Evaluation of document binarization using eigen value decomposition

Deepak Kumar, M N Anil Prasad and A G Ramakrishnan
Medical Intelligence and Language Engineering Laboratory
Department of Electrical Engineering
Indian Institute of Science, Bangalore, INDIA-560012

ABSTRACT

A necessary step for the recognition of scanned documents is binarization, which is essentially the segmentation of the document. In order to binarize a scanned document, we can find several algorithms in the literature. What is the best binarization result for a given document image? To answer this question, a user needs to check different binarization algorithms for suitability, since different algorithms may work better for different type of documents. Manually choosing the best from a set of binarized documents is time consuming. To automate the selection of the best segmented document, either we need to use ground-truth of the document or propose an evaluation metric. If ground-truth is available, then precision and recall can be used to choose the best binarized document. What is the case, when ground-truth is not available? Can we come up with a metric which evaluates these binarized documents? Hence, we propose a metric to evaluate binarized document images using eigen value decomposition. We have evaluated this measure on DIBCO and H-DIBCO datasets. The proposed method chooses the best binarized document that is close to the ground-truth of the document.

Keywords: binarization, evaluation, eigen value decomposition, threshold, degraded documents, document quality measure

1. INTRODUCTION

One of the key steps in document image processing is to form a two class labeled binary image.[11] There are several algorithms for binarization of document images. Majority of the algorithms can be categorized into supervised and unsupervised methods. In supervised methods, a machine learning algorithm is trained to binarize an image based on a training dataset and then is tested on the test dataset. In unsupervised methods, document image statistics are used to segment the document image into two parts. Image statistics used in these methods may be of local or global nature.[12] If a document image is clean (without any degradation), then global statistics is sufficient. For binarizing a degraded document image, local image statistics is necessary. In our experiments, we use unsupervised methods for binarization and the binarized images are evaluated using the proposed metric to rank them.

The second key step in document image processing is the recognition of binarized image. Generally, a page layout analysis is performed on the binarized image. This analysis creates a tree with branches as paragraphs, sub branches as lines and words. Individual word images consist of connected components, which may be parts of a character, full characters or merged characters. Individual connected components are classified using a trained classifier, which needs to be script-specific. Combining all scripts in a single optical character recognition (OCR) engine is an impracticable proposition, from the point of training a classifier, reliability and possible confusions. This restricts recognition by a classifier to one or two scripts.

If there is any error introduced in binarization, then one can observe deterioration in the recognition performance of the system. We can avail a feedback from the second stage to the first, but it results in increased

Further author information: (Send correspondence to Deepak Kumar)
Deepak Kumar: E-mail: deepak@ee.iisc.ernet.in, Telephone: +91 80 2293 2935
M N Anil Prasad: E-mail: anilprasadmn@ee.iisc.ernet.in, Telephone: +91 80 2293 2935
A G Ramakrishnan: E-mail: ramkiag@ee.iisc.ernet.in, Telephone: +91 80 2293 2556

computational time. Suppose a document image is binarized and the classifier mis-classifies connected components due to issues like split or merge of characters or merging of words. Then passing this information back to binarization stage increases the complexity of the system and also the computational time with a loop in the system. Since, the recognition stage is script dependent, the feedback based system also results in script dependent strategies relating to the script.

If we deploy several stable and known binarization algorithms while trying to avoid the merging of characters or words with the background and also to counter the degradations in the document images, a new problem arises as to which binarized image is the best. We have attempted to answer this question through a mathematical framework. Badekas and Papamarkos have tried to combine optimally binarized images using a self-organizing map neural network.[16,17] Su et. al have proposed a method to classify uncertain pixels from binarization methods, using local neighborhood statistics of foreground and background pixels.[29] Instead of combining, we can use some critical measures that exist in the literature to select the best binarized image. These measures are covered in the next section on related work. The measure must be script independent and able to work on degraded documents, different page layouts and historical documents. The evaluation measure can be utilized to choose the best binarized image, which avoids the complication in successive stages. Multiple binarized images can be used in semi-automated systems to reduce manual task at recognition by choosing the best binarized result. Kumar et. al have developed and used semi-automated system to benchmark camera captured word images.[31] Our measure can be an additional tool to automate the manual task between the segmentation and recognition stages.

We have used our measure to evaluate known binarization algorithms on DIBCO[24,30] and H-DIBCO[25] datasets. The results of our measure are tabulated and compared with the ground-truth binarized images.

2. RELATED WORK

Evaluation of image segmentation has been dealt with, in the literature. Generally, these techniques estimate the fitness measure on the segmented regions and the consistency of the fitness measure inside the region helps to choose a segmented image. The fitness measures are evaluated on unsupervised algorithms used for image segmentation. Do we have a similar measure for a binarized document image?

In his paper titled "Document Image Analysis: Automated Performance Evaluation," George Nagy[9] points out that except for OCR, where several methods have been applied to significant volumes of data, little has been attempted by way of automated evaluation of DIA systems.

We have a few measures in the literature for automated evaluation of binarization.[13] Different binarization techniques in the literature are described in Section 3 and their results with this measure are tabulated in the results section. One of the earliest papers on evaluation of algorithms by Palumbo et.al provides insights on global, local, contrast and edge based thresholds.[6] They conclude that global threshold should be applied in "ideal" conditions, where the transition between foreground and background is consistent with minimal amount of noise.

We can broadly classify evaluation strategies into supervised-pixel, supervised-component and recognition-based. In the recognition-based strategy, OCR results are used for evaluation. Trier and Taxt[7] study the performance of local thresholding algorithms. Further, to cover other types, Trier and Jain evaluate different global and local threshold algorithms based on the number of digits properly classified.[8] To study a complete system, after binarization, they perform classification by algorithms considered in the study. He et.al use ABBYY OCR to compare the binarization algorithms on historical documents.[15] Gatos et.al evaluate binarization technique on low quality historical documents.[14]

In supervised-pixel type, image statistical information can be used. Stathis et.al use measures such as mean square error (MSE), signal to noise ratio (SNR), peak signal to noise ratio (PSNR) and pixel error rate (PERR).[21,22] Stathis et.al carried out evaluation on historical documents.[23] Sometimes, synthetically generated noise is added to the document image for evaluation of the binarization algorithms.

In supervised-component type, connected components in the ground-truth of document image are used in the evaluation. Kefali et.al study the performance of binarization algorithms on old Arabic documents.[27] To evaluate

algorithms, they prepare signatures and perform edit-distance between test images and stored signatures. Similarly, Nitrogiannis et.al have extensively evaluated binarization algorithms with different evaluation criteria.[20] In a supervised method, Nitrogiannis et.al compare the skeleton of the outputs of different binarization algorithms with reference to the ground-truth, in terms of f-measure, precision, recall, false alarm, missing text, broken text and deformations. These measures are helpful in benchmarking algorithms on a small set of document images obtained randomly from a huge dataset. Indeed, these measures are used to evaluate binarization algorithms that were submitted for the DIBCO[24] and H-DIBCO[25] challenges. However, we require ground-truth information to evaluate their performance.

Apart from the above mentioned evaluation approaches, we had a new question in mind. All methods in the literature look at the binarized image from the viewpoint of ground-truth or recognized characters. Can we have information about the quality from the binarized image itself? In this aspect, we have designed a novel measure for evaluation of binarized images. In this paper, we have two novelties. First, change in viewpoint for evaluation and second, measure information from binarized images.

3. ANALYZED BINARIZATION ALGORITHMS

Binarization algorithms used in our analysis cover both global and local threshold methods. In global threshold methods, we analyze Otsu, Kapur and Kittler methods. In local threshold methods, we analyze Niblack, Sauvola and Su methods. We do not perform recognition of binarized images, since our measure calculates information directly from the binarized images.

3.1 Otsu's method

Otsu proposed a way to threshold an image using the histogram of the image.[1] Otsu utilized Fischer discriminant function to estimate the threshold. The value of discriminant function is calculated for assumed threshold values from the minimum to the maximum gray value. The gray value at which the discriminant function is maximum is used as the optimal threshold to split the histogram into two parts and binarize the document image.

The discriminant function used to calculate the optimal threshold ($k*$) is:

$$\sigma^2(k*) = \max_k \frac{[\mu_T\omega(k) \ - \ \mu(k)]^2}{\omega(k)[1 \ - \ \omega(k)]} \tag{1}$$

where,

$$\omega(k) = \sum_{i=1}^{k} p_i \tag{2}$$

$$\mu(k) = \sum_{i=1}^{k} ip_i \tag{3}$$

$$\mu_T = \sum_{i=1}^{L} ip_i \tag{4}$$

Here, 'L' is the total number of gray levels and 'p_i' is the normalized probability distribution obtained from the histogram of the image.

3.2 Kapur's method

Kapur proposed an entropy measure to threshold a document image.[3] It is an extended version of Pun's method.[2] Document image is split into foreground and background and the entropy of each is calculated. The optimal threshold is obtained by maximizing the sum of foreground and background entropy:

$$H(k*) = \max_{k}(H_f(k) + H_b(k)) \tag{5}$$

$$H_f(k) = \sum_{i=1}^{k} -\frac{p_i}{p(k)} \log \frac{p_i}{p(k)} \tag{6}$$

$$H_b(k) = \sum_{i=k+1}^{L} -\frac{p_i}{1 \ - \ p(k)} \log \frac{p_i}{1 \ - \ p(k)} \tag{7}$$

$$p(k) = \sum_{i=1}^{k} p_i \tag{8}$$

Here, '$H_f(k)$' and '$H_b(k)$' are foreground and background entropies, respectively.

3.3 Kittler's method

Kittler proposed several methods based on window weights.[4] Here, we assume the entire image as a single window and obtain the pixel weights using the gradient image and compute the threshold. Several local region based algorithms are available as variants for this method. The threshold is calculated as:

$$T = \frac{\sum_{i=1}^{M} \ \sum_{j=1}^{N} g(i,j) * f(i,j)}{\sum_{i=1}^{M} \ \sum_{j=1}^{N} g(i,j)} \tag{9}$$

Here, '$f(i,j)$' is the image gray value at pixel position (i,j) and '$g(i,j)$' is the gradient image obtained by applying Sobel operator on the input image.

3.4 Niblack's method

Niblack proposed an algorithm to calculate a local threshold for each pixel by moving a rectangular window over the whole image.[5] The mean and the standard deviation of all pixels in the window are used in arriving at the threshold. Thus, the threshold is given as:

$$T(x,y) = \mu(x,y) \ + \ k_n \ * \ \sigma(x,y) \tag{10}$$

$$\mu(x,y) = \frac{1}{N_w} \sum_{i} \sum_{j} \ f(x+i, y+j) \tag{11}$$

$$\sigma^2(x,y) = \frac{1}{N_w} \sum_{i} \sum_{j} (f(x+i, y+j) \ - \ \mu(x,y))^2 \tag{12}$$

Here, a user needs to fix the values for 'k_n' and the size of the window. These values are dependent on the document image. We have used $k_n = -0.2$ and $N_w = 25$ as the window size. In our experiment, these values are fixed for all the document images.

3.5 Savuola's method

Sauvola proposed a modification for Niblack algorithm.[10] Documents imaged with a background containing a light texture or variation and uneven illumination were considered to study the binarization. The modification in the estimation of the threshold is :

$$T(x,y) = \mu(x,y) * \left[1 + \ k_s * \left(\frac{\sigma(x,y)}{R} - 1\right)\right] \tag{13}$$

Here, 'R' is the dynamic range of the standard deviation $\sigma(x,y)$ across all the windows. In our experiment, $R = 128$, $N_w = 25$ and $k_s = 0.2$ are the values used in threshold estimation.

3.6 Su's method

Su et.al proposed a binarization method using local image statistics.[26] This method has been used in evaluating H-DIBCO and DIBCO datasets. To suppress local background variation, Su used,

$$D(x,y) = \frac{f_{max}(x,y) \ - \ f_{min}(x,y)}{f_{max}(x,y) \ + \ f_{min}(x,y) + \epsilon} \tag{14}$$

Here, '$f_{max}(x,y)$' and '$f_{min}(x,y)$' are the local maximum and minimum intensities in a 3x3 neighborhood window, respectively and ϵ is a small positive number. Otsu threshold is applied on the suppressed image $D(x,y)$ to obtain high contrast image pixels. The pixel is classified based on

$$B(x,y) = \begin{cases} 1 & N_e \geq N_{min} \ \&\& \ f(x,y) \leq E_{mean} + E_{std}/2 \\ 0 & otherwise \end{cases} \tag{15}$$

where E_{mean} and E_{std} are the mean and the standard deviation of image intensities of detected high contrast image within the neighborhood.

$$E_{mean} = \frac{\sum_{neighbor} f(x,y) \ * \ (1 - E(x,y))}{N_e} \tag{16}$$

$$E_{std} = \sqrt{\frac{\sum_{neighbor}((f(x,y) \ - \ E_{mean}) \ * \ (1 - E(x,y))^2}{N_e}} \tag{17}$$

where N_e is the number of high contrast pixels in the neighborhood window. $E(x,y)$ is equal to 0 if the pixel is detected as a high contrast pixel and 1, otherwise. In our experiment, N_{min}, the minimum number of high contrast pixels and the neighborhood window size are set to 25 and 15 pixels, respectively.

4. PROPOSED EIGEN VALUE DECOMPOSITION METHOD

Here, we provide a mathematical framework based on Fischer discriminant analysis.[18] The proposed method does not yield any new binarization method. But, it evaluates the binarization results from different binarization algorithms. This method is evolved from Principle Component Analysis (PCA).

A generic Fischer linear discriminant has the following criterion function for binary classification:[18]

$$\max \ \mathbf{J} \ = \ \frac{|\mu_1 \ - \ \mu_2|^2}{\sigma_1^2 \ + \ \sigma_2^2} \tag{18}$$

The maximum value of $\mathbf{J}$ will yield the best separation between the two classes. Otsu's method was developed based on this criterion. $\mu_1, \ \mu_2$ are means of classes 1 and 2, respectively. $\sigma_1^2, \ \sigma_2^2$ are variances of classes 1 and 2, respectively.

We assume the variation between the class means is minimal and replace with a constant **K**. This assumption is helpful to prove the theorem stated below. Thus, the modified discrimination function is

$$\max\ \mathbf{J}\ =\ \frac{\mathbf{K}}{\sigma_1^2\ +\ \sigma_2^2} \tag{19}$$

The feature vector for each pixel of a binarized image for each class consists of pixel gray value and the 'x' and 'y' coordinates of the pixel. If only the gray values are considered, then the feature will be a scalar. If spatial information is also added as attributes, then the feature vector will have a dimension of three. Then the feature matrix for the entire image contains three columns. We can also incorporate color information into these feature vectors. We normalize each column in the feature matrix. Thus, the variances of each column lie in the range $[0, 1]$. In the case of a scalar feature,

$$0 < \sigma_1^2, \sigma_2^2 < 1$$

Using the theorem given below, we can show that over the set of binarization methods,

$$\max\ \mathbf{J_1}\ =\ \frac{1}{\sigma_1^2\ +\ \sigma_2^2}\ <\ \min\ \mathbf{J_2}\ =\ \frac{1}{\sigma_1^2\ *\ \sigma_2^2}$$

Theorem 1. The sum of any two non-negative real numbers is greater than their product, if the product is less than '4.'

With strong duality gap, we can modify further and state the discrimination function as,

$$\max\ \mathbf{J_3}\ =\ \sigma_1^2\ *\ \sigma_2^2 \tag{20}$$

Thus, the modified Fischer discrimination function is maximized by maximizing the product of variances. The expected covariance matrix of class-1 feature vector is

$$\Sigma_1\ =\ \frac{1}{N1}\ \sum_{i=1}^{N1}\ (x_i^1\ -\ \mu_1)(x_i^1\ -\ \mu_1)^T \tag{21}$$

where,

$$\mu_1\ =\ \frac{1}{N1}\ \sum_{k=1}^{N1}\ x_k^1 \tag{22}$$

For zero mean ($\mu_1\ =\ 0$) feature vector,

$$\Sigma_1\ =\ \frac{1}{N1}\ \sum_{i=1}^{N1}\ x_i^1 x_i^{1T} \tag{23}$$

The eigen value decomposition of Σ_1 will have $\sigma_{11}^2, ..., \sigma_{1d}^2$ as elements of the diagonal matrix.[19] The expected covariance matrix is symmetric and positive definite. Similar to PCA, product of eigen values is used in our evaluation.

The determinant of Σ_1 is the product of all the diagonal elements in the diagonal matrix obtained from eigen value decomposition.

$$det(\Sigma_1) = \sigma_{11}^2 * \sigma_{12}^2 * ... * \sigma_{1d}^2 \tag{24}$$

The determinant is used, if the feature used is more than just the gray value. More feature vectors can be added to improve the evaluation process. We use normalized gray values, and the row and column indices, whereas, Otsu had used only the gray values.

With only the gray values as a scalar feature,

$$det(\Sigma_1) = \sigma_{11}^2 \tag{25}$$

From equation (20), the product of variances is

$$det(\Sigma_1) * det(\Sigma_2) = \sigma_{11}^2 * \sigma_{12}^2 * ... * \sigma_{1d}^2 * \sigma_{21}^2 * \sigma_{22}^2 * ... * \sigma_{2d}^2 \tag{26}$$

We have used the determinant to form product of eigen values obtained from decomposition. Complexity of incorporating addition of variances is avoided by the products of the variance determinants.

4.1 Preprocessing for evaluation

We obtain binarized results for Otsu, Niblack, Kapur, Kittler, Savuola and Su algorithms. We form a feature matrix for each class from the binarized image. The columns of these feature matrices are the normalized gray value, row and column values of the pixel. Number of rows is equal to the number of pixels belonging to that class. Thus, we have $\mathbf{X}_1$ and $\mathbf{X}_2$ as feature matrices. We estimate the covariance matrices of these two sets of vectors. The eigen value decomposition is performed on the covariance matrices. Then, the product of eigen values is calculated for each binarized image. The binarized image which has maximum product value is chosen as the best binarized image among the group.

5. EXPERIMENTAL RESULTS ON DIFFERENT DATASETS

We consider three document datasets for experimentation, namely DIBCO 2009, H-DIBCO 2010 and DIBCO 2011. DIBCO represents Document Image Binarization Contest organized in International Conference on Document Analysis and Recognition (ICDAR) 2009 and 2011. Similarly, H-DIBCO represents Handwritten Document Image Binarization Contest organized in International Conference on Frontiers in Handwriting Recognition (ICFHR) 2010.

5.1 DIBCO 2009 dataset

The DIBCO testing dataset consists of 5 machine-printed and 5 handwritten images resulting in a total of 10 images.[24] The associated ground-truth for these images is also available. The documents of this dataset have been collected from the following libraries: The Goettingen State and University Library (UGOE), The Bavarian State Library, the British Library and the Library of Congress. The images in the dataset were artificially chosen to contain degradations such as variable background intensity, shadows, smear, smudge, low contrast, bleed-through and show-through.

5.2 H-DIBCO 2010 dataset

The H-DIBCO testing dataset consists of 10 handwritten document images.[25] The ground-truth is available for these images also, which can be used for the evaluation. The document images of this dataset are from the collections of the Library of Congress. These images were also selected to contain different degradations listed above.

5.3 DIBCO 2011 dataset

The DIBCO 2011 testing dataset consists of 8 machine-printed and 8 handwritten images resulting in a total of 16 images.[30] The associated ground-truth for these images is also available for the evaluation. The documents of this dataset are collected from the following libraries: The Goettingen State and University Library (UGOE), The Bavarian State Library, the British Library and the Library of Congress. These images were selected to contain different degradations listed above, and are different from earlier two datasets.

5.4 Existing evaluation measures

Evaluation of binarized image can be carried out based on several measures. We discuss a few evaluation measures used in the literature, which were also evaluated along with the proposed measure. These measures were used in DIBCO and H-DIBCO competitions.

5.4.1 F-measure

F-measure(FM) is the harmonic mean of precision (p) and recall (r). F-measure is calculated at the pixel level. Each pixel is classified as one of these labels, namely True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

$$FM = 2 * p * r/(p \ + \ r) \tag{27}$$

where,

$$recall = \frac{TP}{TP \ + \ FN} \tag{28}$$

$$precision = \frac{TP}{TP \ + \ FP} \tag{29}$$

5.4.2 Negative Rate Metric

The negative rate metric NRM is based on pixel-wise mismatches between the ground-truth and the prediction. It combines false negative rate NR_{FN} and false positive rate NR_{FP}. It is denoted as follows:

$$NRM = \frac{NR_{FN} \ + \ NR_{FP}}{2} \tag{30}$$

where,

$$NR_{FN} = \frac{N_{FN}}{N_{FN} \ + \ N_{TP}} \tag{31}$$

$$NR_{FP} = \frac{N_{FP}}{N_{FP} \ + \ N_{TN}} \tag{32}$$

N_{TP}, N_{FP}, N_{TN} and N_{FN} denote the number of true positives, false positives, true negatives and false negatives, respectively.

5.4.3 Mean Square Error

Mean square error (MSE) is a traditional image statistics measure. It is given by,

$$MSE = \frac{\sum_{x=1}^{M} \ \sum_{y=1}^{N} \ (I(x,y) \ - \ I^{'}(x,y))^2}{MN} \tag{33}$$

Here, $I(x,y)$ and $I^{'}(x,y)$ represent the ground-truth and binarized image pixels, respectively.

5.4.4 Peak Signal to Noise Ratio

The peak signal to noise ratio (PSNR) is defined as,

$$PSNR = 10 \log \frac{C^2}{MSE} \tag{34}$$

Here, C is a constant that denotes the difference between foreground and background. This constant is set to 1.

Table 1. Performance evaluation of six binarization algorithms on DIBCO 2009 dataset using F-measure, NRM, MSE and PSNR with EVD ranks as evaluation measures

Algorithm	F-measure	NRM	MSE	PSNR	EVD1	EVD3
Su	86.86	5.29	2.85	16.52	11	11
Sauvola	85.02	7.99	2.57	16.34	19	19
Kapur	82.41	5.12	3.23	15.19	39	36
Kittler	83.92	8.13	2.76	16.37	39	41
Otsu	78.60	5.64	5.74	15.31	42	43
Niblack	46.04	14.41	20.85	6.96	–	–

Table 2. Binarization performance evaluation of six algorithms on H-DIBCO 2010 dataset using F-measure, NRM, MSE and PSNR with EVD ranks as evaluation measures

Algorithm	F-measure	NRM	MSE	PSNR	EVD1	EVD3
Su	85.18	4.96	2.20	16.81	11	11
Sauvola	73.94	17.35	3.01	15.87	21	26
Kapur	86.08	6.23	1.92	17.30	40	35
Otsu	85.43	9.36	1.82	17.52	39	36
Kittler	83.84	12.52	1.93	17.37	39	42
Niblack	35.55	15.86	23.06	6.45	–	–

Table 3. Evaluation of the performance of different binarization algorithms on DIBCO 2011 dataset, using F-measure, NRM, MSE and PSNR with EVD ranks as evaluation measures

Algorithm	F-measure	NRM	MSE	PSNR	EVD1	EVD3
Su	83.11	6.65	3.32	15.64	18	20
Sauvola	82.44	9.81	3.11	15.76	42	42
Kapur	82.73	6.84	3.80	15.56	46	45
Kittler	80.69	10.04	3.55	15.35	68	66
Otsu	82.10	8.16	4.24	15.72	66	67
Niblack	42.71	16.38	22.67	6.63	–	–

5.4.5 Results

The six algorithms were separately used to binarize the DIBCO and H-DIBCO datasets. Global thresholding algorithms like Otsu, Kapur and Kittler methods do not require any parameter tuning, unlike the local thresholding ones. Sokratis has described a tool to tune parameters for binarization,[28] where a user provides his or her feedback, when binarization is performed on a particular image with selected algorithms. User feedbacks are the input parameters to set proper binarization, for a document image, by an algorithm. In this paper, we have calculated F-measure, Negative Rate Metric, Mean Square Error and Peak Signal to Noise Ratio for all the tested algorithms on DIBCO and H-DIBCO datasets. These calculated evaluation measures are tabulated in Tables 1, 2 and 3. The algorithms are ranked based on the proposed measure. The binarization by Niblack method requires tuning of parameters, for each document, depending on the page layout and the font size. Tuning the parameters for each document improves evaluation measures. Since, it requires user feedback, we do not use Niblack method in ranking. Other algorithms are ranked based on the product of eigen values. In Tables 1, 2 and 3, EVD1 and EVD3 represent proposed measure calculated using Eqs. (25) and (26), respectively. EVD1 considers only gray values of the document image for calculation of proposed measure, while EVD3 considers the two spatial coordinate values in addition. In our ranking system, algorithms are ranked in descending order using proposed measure. EVD1 and EVD3 are the total sum of ranks of all the document images in the dataset. Algorithms in Tables 1, 2 and 3 are placed in EVD3 rank order.

6. DISCUSSION

In this paper, we have neither proposed any binarization method nor any strategies for combining binarization techniques. We have made a genuine attempt towards selecting the better binarized image using image statistics. The selected binarized image may or may not be good for recognition. But our method, in essence, provides an objective measure to choose from a set of binarized images. Further, this set of images can be used to form a

Table 4. Ranking of ground-truth image and global thresholding algorithms on DIBCO 2009, H-DIBCO 2010 and DIBCO 2011 datasets.

	DIBCO 2009		H-DIBCO 2010		DIBCO 2011	
Algorithm	EVD1	EVD3	EVD1	EVD3	EVD1	EVD3
Ground-truth	11	10	10	11	19	21
Kapur	28	26	31	28	37	36
Kittler	29	31	29	32	52	51
Otsu	32	33	30	29	52	52

combined binary image. Mathematically, we have shown that the image statistics have the capability to evaluate the binarization of a document image. In Tables 1, 2 and 3, algorithms are ordered according to EVD3 values. We can observe that, there is a slight mismatch between EVD1 and EVD3 values in the ranking process. Based on the image statistics used in the calculation of proposed measure, the binarized image rank also varies. We have used gray and spatial coordinate values for ranking a binarized image. A few researchers have tried to combine different binarization methods.[17,29] Top ranked algorithms in Tables 1, 2 and 3, if combined judiciously, may improve the evaluation measures. Iteratively, one can combine binarized images, calculate and compare new EVD measure with others. This combining procedure requires a user intervention to check whether combined image is close enough to the ground-truth image and EVD measure.

As a black box approach, we passed the ground-truth also for evaluation along with the outputs of algorithms such as Otsu, Kapur and Kittler, which rely on global threshold. Ranking of these methods are reported in Table 4. The proposed measure always ranked the ground-truth as first by placing it on top of the table. Since local threshold based algorithms require parameter tuning, we skipped them in black box approach.

7. CONCLUSION

We have proposed a novel measure to evaluate document binarization. Our measure is calculated through eigen value decomposition. The results tabulated indicate the feasibility of choosing either a binarized image or a set of top performing binarized images. Proposed measure is evaluated on three datasets namely DIBCO2009, H-DIBCO2010 and DIBCO2011. These datasets consist of document images with most types of degradations. Without any aid from a ground-truth binary image, EVD measure has pointed to the better binarized image from the set.

APPENDIX A.

Theorem 1. The sum of any two non-negative real numbers is greater than their product, if the product is less than '4.'

Proof - Let assume two distinct non-negative numbers 'a' and 'b'. The modulus of their difference is greater than zero.

$$| \ a \ - \ b \ | > 0$$

Squaring on both sides,

$$(a \ - \ b)^2 > 0$$

$$a^2 \ - \ 2ab \ + \ b^2 > 0$$

Adding by '$4ab$' from both side.

$$a^2 \ + \ 2ab \ + \ b^2 > 4ab$$

$$(a \ + \ b)^2 > 4ab$$

$$(a \ + \ b) > 2\sqrt{ab} \tag{35}$$

From the theorem, if we assume

$$a \ + \ b > ab \tag{36}$$

By combining Eqs. (35) and (36)

$$2\sqrt{ab} > ab \tag{37}$$

Thus, it results in

$$ab < 4 \tag{38}$$

From the theorem, we have

$$a \ + \ b > ab \tag{39}$$

$$\frac{1}{a \ + \ b} < \frac{1}{ab} \tag{40}$$

Over the set of different values for a and b,

$$\max_i \ \frac{1}{a_i \ + \ b_i} \ < \ \min_i \ \frac{1}{a_i b_i}$$

REFERENCES

[1] Otsu, N., "A Threshold Selection Method from Gray-level Histograms," *IEEE Trans. SMC*, **9**, 62–66 (1979).

[2] Pun, T., "A new method for gray-level picture threshold using the entropy of the histogram," *Signal Processing*, **2**, 223–237 (1980).

[3] Kapur, J.N., Sahoo, P.K. and Wong, A.K.C., "A new method for gray-level picture threshold using the entropy of the histogram," *Computer Vision, Graphics and Image Processing*, **29**, 273–285 (1985).

[4] Kittler, J., Illingworth, J. and Föglein, J., "Threshold Selection Based on a Simple image Statistic," *Computer Vision, Graphics and Image Processing*, **30**, 125–147 (1985).

[5] Niblack, W., [*An Introduction to Digital Image Processing*], Englewood Cliffs, N.J. Prentice Hall, 115–116 (1986).

[6] Palumbo, P.W., Swaminathan, P. and Srihari, S.N., "Document image binarization: Evaluation of algorithms," *Applications of Digital Image Processing IX*, **SPIE** ,**697**, 278–285 (1986).

[7] Trier, Ø. D. and Taxt, T., "Evaluation of binarization methods for document images," *IEEE Trans. PAMI*, **17**, 312–315 (1995).

[8] Trier, Ø. D. and Jain, A.K., "Goal-directed evaluation of binarization methods," *IEEE Trans. PAMI*, **17**, 1191–1201 (1995).

[9] Nagy, G., "Document Image Analysis: Automated Performance Evaluation," *Document Analysis Systems*, A. L. Spitz and A. Dengel, Eds., 137–156 (1995).

[10] Sauvola, J. and Pietikainen, M., "Adaptive document image binarization," *Pattern Recognition*, **33**, 225–236 (2000).

[11] Nagy, G., "Twenty Years of Document Image Analysis in PAMI," *IEEE Trans. PAMI*, **22**, 38–61 (2000).

[12] Gonzalez, R.C. and Woods, R.E., [*Digital Image Processing*], Second Edition, Pearson Education (2002).

[13] Leedham, G., Yan, C., Takru, K. and Mian, J.H., "Comparison of Some Thresholding Algorithms for Text/Background Segmentation in Difficult Document Images," in Proc. 7th *Intl. Conf. Document Analysis and Recognition*, 859–865 (2003).

[14] Gatos, B., Pratikakis, I. and Perantonis, S.J., "An Adaptive Binarization Technique for Low Quality Historical Documents," in Proc. 6th *Intl. Workshop on Document Analysis Systems (DAS'04)*, 102-113 (2004).

[15] He, J., Do, Q.D.M., Downtown, A.C. and Kim, J.H., "A Comparison of Binarization Methods for Historical Archive Documents," in Proc. 8th *Intl. Conf. Document Analysis and Recognition*, 538–542 (2005).

[16] Badekas, E. and Papamarkos, N., "Automatic Evaluation of Document Binarization Results," in Proc. 10th *Iberoamerican Congress on Pattern Recognition (CIARP) 2005*, 1005–1014 (2005).

[17] Badekas, E. and Papamarkos, N., "Optimal combination of document binarization techniques using a self-organizing map neural network," *Engineering Applications of Artificial Intelligence*, **20**, 11–24 (2006).

[18] Duda, R.O., Hart, P.E. and Stork, D.G., [*Pattern Classification*], Second Edition, Wiley (2006).

[19] Strang, G., [*Linear Algebra and Its Applications*], Fourth Edition, Cengage Learning (2006).

[20] Ntirogiannis, K., Gatos, B. and Pratikakis, I., "An Objective Evaluation Methodology for Document Image Binarization Techniques," in Proc. 8th *Intl. Workshop on Document Analysis Systems (DAS'08)*, 217–224 (2008).

[21] Stathis, P., Kavallieratou, E. and Papamarkos, N., "An Evaluation Technique for Binarization Algorithms," *Journal of Universal Computer Science (UCS)*, **18**, 3011–3030 (2008).

[22] Kavallieratou, E., "An objective way to evaluate and compare binarization algorithms," in Proc. *ACM Symposium on Applied Computing (SAC) 2008*, 397–401 (2008).

[23] Stathis, P., Kavallieratou, E. and Papamarkos, N., "An Evaluation Survey of Binarization Algorithms on Historical Documents," in Proc. 19th *Intl. Conf. on Pattern Recognition (ICPR) 2008*, 1–4 (2008).

[24] Gatos, B., Ntirogiannis, K. and Pratikakis, I., "ICDAR 2009 Document Image Binarization Contest (DIBCO 2009)," in Proc. 10th *Intl. Conf. Document Analysis and Recognition*, 1375–1382, http://www.iit.demokritos.gr/∼bgat/DIBCO2009/benchmark (2009).

[25] Pratikakis, I., Gatos, B. and Ntirogiannis, K., "H-DIBCO 2010 Handwritten Document Image Binarization Competition," in Proc. *Intl. Conf. Frontier Handwritten Recognition*, 727–732, http://www.iit.demokritos.gr/∼bgat/H-DIBCO2010/benchmark (2010).

[26] Su, B., Lu, S. and Tan, C.L., "Binarization of Historical Document Images Using the Local Maximum and Minimum," in Proc. 9th *Intl. Workshop on Document Analysis Systems (DAS'10)*, 159–166 (2010).

[27] Kefali, A., Sari, T. and Sellami, M. "Evaluation of several binarization techniques for old Arabic documents images," *Intl. Symposium on Modelling and Implementation of Complex Systems (MISC'10)*, 88–99 (2010).

[28] Sokratis., V. and Kavallieratou, E., "A Tool for Tuning Binarization Techniques," in Proc. 11th *Intl. Conf. Document Analysis and Recognition*, 1–5 (2011).

[29] Su, B., Lu, S. and Tan, C.L., "Combination of Document Image Binarization Techniques," in Proc. 11th *Intl. Conf. Document Analysis and Recognition*, 22–26 (2011).

[30] Pratikakis, I., Gatos, B. and Ntirogiannis, K., "ICDAR 2011 Document Image Binarization Contest (DIBCO 2011)," in Proc. 11th *Intl. Conf. Document Analysis and Recognition*, 1506–1510,http://utopia.duth.gr/∼ipratika/DIBCO2011/ (2011).

[31] Kumar, D., Anil Prasad, M.N. and Ramakrishnan, A.G., "Benchmarking recognition results on word image datasets," in *CoRR*, vol. abs/1208.6137, http://arxiv.org/abs/1208.6137 (2012).