A Binarization Algorithm for Historical Arabic Manuscript Images using a Neutrosophic Approach

Khalid M. Amin[1,5], Mohamed Abd Elfattah[2,5], Aboul Ella Hassanien[3,5] and Gerald Schaefer[4]
[1]Faculty of Computers and Information, Menofiya University, Egypt
[2]Faculty of Computers and Information, Mansoura University, Egypt
[3]Faculty of Computers and Information, Cairo University, Egypt
[4]Department of Computer Science, Loughborough University, U.K.
[5]Scientific Research Group in Egypt (SRGE), http://www.egyptscience.net

***Abstract*—In this paper, an improved thresholding approach based on neutrosophic sets (NSs) and adaptive thresholding is proposed. This is applied to degraded historical documents imaging and its performance evaluated. The input RGB image is transformed into the NS domain, which is described using three subsets, namely the percentage of truth in a subset, the percentage of indeterminacy in a subset, and the percentage of falsity in a subset. The entropy in NS is employed to evaluate the indeterminacy with a λ-mean operation used to minimize indeterminacy. Finally, the historical document image is binarized using an adaptive thresholding technique. Experimental results demonstrate that the proposed approach is able to select appropriate image thresholds automatically and effectively, while it is shown to be less sensitive to noise and to perform better compared with other binarization algorithms.**

Keywords: image binarization, thresholding, historical manuscript image, neutrosophic theory.

I. Introduction

Ancient Arabic documents typically suffer from various degradations due to both ageing and uncontrolled environmental conditions [1]. The main artefacts encountered in digitally captured images of historical documents are [1], [2]: shadows, non-uniform illumination, smear, strain, bleed-through and faint characters. Fig. 1 illustrates some examples of degraded historical Arabic manuscript images.These degradations arise either due to the physical storage conditions of the original manuscript, or because of writers having used different quantities of ink and pressure resulting in characters with different intensities and/or thicknesses as well as faint characters [1]. In addition, some documents contain extra details such as diacritics, decorations, or have writing in multiple colors.

Binarization of such document images is typically an essential pre-processing task, and hence important for document analysis systems. It converts an image into bi-level form in such way that the foreground (text) information is represented by black pixels and the background by white ones [3]. Although document image binarization has been studied for many years, thresholding of historical document images is still a challenging problem due to the complexity of the images and the above mentioned degradations [2].

Neutrosophic set (NS) approaches are relatively new and have been applied to various image processing tasks such as thresholding, segmentation and denoising [4]. In this paper, a new hybrid algorithm for binarization of degraded Arabic manuscript image is proposed, which modifies the previous algorithm of [5] to work in an adaptive manner. Experimental results demonstrate that the proposed approach is able to select appropriate image thresholds automatically and effectively, while it is shown to be less sensitive to noise and to performs better compared with other binarization algorithms.

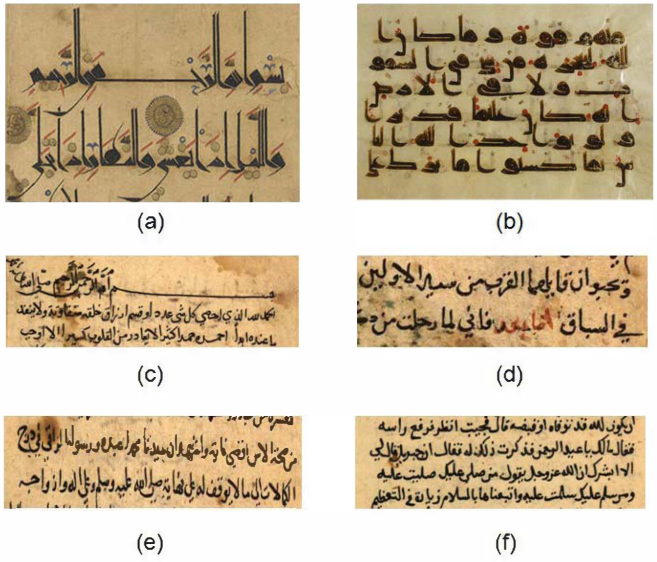

Fig. 1. Examples of manuscript images containing multi-colored text lines with different degradations.

The remainder of the paper is structured as follows: Sections II and III present related work on image binarization and neutrosophic sets. In Section IV, the proposed method for binarization of historical document images is presented. Section V gives experimental results, while Section VI concludes the paper.

II. Document Image Binarization

Approaches for document image binarization can be grouped into two main categories: global and local approaches. Global algorithms select a single threshold value for the entire image. This gives good results if there is a good separation between foreground and background. However, for historical documents, this approach is not robust enough [3]. To deal

with degradations, the current trend is to use local information that guides the threshold value, often pixel-wise, in an adaptive manner [6], [7]. Hybrid methods [3] combine global and local information to assign pixels to one of the two classes (text or background).

Otsu's thresholding algorithm [8] is the most widely employed global method. It tries to find the threshold t which separates the gray-level histogram, in an optimal way, into two segments. It maximizes the inter-class variance and minimizes the intra-class variance with the calculation of inter-class and intra-class variances based on the normalized histogram of the image $H = [\mathrm{h}_0 ... \mathrm{h}_{255}]$ where $\sum h(i) = 1$. The inter-class variance is given by

$$\sigma^2_{\text{inter}} = q_1(t) \times q_2(t) \times [\mu_1(t) - \mu_2(t)]^2, \tag{1}$$

where

$$\mu_1(t) = \frac{1}{q_1(t)} \sum_{i=0}^{t-1} h(i) \times i, \tag{2}$$

and

$$\mu_2(t) = \frac{1}{q_2(t)} \sum_{i=t}^{255} h(i) \times i, \tag{3}$$

with

$$q_1(t) = \sum_{i=0}^{t=1} h(i), \tag{4}$$

and

$$q_2(t) = \sum_{i=t}^{255} h(i). \tag{5}$$

Niblack's algorithm [9] calculates a pixelwise threshold in a sliding window fashion. The threshold t is computed by using the mean μ and standard deviation σ of all the pixels in the window, and is derived as

$$t = \mu + k \times \sigma, \tag{6}$$

where k is a constant in $[0; 1]$ that determines how much of the total print object edge is retained.

Sauvola's algorithm [10] is a modification of Niblack's approach claimed to give improved performance on documents where the background contains light texture, big variations and uneven illumination. A threshold is computed based on the dynamic range R of the standard deviation as

$$t = \mu \times (1 + k(\sigma/R - 1)), \tag{7}$$

where k is a fixed value. According to [11], this method is shown to be more effective than Niblack's algorithm when the gray-level of the text is close to 0 and that of background close to 255. However, in images where the gray-levels of background and text pixels are close, the results are unsatisfactory.

III. Neutrosophic Theory

Neutrosophy theory considers an event, concept, or entity A in relation to its opposite Anti-A and neutrality Neut-A, which is neither A nor Anti-A. Neut-A and Anti-A are referred to as Non-A. Every idea A tends to be neutralized and balanced by Anti-A and Non-A. For example, if A = "white", then Anti-A = "black", Non-A = "blue, yellow, red, black, etc." (any color except white), and Neut-A = "blue, yellow, red, etc." (any color except white and black) [12], [13].

A. Neutrosophic Set

Let U be a universe of discourse, and a neutrosophic set A included in U. An element x in the set M is denoted as $x(T, I, F)$. T, I and F are real standard or non-standard subsets of $]-0, 1+[$ with $\sup T = t\_\sup$, $\inf T = t\_\inf$, $\sup I = i\_\sup$, $\inf I = i\_\inf$, $\sup F = f\_\sup$, $\inf F = f\_\inf$, $n\_\sup = t\_\sup + i\_\sup + f\_\sup$, and $n\_\inf = t\_\inf + i\_\inf + f\_\inf$. T, I and F are referred to as neutrosophic components.

$x(T, I, F)$ belongs to A in the following way: it is $t\%$ true in the set, $i\%$ indeterminate, and $f\%$ false, where t varies in T, i varies in I, and f varies in F. T, I and F are subsets, while T, I and F are functions/operators depending on known or unknown parameters. T, I and F are not necessarily intervals, and may be discrete or continuous, single-element, finite, or countable infinite, union or intersection of various subsets, etc. They may also overlap.

B. Neutrosophic Image

Let U be a universe of discourse, and W be a set included in U, which is composed of bright pixels. A neutrosophic image P_{NS} is characterized by three subset T, I and F. A pixel P in the image is described as $P(T, I, F)$ and belongs to W in the following way: it is $t\%$ true in the bright pixel set, $i\%$ indeterminate, and $f\%$ false. The pixel $P(i, j)$ in the image domain is transformed into neutrosophic domain by

$$P_{NS}(i, j) = \{T(i, j), I(i, j), F(i, j)\}, \tag{8}$$

where $T(i, j)$, $I(i, j)$ and $F(i, j)$ are the probabilities of belonging to the bright, indeterminate and non-bright set, respectively, which are defined as

$$T(i, j) = \frac{g(i, j) - g_{\min}}{g_{\max} - g_{\min}}, \tag{9}$$

$$F(i, j) = 1 - T(i, j), \tag{10}$$

and

$$I(i, j) = 1 - \frac{Ho(i, j) - Ho_{\min}}{Ho_{\max} - Ho_{\min}}, \tag{11}$$

with

$$Ho(i, j) = |e(i, j)|, \tag{12}$$

where $Ho(i, j)$ is the homogeneity value of T at (i, j), which is described by the local gradient value $e(i, j)$, $g(i, j)$ is the intensity value of the pixe l(i, j), $g_{\min}$ and $g_{\max}$ are the minimum and maximum value of $g(i, j)$ respectively.

C. Neutrosophic Image Entropy

Th entropy is utilized to evaluate the distribution of different gray levels in an image. If the entropy is maximal, the different intensities have equal probability and the intensities thus distribute uniformly. On the other hand, if the entropy is small, the intensities have different probabilities and their distributions are non-uniform.

Neutrosophic image entropy is defined as the summation of the entropies of the three subset T, I and F, and is employed to evaluate the distribution of the elements in the NS domain:

$$En_{NS} = En_T + En_I + En_F, \tag{13}$$

with

$$En_T = \sum P_T(i) \ln P_T(i), \tag{14}$$

$$En_I = \sum P_I(i) \ln P_I(i), \tag{15}$$

and

$$En_F = \sum P_F(i) \ln P_F(i), \tag{16}$$

where En_T, En_I and En_F are the entropies of subsets T, I and F, respectively, and $P_T(i)$, $P_I(i)$ and $P_F(i)$ are the probabilities of element i in T, I and F, respectively. En_T and En_F are utilized to measure the distribution of the elements in the neutrosophic set, and En_I is employed to evaluate the indetermination distribution.

D. λ-mean Operation

As mentioned, the value of $I(i,j)$ is employed to measure the indeterminate degree of $P_{NS}(i,j)$. To make I to be correlated with T and F, changes in T and F influence the distribution of elements in I and the entropy of I. In the gray level domain, a λ-mean operation for image X can be defined as

$$\overline{X}(i,j) = \frac{1}{w \times w} \sum_{m=i-w/2}^{i+w/2} \sum_{n=j-w/2}^{j+w/2} X(m,n), \tag{17}$$

where w is the local window size.

A λ-mean operation for P_{NS} is defined as:

$$\overline{P}_{NS}(\lambda) = P(\overline{T}(\lambda), \overline{I}(\lambda), \overline{F}(\lambda)), \tag{18}$$

with

$$\overline{T}_\lambda(i,j) = \frac{1}{w \times w} \sum_{m=i-w/2}^{i+w/2} \sum_{n=j-w/2}^{j+w/2} T(m,n), \tag{19}$$

$$\overline{F}_\lambda(i,j) = \frac{1}{w \times w} \sum_{m=i-w/2}^{i+w/2} \sum_{n=j-w/2}^{j+w/2} F(m,n), \tag{20}$$

and

$$\overline{I}_\lambda(i,j) = 1 - \frac{\overline{Ho}(i,j) - \overline{Ho}_{\min}}{\overline{Ho}_{\max} - \overline{Ho}_{\min}}, \tag{21}$$

where $\overline{Ho}(i,j)$ is the homogeneity value of $\overline{T}(\lambda)$ at (i,j).

After the true subset T is handled using the λ-mean operation, noise in T is removed and T becomes more homogeneous, and consequently more suitable to segment T precisely even using a simple thresholding method.

IV. Proposed Approach

Fig. 2 summarises the various steps of the proposed algorithm. Captured manuscript images RGB images P_{RGB} (of different sizes and stored in simple bitmap format) are converted to gray level images P_G using the NTSC standard method.

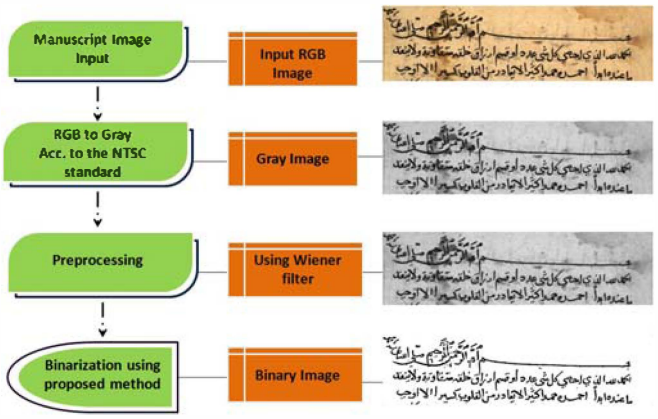

Fig. 2. Overview of the proposed document analysis approach.

A. Pre-processing

Since historical document images are usually of low quality, a pre-processing stage is essential in order to eliminate noise areas, smooth the background texture and better highlight the contrast between background and text areas [14]. The use of a low-pass Wiener filter has proved efficient in this context [15], with the window size of the filter selected according to the minimum character line width [14]. In our method, the window size is selected as 3×3. The filtered gray level image P_{gw} can the be binarized in the next stage.

B. Binarization

This is the stage where we employ a modified version of the neutrosophic thresholding method of [5]. The filtered gray image P_{gw} is transformed into the neutrosophic domain, giving P_{NS} as described in Section III-B. Then, the λ-mean operation is employed to reduce the indetermination degree of the image P_{NS} which is evaluated by the entropy of the indeterminate subset. Thus, the image becomes more uniform and homogenous and more suitable to be thresholded. We then use the adaptive thresholding method of [10] to obtain the binary image.

C. Post-processing

Adaptive thresholding produces a noisy binary image. Consequently, a post-processing stage is required to remove the noise. For this, a median filter is used with a window size of 3×3" to enhance the binary image.

D. Summary

Our proposed document analysis algorithm is summarized in Algorithm 1.

Algorithm 1 Proposed document analysis algorithm

1: Read in RGB image $P_{RGB}(x, y)$.
2: Convert to gray image $P_g(x, y)$ using NTSC standard.
3: Apply Wiener adaptive filter as pre-processing step to obtain $P_{gw}(x, y)$.
4: Transform $P_{gw}(x, y)$ into neutrosophic domain to obtain $P_{NS}(x, y) = \{T(x, y), I(x, y), F(x, y)\}$ according to the entropy of $P_{gw}(x, y)$ and its mean.
5: Measure the entropies of the three subsets T, I, and F.
6: Apply a λ-mean operation on $P_{NS}(x, y)$ to decrease its indetermination.
7: Segment the true subset T using an adaptive thresholding technique.
8: Apply a median filter to remove noise as post-processing step.

V. Experimental Results

In this section, the proposed method is evaluated and compared with other binarization methods from the literature.

Our dataset contains samples collected from both the database of [2] and from the electronic Arabic manuscripts of [16]. In our evaluation, we focus on images that have several degradations such as multi-colored text lines, stains in the background, degraded characters and marks, and character diacritics.

For evaluation of our results, ground truth images are generated using a similar method to the recent work of [17].

We compare the performance of our proposed algorithm with those obtained by other binarization methods, namely Otsu's [8], Niblack's [9], Sauvola's [10], and Guo's [5]. The latter uses a similar neutrosophic approach but the binary image is obtained using global thresholding.

Fig. 3 demonstrates the results obtained for an example image of the dataset.

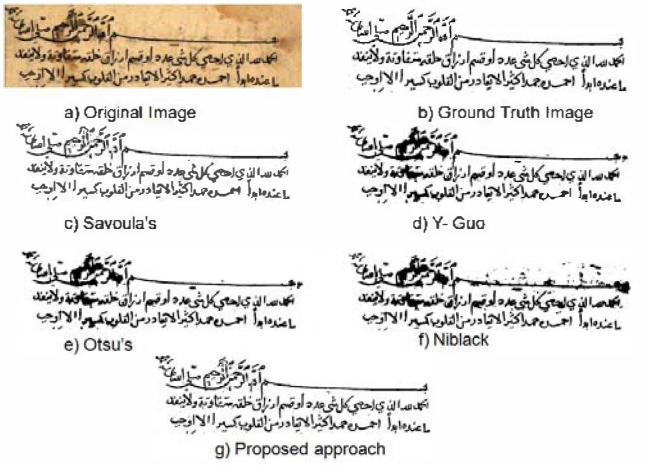

Fig. 3. Example results obtained by the different binarization algorithms.

As can be seen, our algorithm outperforms the other approaches in terms of preservation of meaningful textual information. The other methods either fail to segment the foreground text, especially in the region of stains (Fig. 3(d), (e), and (f)), or segment foreground from background but add excessive noise (Fig. 3(f)). On the other hand, our final binary image suffers from some stroke-like pattern noise (SPN), which is due to the used Arabic manuscripts including diacritics. SPN [18], [19] is similar to diacritics, and hence its presence near textual components may change the meaning of a word.

For the images of our dataset, the average processing time for our method was about 0.4 seconds for images of an average size of 500×500 (on an Intel Core i3-2310- CPU@2.10 GHZ with 3GB RAM 3.00, running Windows 7 and the algorithm implemented using Matlab R2009a).

Several methods have been presented for the evaluation of document image binarization techniques, and can be classified into three main categories [1]. Evaluation can be performed by visual inspection by one or more human evaluators. Here, symbols that are broken or blurred, and loss of objects as well as noise in background and foreground are used as visual evaluation criteria. Clearly, this approach is time consuming and subjective. Evaluation based on optical character recognition (OCR) performance, applies OCR on the result image and uses the obtained character or word recognition accuracy. Applying OCR as an evaluation criterion is not possible for our experiments due to the lack of an efficient commercial software for recognizing handwritten Arabic manuscript writing [20]. Finally, direct evaluation of the binarization can be performed by taking into account the pixel-to-pixel correspondence between the ground truth and the binarized image. For this, several measures can be employed [21]–[23], of which we use the following in our tests:

- F-measure [24], defined as

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{22}$$

where

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{23}$$

and

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{24}$$

Here, TP denotes true positives, TN true negatives, FP false positive, and FN false negatives, respectively.

- F_β-measure, defined as the weighted harmonic mean between precision and recall

$$F_\beta = \frac{(1 + \beta)^2 \times TP}{(1 + \beta)^2 \times TP + \beta^2 \times FN + FP}. \tag{25}$$

- PSNR, a similarity measure between two images, defined as [21], [25]

$$\text{PSNR} = 10 \log_{10} \left(\frac{R^2}{MSE} \right), \tag{26}$$

with

$$\text{MSE} = \frac{\sum [I_1(m, n) - I_2(m, n)]^2}{M \times N}, \tag{27}$$

where I_1 and I_2 are the two images, and M and N the image dimensions.

- Negative metric rate (NRM), which is based on pixelwise mismatches between the ground truth and the binarized image [26], and combines the false negative rate N_{FN} and the false positive rate N_{FP} as

$$\text{NRM} = \frac{\frac{N_{FN}}{N_{FN}+N_{TP}} + \frac{N_{FP}}{N_{FP}+N_{TN}}}{2}. \quad (28)$$

A better binarization quality is characterised by a lower NRM value.

- Distance reciprocal distortion (DRD) metric, defined as

$$\text{DRD} = \frac{\sum_{K=1}^{S} \text{DRD}_k}{\text{NUBN}}, \quad (29)$$

where DRD_k is the distortion of the k-th flipped pixel and is calculated using a 5×5 normalized weight matrix WN_m [29]. DRD_k equals the weighted sum of pixels in the 5×5 block of the ground truth GT that differ from the centered k-th flipped pixel at (x, y) in the binarization result B

$$\text{DRD}_k = \sum_{i=-2}^{2} \sum_{j=-2}^{2} |GT_k(i,j) - B_k(i,j)| \times WN_m(i,j). \quad (30)$$

NUBN is the number of the non-uniform (not all black or white pixels) 8×8 blocks in the GT image [27].

Table I summarizes the results (averages over all dataset images) of the various binarization algorithms. As is clear from the obtained measures, our proposed approach provides the best results with respect to all performance metrics.

TABLE I
PERFORMANCE COMPARISON OF ALL BINARIZATION METHODS

method	F	F_β	PSNR	NRM	DRD
Niblack	84.47	88.63	13.51	4.05	0.072
Otsu	90.36	90.72	15.88	2.65	0.037
Sauvola	94.76	95.34	20.96	1.20	0.033
Guo	90.44	90.89	16.53	2.60	0.037
Proposed	95.59	95.87	23.57	1.01	0.028

VI. CONCLUSIONS

In this paper, we have proposed a hybrid thresholding technique for degraded images of historical Arabic manuscripts. Our method combines a neutrosophic set approach with an adaptive thresholdinging method. Experiment results show that the proposed method provides good binarization performance for complex images that contain various challenges including stains, ink seeping, and characters written by different ink colors. Future work will focus on removal of stroke-like pattern noise to further improve the binarization results.

ACKNOWLEDGMENTS

Thanks to Dr. Y. Guo, St. Thomas Univ., USA, for his kind help and useful discussions.

REFERENCES

[1] K. Nitrogiannis, N. Gatos, and I. Pratikakis, "Performance Evaluation methodology for historical document image binarization", IEEE Trans Image Processing, Vol. 22(2), 2013.
[2] K.M. Amin, M.A. Ahmad, and A. Ali, "A Novel Binarization Algorithm for Historical Arabic Manuscripts using Wavelet Denoising", Int. J. of Computing and Inf. Sciences, Vol. 13(1), 2013.
[3] P. Stathis, E. Kavallieratou, and N. Papamarkos, "An Evaluation Technique for Binarization Algorithms", Journal of Universal Computer Science, Vol. 14(18), pp. 3011-3030, 2008.
[4] J. Mohan, V. Krishnaveni, Y. Guo, "A New Neutrosophic Approach of Wiener Filtering for MRI Denoising", Measurement Science Review, Vol. 13(4), pp. 177-168, 2013.
[5] Y. Cheng, Heng-Da, and Yanhui Guo, "A new neutrosophic approach to image thresholding", New Mathematics and Natural Computation, Vol. 4(3), pp. 291-308, 2008.
[6] M-L. Feng and Y-P. Tan, "Adaptive binarization method for document image analysis", IEEE Int. Conf. on Multimedia and Expo, Vol. 1, pp. 339-342, 2004.
[7] S. Faisal, K. Daniel, B. Thomas, "Efficient implementation of local adaptive thresholding techniques using integral images", Electronic Imaging, pp. 681510-681510, 2008.
[8] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms ", IEEE Transactions on Systems, Man and Cybernetics, Vol. 9(1), pp. 62-66, 1979.
[9] W.Niblack, "An Introduction to Digital Image Processing", Prentice Hall, Englewood Cliffs, 1986.
[10] J. Sauvola, M. Pietikainen, "Adaptive document image binarization ", Pattern Recognition, Vol. 33(2), pp. 225-236, 2000.
[11] K. Khurram, "Comparison of Niblack inspired Binarization methods for ancient documents", IST/SPIE Electronic Imaging, pp. 72470U-72470U, 2009.
[12] Y. Guo, and H.D. Chenga, "New neutrosophic approach to image segmentation", Advances in Multimedia, Vol. 42(5), pp. 587-595, 2009.
[13] M. Zhang, "Novel Approaches to Image Segmentation Based on Neutrosophic Logic", Doctoral Dissertation, Utah State University, 2010.
[14] B. Gatos, I. Pratikakis, and S. Perantonis, "An Adaptive Binarization Technique for Low Quality Historical Documents", DAS 2004, LNCS 3163, pp. 102-113, 2004.
[15] A. Jain, "Fundamentals of Digital Image Processing", Pr. Hall, 1989.
[16] http://www.wqf.me.com
[17] H. Nafchi, S. Ayatollahi, R. Farrahi, and M. Cheriet, "An efficient ground truthing tool for binarization of historical manuscripts", 12th Int. Conf. on Document Analysis and Recognition, pp. 807-811, 2013.
[18] A. Farahmand, A. Sarrafzadeh, and J. Shanbehzadeh, "Document Image Noises and Removal Methods", IMECS, pp. 436-440, 2013.
[19] M. Agrawal and D. Doermann, "Stroke-like Pattern Noise Removal in Binary Document Images", 11th Int. Conf. on Document Analysis and Recognition, pp. 17-21, 2011.
[20] F. Smarandache, "A Unifying Field in Logics Neutrosophic Logic. Neutrosophy, Neutrosophic Set, Neutrosophic Probability", 3rd Ed., American Research Press, 2003.
[21] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 document image binarization contest (DIBCO 2009)", Int. Conf. on Document Analysis and Recognition, pp. 1375-1382, 2009.
[22] H. Lu, A. C. Kot, Y. Q. Shi, "Distance-Reciprocal Distortion Measure for Binary Document Images", IEEE Signal Processing Letters, Vol. 11(2), pp. 228-231, 2004.
[23] B. Su, S. Lu, and C. Lim, "A Robust Document Image Binarization Technique for Degraded Document Images", IEEE Trans. on Image Processing, Vol. 22(4), 2013.
[24] N. Chinchor, "MUC-4 evaluation metrics", 4th Message Understanding Conference, pp. 22-29, 1992.
[25] H. Lu, A. C. Kot, and Y.Q. Shi, "Distance-reciprocal distortion measure for binary document images", IEEE Signal Processing Letters, Vol. 11, No. 2, pp. 228-231, 2004.
[26] J. Aguilera, H. Wildenauer, M. Kampel, M. Borg, D. Thirde, and J. Ferryman, "Evaluation of motion segmentation quality for aircraft activity surveillance", 2nd Joint IEEE Int.Workshop on Visual Surveillance and Performance Evaluation of Tracking & Surveillance, pp. 293-300, 2005.
[27] I. Pratikakis, B. Gatos and K. Ntirogiannis, "ICDAR 2011 Document Image Binarization Contest", International Conference on Document Analysis and Recognition, pp. 1506-1510, 2011.