का समुद्रा कर है कीएक दुस्ती है। जाने परिवार देखा है। जाने परिवार देखा है। जाने परिवार देखा है। जाने परिवार देखा ने परिवार के किए के प्रतिकृति के प्र

भागा जिया है दूर में सहर्यों किस वर्ष में मार्गा जिया है कर के सामग्रे दूर है जिया रहा जो कर मार्ग कर में कर के सामग्रे दूर है जिया रहा कर है जो कर मार्ग कर में कर दे कर एक है दिस्त रहा कर है के मार्ग कर में कर दूर के कि सामग्रे कर मार्ग कर है किहें कर कर कर के मार्ग कर के मार्ग कर है कर में के सामग्रे कर के कर मार्ग कर है के स्वाप कर है है जो के सामग्रे कर में कर मार्ग कर मार्ग कर है कर मार्ग क

देश।"

क्षिता न वहूँ मामानी ज मानार्रक्ता वा मानार्यक्त मानार्यक्

#### Otsu

#### Poisson Noise:-

है। जनमा कृतवा तो बेनात एक पुरान के बनाव पर्याप स्थाप है। सन्दर्भ कर हो सीमित दूसर है। जनसे परिशार के पश माराव रीड क्योजित की भी किर्दाल है। जनसे परिशार के पश माराव रीड क्योजित की प्रति है। "जर प्रयोग कर हो जो की किल की बिनार हून कर की नहीं की को पार्ट कुमारी कर हो। जा की बिनार हून कर की नहीं की को पार्ट हुमारी कर हा है। "हा साजा की है किशानी की स्थापन हिनार में पूर्णवा माराव हूं। मारावे हो हुम हमें बाले सीम्या की स्थापन की स्थापन

चिता हैं
गानिया ने अब अपने दूत भी संदर्शि विकार हो
गानिया ने अस्ति देता अस्ति देता विकार है
गानिया ने अस्ति अस्ति अस्ति अस्ति देता विकार वर्ष भा कर पात्रा अस्ति अस्ति अस्ति अस्ति अस्ति देता विकार स्ति क्षेत्री कर स्ति के दूर हो भी कि आधान दहता स्ति है। यह दूसाय क्या करोशा ! यहारिया विकार, क्या हो इस्ति हिन्दे क्यों अस्ति उस्ति अस्ति प्राप्त भी प्राप्त अस्ति क्षा सि से सामिया है। वर्षन स्ताति अस्ति प्राप्त क्षा स्ति अस्ति पार्य एस स्मामार है। वर्षन स्तात अस्ति प्राप्त होंने स्ति

। "ऐता न वहुँ, माशाजी ! साता-पिता का साबा हर समय

हिंशा न बहुँ, माठाली ! जाता-चिता का तामा हा तमस् सिर रर स्वा भाहिए। अने लिए मेही वो अने बन्धों पर हो उस्त सारी हुक्या ऐवे विचार कन में न नामा करें। उससे बड़ा दुख होता है ' रशी बीच असेबों में जाए। अस्व-विच्हुओं परे नीक्से हुए जानोर्स बोली, 'अध्या बेटा, ऐसा ही सही।'

## उपन्यसम्माः, अध्यावदा, एका ही सही।" कृत समय बातावरण शास्त्र रहा। पुनः सातावी वीनी, "अतर तुम्हारी वृष्टि में कोई बोध्य सद्वत आधा हो तो बताओ।"

#### Otsu

## Speckle Noise:-

है। जबका बहुत्या तो बेबार एक-पूर्व के बतान मुंगार व महत्त्वत्र जहाँ हो बीची पहुंचा है। जबकी परिशार है बता मामार डोजन्मों में कि भी की परिशार होता है। जनकी प रोग मामार डोजन्मों में की भी की परिशास होने, जनकी प रोग प्रवाद पूर्वों प रोगों अभी तक जी बिचार सु तब भी मीत रोगों में हों भी अभी तक जी बिचार सु तब भी पूर्वी नारांगी हैं मिताओं और सावश्री विचार से पूर्णका मामार पूर्वें भी अभी तक मामार बिचार से पूर्णका मामार डोजां में सावश्री होता है। सावश्री विचार से पूर्णका मामार डोजां में सावश्री होता है। सावश्री विचार से प्रवाद स्थार से पूर्णका

साध्य हूँ । अपने होते हुए हों जारे पोरण्य में साध्य में क्षा कर भी हुए हों जारे पोरण में प्राव किया होते । जारे का निकार हाते हैं के स्थान हात्य है जी ते । जारे का निकार हात है जी ते । जारे का निकार हात है जी ते । जारे का निकार हात है जी ते आप होता है जी ते आप होता है जारे हात हो है जी ते आप होता है जी ते आप हता है जी ते आप होता है जी तही है जी ते जी हता है जी तही है जी तही

#### Otsu

#### Adaptive

#### g piant

. नक्ष्म है। उनका क्यूचन तो केवल एक-कूलरे के जनकर/जार व मकत नुता जह है। बीहित पहल है। उनको परिवार केवला मजार ! अमेन जीन हैं भो की परिवार नहीं अमेन मजार ! क्यूचने में हैं भो की परिवार नहीं अमेन करा मजार हुआ रहती जाने कर मौं हिमार मुझ्क मही मों। ओ पारे, मुझ्क रों का पार है!" मुझ्क मों हैं भी मजारी मों राज्य कर में महान हों। अमारे हैं के बुद्ध में कारी मों राज्य मां मानक हूं। अमारे हैं के बुद्ध में कारी मों राज्य मां

#### Adaptive

#### 8 संदर्ध

करण है। आपके होते हुए हैं जो कर दिल्ला की प्रात्त किया हैं। अपने होते हुए हैं जो कर दिल्ला की प्रात्त किया हैं। अपने हिन्दी होता की प्रात्त की प्रात्त किया है। अपने हमें किया होता कर होता होता कर है। अपने प्रात्त के जीन हमें हमाने कर होता है। जी कर होता है। अपने हमाने के जीन हमें हमें के प्रात्त है। जी कर हो भी हमाने हमाने हमाने हमें हमाने हमाने

## Adaptive

# 5 PM है। जनक कहुता को बेनल हर-तुमी के काम-पूर्वार समझान कर है तीमित हरते हैं। जन्म तरिवार केसा समझान कर है तीमित हरते हैं। जन्म तरिवार केसा समझान है। जन्म तरिवार केसा है। जन्म तरिवार केसा है। जन्म तरिवार केसा तिवार केसा तिवार केसा तरिवार केसा तिवार तिवा

मार्गा कि एवं है दूर में मुश्ती कि दिने में स्थान कि दूर में मुश्ती कि दिने के स्थान के दूर में मुश्ती कि दिने के स्थान के दूर में कि दिने के स्थान के दूर में कि दिने के स्थान के दूर में कि दूर में कि दिने के स्थान के दिने के दूर में कि स्थान के स्थान के

नेवा! 'ही। न बहूँ शावती ! माम हैता सर समा हूँ तमस दित पर पूरा चाहिए। बारे किए महितों करा बच्ची पर हित पर बारे इसार ही मितार में दें द बात करें। इसे बार पूर्व हैता है! 'ही कि मोडी में बार म्म दें द बात करें। इसे बार बार हो हैते हैं! 'हा बार करा करा माने हैं। हो हैता करा करा है। पुज्य कर कारपार करान पूछ। वह नामार्थ शेल, पर तुमारे हुं होटे में से सेम बहम माम हो हो बारों।'

#### Sauvola

है। उनका बनुष्य तो केवल एक-तुसरे के बनाव-प्रंगार व सकत-पूरत तक ही सीमित रहता है। उनको परिसार से नया मत्रवय । अमे-बासि से भी कोई सरोक्तार नहीं। ऊंच-नीच नया बना है, सारे वे कोसी हुए रहते हैं।" "मधर हुआरे पर में तो अभी तक नये विचार कू तक भी

"ज्यार पूर्वार कर जैंगे हो को कर नहें निवार कु वक में भी के को जो कर मुक्त देवा जात है! "तुंत नामानी में निवारों को का जाने निवार के कुनेवा अपना कर हैं। आपने हैं में हुत हुई में तुने देवा कर में क्या निवार हैं! अपने हुत हुई में तुने देवा कर में क्या निवार हैं! जो कर अपने तुन नी अपनोंत निकार में की जातें को कमना हुई कारी आधारतारों पूर्व के निवार नाम निवार में जो हैं. जो में कहा हो भी कि वाचना कहते हुत है जो हैं. जो है कहा हो भी कि वाचना कहते कुन मी के हो जो निवार कर में की कि वाचना कहते हुत मी के हो जो कहते अपने में प्राप्त होंगे हैं। को की कारण कर हुत हुत मी के हो जो कहती अपने मा कारण के प्राप्त कर होते हैं। यह तो के सामानिवार की प्राप्त प्राप्त होंगे हैं। ध्यारा स्थयं समझदार है। अपने-अस्य अपने माई-बहनों को देख

सेना।"
"हैगा न वहुँ सावाजी ! सावाजिया का सामा हर समय सिर पर राजा भाकिए। अपने लिए अवृति तो अपने बच्चों पर ही उस सामा अपना देशे शिवार भन में क आया करें। इससे बहा कुछ होता है।" ही बीच आयों में बार अस्वित्युओं को नीधने हुए आपने में सामा अपने हैं। हो सा है। हो।" असमा अस्वत्यास्त्र सामा करने स्वत्यास्त्र सामा सी हो हों

...., ज्व्छा बदा, एसा ही सही।" हुळ समय बातावरण जान्स रहा। पुनः माताजी बोर्नी, "अनर तुम्हारी दृष्टि में कोई बोध्य लड्का आया हो तो बताजो।"

#### Sauvola

#### 8 situat

है। जबका बहुजा तो बेजार एक-तुरारे के बातक-पूरार व सकत्तुत्वत कहा हो होतित बहुता है। उठकरी जरियार होता साजार । उठकरी जो भी को दि राष्ट्रास होता है। उठकरी जरियार होता बाता है, जाती के कोनें हुए रही हैं। "पार हुवारे पत्नी आजे तक जो निचार कुछ के मही होने को 'जर्र हुवारे देखा ता है।" "ता, माहानों हैं। साजार की ता आपके विचार के सहाज हुवार पहला की का साजार है। विचार की सहाज हुवार है। अपके होते हुत हुने कारी चीजर की का

## Sauvola

#### Localvar Noise:-

हैं। उनका अनुभव तो केवल एक-दूसरे के बनाव-प्रंगार वे शक्त-पूरत तक ही सीमित रहता है। उनको परिवार से नया मततव । धर्म-जाति से भी कोई सरोकार नहीं। ऊँच-नीच क्या बता है, इससे वे कोमों दूर रहते हैं।"

"मगर हमारे घर में तो अभी तक नवे निचार खूतक भी नहीं यथे। वर्षों प्यारे, तुम्हारी क्या राय है?" "हाँ, माताओं! मैं पिताजी और आपके विचारों से पूर्णतया सहमत हूँ। आपके होते हुए हमें अपने भनित्य की क्या

भागा-पिया को जब अपने पुत्र की सहमति मिल गई ते जब्दे बड़ी समलता हुई अपने आसाकारी पुत्र के निवार जा रूप है कि स्वार अपनी अंत को जारे पखती हुई बेंक्स साहब के बोली—'ही, तो मैं बढ़ रही भी कि आपका कहन होत्र की हो अब होता का रेसी में पुत्र के लिखा है हैं हाल पीते हो अब होता का रेसी मा पुत्र का जाए और सदा मुख रहे। इसी से सत्तर देना को सामित प्रत्य होती है। आं भारत वर्ष समझदार है। अपने आप अपने भारत्वहारी को की

लेगा।"
"ऐसान कहें, माताजी ! माता-पिसाका सावा हर सक पिसान कहें, माताजी ! माता-पिसाका सावा हर सक पिसपर रहना चाहिए। अपने लिए नहीं तो अपने बच्चों हो तरस खार्ये। इत्या ऐसे विचार मन में न लाया करें। इन बहा दू:ख होता है।"

क्षा पुरुष होता हु। इसी बीच अधिं में आए अधु-बिन्दुओं को पॉछते हुए शावबंदें बोसी, 'अच्छा बेटा, ऐसा ही सही।'' कुछ समय बाताबरण झालर रहा। युनः माताओं बोसी, ''अगर तुम्हारी बृष्टि में कोई बोध्य सकुका आया हो ती

Otsu

हैं। जनका अनुभव तो केवल एक-दूसरे के बनाव-शूंबार व मनन-सुरत तक ही सीमित रहता है। उनको परिवार से क्य मतसब ! अमै-जाति से भी कोई सरोकार नहीं। जैन-नीच नय बना है, इससे वे क्सेसों दूर रहते हैं।"

नहीं गये। क्यों प्यारे, तुम्हारी क्या राग है ?" "हाँ, माताजी ! मैं पिताजी और आपके विचारों से पूर्णतय सहमत हूँ। आपके होते हुए हमें अपने भविष्य की क्य

ाच्या शिकार को जब अपने पुत्र की सहस्रोत्त मिल गर्दे अ प्रश्नित अपनामा हुई जाने सोताकारी देव के शिकार देवा कर एक पर पाता अपनी आताको वार्त एक्सी रही काला सहस्र के सोला- "हुई, तो मैं कहा पुरत्ती की काला सहस्र प्राह्म के सोला- "हुई, तो मैं कहा पुरत्ती की काला स्वार्त करें हुई । अब हुमारा बचा परोसा ! पहाने निवार दिया, अब क हुमा पत्ती को जाने सहस्रोत की पर जाए और सारा हुई पहें। सो हो महात हुई आता अपने पर जाए और सारा हुई पहें। सो हो महात हुई अपने आता अपने मार्ट-कहाँ को है

"ऐसान कहुँ, माताजी ! माता-पिता का सामा हर सम् सिर पर रहना चाहिए। अपने लिए नहीं तो अपने बच्चों प ही तरस खामें। क्रुप्या ऐसे विचार मन में न लाया करें। इन्हें बड़ा दुःख होता है।" इसी बीच आंखों में आए अबु-सिंग्डुओं को पॉक्टते हुं। बातदेडें अधि "अच्चा कर ग्रेस की समी ।"

यका दुःज होता है।" इसी शीच जीवों में आए अब्दु-विन्दुओं को पींछते हुए झानदेंदें जोती, 'अच्छा बेटा, ऐसा ही मही।' कुछ समय बातावरण कान्त रहा। पुत: माताजी योतीं, ''अयर पुन्हारी दुष्टि में कोई योध्य लड़का आया हो तो

Adaptive

है। जनका अनुसन तो केनल एक-पूसरे के बनाब-शूंगार व गनज-मूरत तक ही सीमित रहता है। जनको परिवार से बचा मतनब ! अमे-जाति से भी कोई सरोकार नहीं। ऊँच-भीच वय जना है, इससे के बोमों दर रहते हैं।"

"सगर हवार घर मता आमातक नव ावचार छूतक नहीं गये। नवों प्यारे, सुन्हारी नवा राय है ?" "हां, माताओं! मैं पिताओं और आपके विचारों से पूर्णत सहमत हूं। आपके होते हुए हमें अपने भनित्य की व

मता-चिवा को जब अपने पुत्र की बादारित मिल पाँड तो अपने बात सम्माना हुँद अपने आताकारी पुत्र के विध्याना हुँद अपने आताकारी पुत्र के विध्याना हुँद अपने आताकारी पुत्र के विध्याना कर है। इस पर माता अपनी बात को आदी पत्रकी हुँद अंतर कर माद्रकी सोनी—"हुँत तो मैं कह पुत्रे की कि अपने कर है। शिव्ह है। अब हुमारा नवा परीसा! पद्मानीच्या दिवा, अब तो तो हुए पार्थित है। अब हुंद ते कहा की अपने पत्र पत्र होते हैं। पूर्वी दे से बता-दिवा को बारित प्राप्त होते हैं। अपने प्राप्त स्वर्ध नामस्वार है। अपने अपने भाई-अहते होते

प्रभाग कहुँ, माताजी ! माता-पिता का साया हुर सक् विर पर रहना चाहिए। अपने लिए नहीं तो अपने यच्चों प ही तरस खायें। हुपया ऐसे विचार मन में न साया करें। इस

बहा हु:बा होता है!" इसी यीज भाषों में आए अब्दु-चिन्दुओं को पीछते हुए शानवेद वोली, "अब्छा बेटा, ऐसा ही सही।" कुछ समय बातावरण शान्त रहा। दून: माताओ वोली, "अगर सुन्हारी इंटिंग में कोई योग्य सहका आया हो तो

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					
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)

Number of Errors (Sauvoia)					
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.

- Adaptive is working better on Gaussian, Poisson and Localvar noisy document images because errors are less as compared to Otsu and Sauvola.
- 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," 18<sup>th</sup> International Conference on Pattern Recognition (ICPR'06) vol. 3, pp.742-745, 2006
- 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
- Pavlos Stathis, Ergina Kavallieratou and Nikos Papamarkos "An Evaluation Survey of Binarization Algorithms on Historical Documents" 19<sup>th</sup> 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).