Binarization of Degraded Historical Document Images

Zineb Hadjadj
Université de Blida
Blida, Algérie
hadjadj_zineb@yahoo.fr

Mohamed Cheriet
École de Technologie Supérieure
Montréal, Canada
mohamed.cheriet@etsmtl.ca

Abdelkrim Meziane
Centre de Recherche sur l'Information Scientifique et Technique (CERIST)
Alger, Algérie
a.meziane@mail.cerist.dz

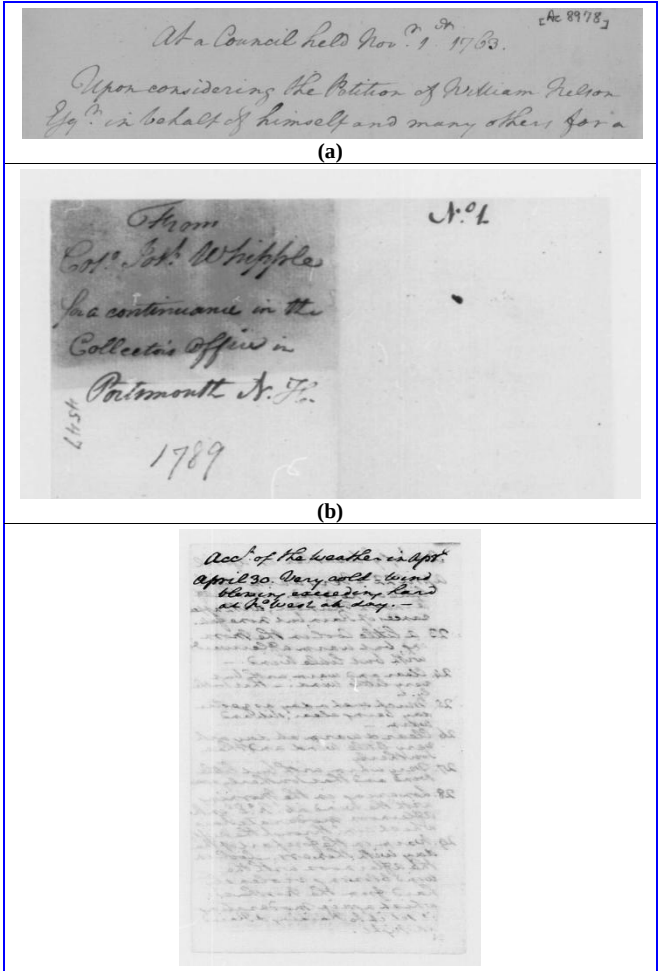
***Abstract*—Document images often suffer from different types of degradation that renders the document image binarization a challenging task. In this paper, a new binarization algorithm for degraded document images is presented. The method is based on active contours evolving according to intrinsic geometric measures of the document image; Niblack's thresholding is also used to control the active contours propagation. The validity of the proposed method is demonstrated on both recent and historical document images including different types of degradations, the results are compared with a number of known techniques in the literature.**

Keywords—Document image; binarization; active contours; level sets method; Niblack's thresholding.

I. INTRODUCTION

In document images binarization, two distinct regions are defined as characters (foreground) and backgrounds. Characters are objects that we desire to extract, recognize, and represent. The remaining regions are backgrounds of these objects.

Though document image binarization has been studied for many years, the thresholding of degraded document images is still an unsolved problem. This can be explained by the difficulty in modeling different types of document degradation such as uneven illumination, image contrast variation, bleeding-through, and smear that exist within many document images as illustrated in Fig. 1.



(a) **(b)** **(c)**

Figure 1. Example document images in DIBCO's [(1)] dataset that illustrate document degradation including: (a) faint characters; (b) uneven background; (c) bleed-through

Many document image binarization methods have been proposed which are usually classified in two main categories, namely global thresholding and local adaptive thresholding techniques.

[(1)]Available online: http://users.iit.demokritos.gr/~bgat/DIBCO2009/benchmark/.

Global thresholding methods use a single intensity threshold value for all the image. This value is calculated based on some heuristics or statistics of some global image attributes to classify image pixels into foreground (text) or background (non-text) pixels. Otsu's algorithm (Otsu, 1979) [1] is the most popular global thresholding technique. Moreover, there are many popular thresholding techniques such as (Kapur et al., 1985) [2], and (Kittler and Illingworth, 1986) [3]. The main issue with global methods is that they cannot adapt well to uneven illumination and noise, hence do not perform well on low quality document images.

Local thresholding methods compute a threshold for each pixel in the image on the basis of the content in its neighborhood. Niblack's method (Niblack, 1986) [4] can be considered as the first local threshold method. It has the advantage to detect the text but it introduces a lot of background noise. (Sauvola and Pietikinen, 2000) [5] modified the Niblack threshold to decrease the background noise but the text detection rate is also decreased while bleed-through still remains in most cases. (Bernsen, 1986) [6], (Wolf and Jolion, 2003) [7] and (Feng and Tan, 2004) [8] also tried to improve the results of Niblack's method.

The global thresholding method of Otsu and the local methods of Niblack and Sauvola are widely incorporated in binarization methods that followed (e.g. (Kim et al., 2002) [9]; (Gatos et al., 2006) [10]; (Lu et al., 2010) [11]; (Ntirogiannis et al., 2012) [12]).

- ✓ (Kim et al., 2002) [9] consider the original image as a 3D terrain, with valleys and mountains corresponding to text and background regions, on which water was poured to fill the valleys that represented the textual components. The final binarization result was produced by applying Otsu to the compensated image, i.e. the difference between the original image and the water-filled image.

Certain binarization methods have incorporated background estimation and normalization steps:

- ✓ (Gatos et al., 2006) [10] estimate the background surface based on the binary document image generated by Sauvola's thresholding method, and use wiener filter for pre-processing. The final threshold was based on the difference between the estimated background and the preprocessed image while post-processing enhanced the final result.
- ✓ (Moghaddam et al., 2009) [13] estimate the document background surface through an adaptive and iterative image averaging procedure.
- ✓ (Lu et al., 2010) [11] estimate the document background surface through an iterative polynomial smoothing procedure. Then, the original image was normalized and Otsu was performed to detect the text stroke edges. Furthermore, their local threshold formula was based on the local number of the detected text stroke edges and their mean intensity.
- ✓ In (Ntirogiannis et al., 2012) [12] method, background estimation is applied along with image normalization based on background compensation. The normalized image is used in the global and local binarization steps that follow. Finally, the two binarization outputs are combined at connected component level to produce the final binarization result.

Certain binarization methods make use of the image edges that can usually be detected around the text stroke boundary:

- ✓ (Chen et al., 2008) [14] propose to first detect and close image edges and then obtain a primary binary document images based on the determined edge information
- ✓ (Moghaddam et al., 2009) [13] make use of the edge profile to locate the text region and accordingly estimate the local image threshold.
- ✓ (Su et al., 2010) [15] locate the text stroke edges by using an image contrast that is evaluated based on the local maximum and minimum.
- ✓ In (Rivest-Hénault et al., 2011) [16] method, local linear models are used to estimate both the expected stroke and the background pixel intensities. This information is then used as the main driving force in the propagation of an active contour. In addition, a curvature-based force is used to control the viscosity of the contour and leads to more natural-looking results.

In our proposed binarization algorithm, Niblack's binarization method is used to control the driving forces (the contour evolution) on the uneven background degraded document images.

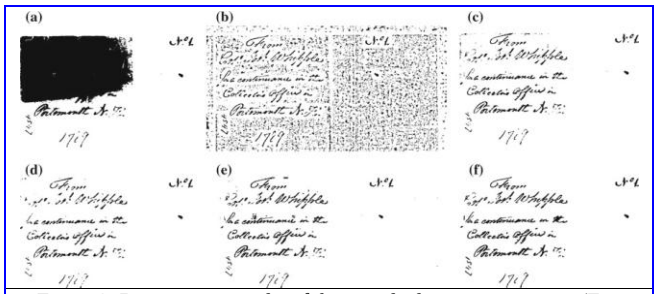


Figure 2. Binarization results of the sample document image in (Fig. 1.b) by using :
Otsu's method [1] in (a), Niblack's method [4] in (b), Sauvola's method [5] in (c), Gatos's method in (d), Su's method [15] in (e), and Lu's method [11] in (f)

The rest of the paper is structured as follows. In Section 2, the proposed binarization method based on an active contour model is presented. Simulation results of the proposed method applied to real documents and the performance comparison with other binarization methods are shown in Section 3. Finally conclusions are given in Section 4.

II. PROPOSED METHOD

The section describes the proposed document image binarization technique. In particular, we divide this section into three subsections.

A. *Active Contours*

In this subsection, we focus on boundary detection of objects (text) by a dynamic model known as the 'Ron Kimmel's geodesic active contour' introduced in [17].

Geodesic active contours were introduced as a geometric alternative for 'snakes'. Snakes [18] are deformable models that are based on minimizing an energy along a curve. The curve, or snake, deforms its shape so as to minimize an 'internal' and 'external' energy along its boundary. The internal part causes the boundary curve to become smooth, while the external part leads the curve towards the edges of the object in the image.

In [19] [20], a geometric alternative for the snake model was introduced, in which an evolving curve was formulated by the Osher-Sethian level set method [21].

The geodesic active contour model was born latter. It is both a geometric model as well as energy functional minimization.

B. *Ron Kimmel's Geodesic Active Contour Model*

Ron Kimmel's model [17] is a geodesic active contour model, it's based on deforming an initial contour C_0 towards the boundary of the object to be detected. In this model, we search for a contour (curve) (Fig. 3), $C: [0, L] \to \mathcal{R}^2$, given in a parametric form $C(s) = \{x(s), y(s)\}$, where s is an arclength parameter, $I: [0, a] \times [0, b] \to \mathcal{R}^+$ is a given image in which we want to detect the objects boundaries.

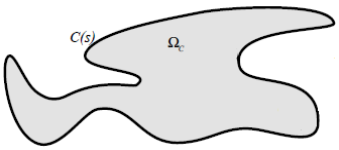


Figure 3. A closed curve C and Ω_C the area inside the curve.

The model, in Eq.(**1**), incorporates the alignment force as part of other driving forces of an active contour, together with the geodesic active contour model for regularization, and the minimal variance criterion suggested by Chan and Vese [22].

$$E(C, c_1, c_2) = E_{AR}(C) - \alpha E_{GAC}(C) - \beta E_{MV}(C, c_1, c_2) \qquad (\mathbf{1})$$

Where, the two constants, c_1 and c_2, get the mean intensities in the interior (inside) and the exterior (outside) the contour C, respectively. α and β are constants.

- The robust alignment term is given by the functional

$$E_{AR}(C) = \oint_C |\langle \nabla I, \vec{n} \rangle| \, ds \qquad (\mathbf{2})$$

The gradient ∇I direction is a good estimator for the orientation of the edge contour C (Fig. 4). The alignment term gets high values if the curve normal $\vec{n}$ aligns with the image gradient direction, so, our goal would be to find curves that maximize this alignment functional.

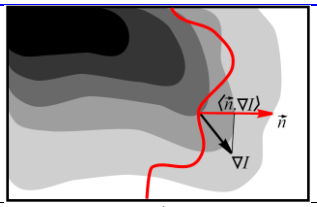


Figure 4. The curve C, its normal $\vec{n}$ at a specific point, and the image gradient ∇I at that point.

- The geodesic active contour term is defined by the functional

$$E_{GAC}(C) = \oint_C g(C(s)) ds \qquad (\mathbf{3})$$

It is an integration of an inverse edge indicator function, like $g(x, y) = 1/(1 + |\nabla I|^2)$, along the contour. The search, in this case, would be for a curve along which the inverse edge indicator gets the smallest possible values. That is, we would like to find the curve C that minimizes this functional.

- Chen and Vese proposed a minimal variance criterion given by [22]

$$E_{MV}(C, c_1, c_2) = \frac{1}{2} \iint_{\Omega_C} (I(x, y) - c_1)^2 dxdy + \frac{1}{2} \iint_{\Omega/\Omega_C} (I(x, y) - c_2)^2 dxdy \qquad (\mathbf{4})$$

This functional serves to find the best separating curve. The optimal curve would best separate the interior and exterior with respect to their relative average values.

By using Eq.(2), Eq.(3), and Eq.(4), the Eq.(1) can be written as follows:

$$\begin{aligned} E(C, c_1, c_2) &= E_{AR}(C) - \alpha E_{GAC}(C) - \beta E_{MV}(C, c_1, c_2) \\ &= \oint_C |\langle \nabla I, \vec{n} \rangle| \, ds - \alpha \oint_C g(C(s)) ds \\ &\quad - \beta \frac{1}{2} \left(\iint_{\Omega_C} (I - c_1)^2 dxdy + \iint_{\Omega/\Omega_C} (I - c_2)^2 dxdy \right) \end{aligned} \qquad (\mathbf{5})$$

The curve C evolution is given by computing the first variation of each functional [17]

$$C_t = \left[sign(\langle \nabla I, \vec{n} \rangle) \Delta I + \alpha (g(x, y) k - \langle \nabla g, \vec{n} \rangle) + \beta (c_2 - c_1) \left(I - \frac{c_1 + c_2}{2} \right) \right] \vec{n} \qquad (\mathbf{6})$$

$$c_1 = \frac{1}{|\Omega_C|} \iint_{\Omega_C} I(x, y) dxdy \qquad c_2 = \frac{1}{|\Omega/\Omega_C|} \iint_{\Omega/\Omega_C} I(x, y) dxdy$$

Where $|\Omega_C|$ denotes the area of the region Ω_C.

We embed the curve C in a higher dimentional $\Phi(x,y)$ function (Fig. 5), which implicitly represents the curve C as a zero set, i.e. $C = \{(x,y)/\Phi(x,y) = 0\}$. Level set method [21] can be employed to implement the curve propagation toward its optimal location.

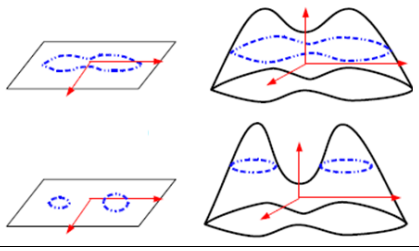


Figure 5. Arbitrary contours (*left*) and the associated level set functions (*right*)

The level set formulation of the curve evolution equation is [17]

$$\Phi_t = \left[sign(\langle \nabla I, \nabla \Phi \rangle)\Delta I + \alpha div\left(g(x,y)\frac{\nabla \Phi}{|\nabla \Phi|}\right) + \beta(c_2 - c_1)\left(I - \frac{c_1 + c_2}{2}\right)\right]|\nabla \Phi| \qquad (7)$$

Next, we approximate the time derivative using the approximation

$$\Phi_t \approx \frac{\Phi^{n+1} - \Phi^n}{\Delta t} \qquad (8)$$

That yields the explicit scheme

$$\Phi^{n+1} = \Phi^n + \Delta t\left(\left[sign(\langle \nabla I, \nabla \Phi \rangle)\Delta I + \alpha(g(x,y)k) + \beta(c_2 - c_1)\left(I - \frac{c_1 + c_2}{2}\right)\right]|\nabla \Phi|\right) \qquad (9)$$

Where $k = div\frac{\nabla \Phi}{|\nabla \Phi|}$

The active contour model described in this subsection was implemented in Matlab. The section 3 describes the model implementation.

C. *Niblack's Thresholding*

In the proposed technique, Niblack thresholding [4] is used to solve, the active contours propagation in the degraded regions around the text, problem (Fig. 6 (b)).

Niblack is a local thresholding algorithm that adapts the threshold according to the local mean and the local standard deviation over a specific window size around each pixel location. The local threshold at any pixel (i,j) is calculated as:

$$T(i,j) = m(i,j) + k * s(i,j) \qquad (10)$$

Where $m(i,j)$ and $s(i,j)$ are the local sample mean and variance, respectively. The size of the local region (window) is dependent upon the application.

The value of the weight 'k' is used to control and adjust the effect of standard deviation due to objects features.

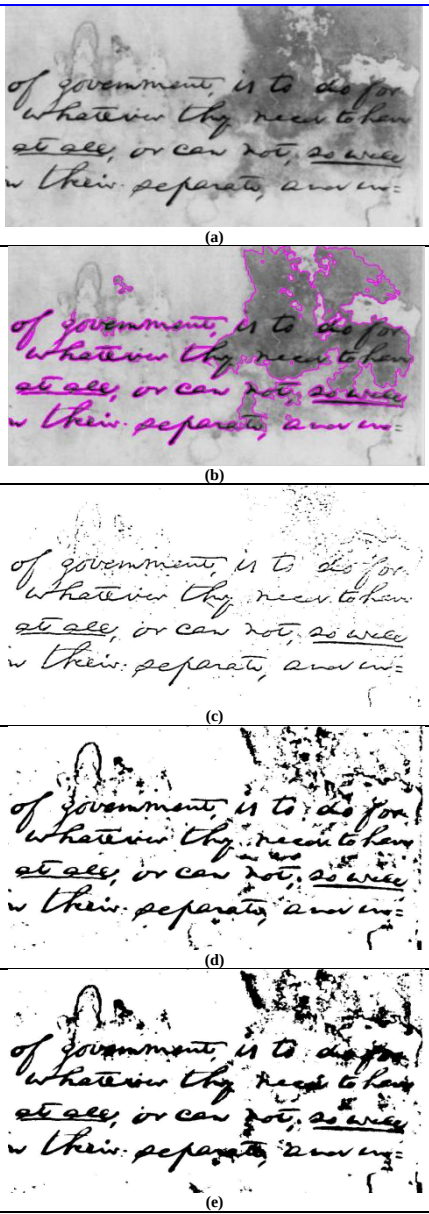


Figure 6. Image H04 from the DIBCO'09 test dataset
(a) input image, (b) final active contour obtained using Ron Kimmel's deformable model, (c) mask obtained using Niblack's method (10x10 neighbourhood), (d) 50x50 neighbourhood, (e) 100x100 neighbourhood

To solve the active contours propagation problem (Fig. 6 (b)), we applied to the input document image a mask which is obtained using the Niblack thresholding before using Ron Kimmel's deformable model.

The results of binarization using Niblack's algorithm for different window sizes (10x10, 50x50 & 100x100) are shown in Fig. 6 (c)(d)(e).

The Niblack's thresholding creates separated small areas in the degraded regions around the text, so it stops the active contours propagation.

By looking at results of different window sizes, we have observed that by increasing window sizes, we get the unnecessary black pixels eliminated from the image

background in a better way while filling out characters and vice versa. In this way, we have found window size of 50x50 to be appropriate for this kind of images as can be observed from the resulting images as shown.

The results of this proposed method are presented and discussed in the following section.

III. Experimental results and discussion

The proposed method has been tested over the handwritten images of the dataset that is used in the Document Image Binarization Contest (DIBCO'09). The dataset is composed of a number of representative document images that suffer from different types of document degradation. We compare our method with other well-known binarization methods including Otsu's global thresholding method [1] and Niblack's and Sauvola's adaptive thresholding methods [4], [5].

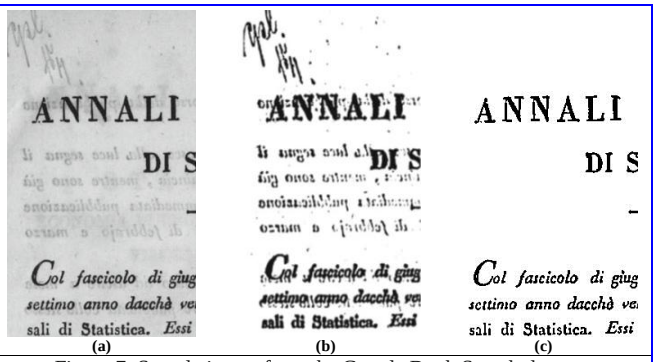


Figure 7. Sample image from the Google Book Search dataset
(a) input image, (b) binarization results obtained using Sauvola's method and (c) the proposed method

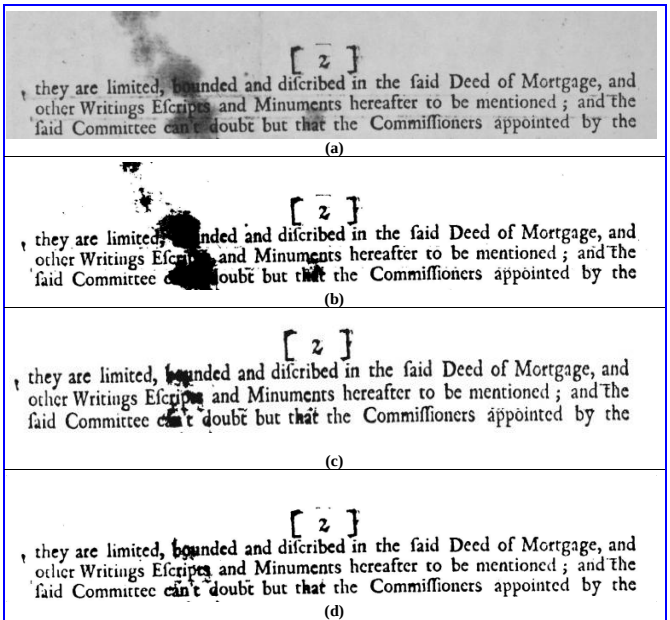


Figure 8. Image P04 from the DIBCO'09 test dataset
(a) input image, (b) binarization result obtained using Otsu's method, (c) Rivest-Hénault algorithm and (d) the proposed method

This image is rather noisy and part of the text boundary is weak. Our method successfully extracts the text in this image.

Fig. 7, 8 and 9 further show three document binarization examples. As shown in the five figures, our proposed method extracts the text properly from document images that suffer from different types of document degradation. On the other hand, the other methods often produce a certain amount of noise due to the variation within the document background.

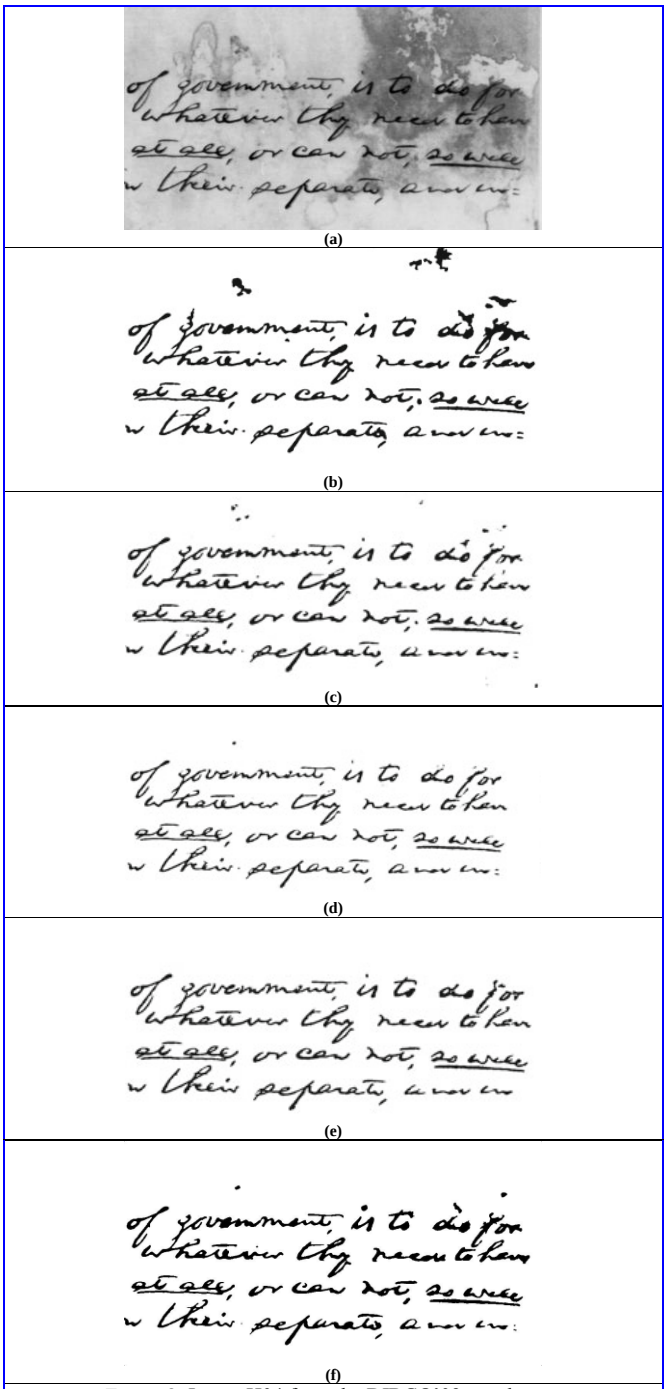


Figure 9. Image H04 from the DIBCO'09 test dataset
(a) input image, (b) binarization result obtained using Rivest-Hénault algorithm, (c) Gatos's method, (d) Su's method, (e) Lu's method and (f) the proposed method

The evaluation measures are adapted from the DIBCO report [23] including

- ✓ $F-measure$
- ✓ peak signal-to-noise ratio ($PSNR$)
- ✓ negative rate metric (NRM)
- ✓ misclassification penalty metric (MPM)

Therefore, we chose to concentrate on the $F-measure$, because it iswell-widely accepted and simple, and so is easy to interpret.

The $F-measure$ is defined as follows:

$$F-measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (11)$$

Where,

$$Recall = \frac{TP}{TP+FN} \qquad Precision = \frac{TP}{TP+FP}$$

and TP, FP, and FN representing the number of true positive, false positive and false negative values respectively.

TABLE I. AVERAGE $F-measure$ FOR THE TEN IMAGES OF THE TEST DATASET FROM THE DIBCO'09 CONTEST

Method	$F-measure$
Otsu's	78.72%
Niblack's	55.82%
Sauvola's	85.41%
Gatos's	85.25%
Su's	91.06%
Lu and Tan algorithm[(1)] [23]	91.24%
Fabrizio and Marcotegui algorithm[(2)] [23]	90.06%
Rivest-Hénault algorithm[(3)] [23]	89,34%
Proposed	90.54%

The performances of Otsu's, Niblack's, Sauvola's Gatos's and Su's methods are reported from [11]. Other results are available in [23]

Experiment results are shown in TABLE I. Compared with other methods, our proposed method performs better than some other in term of the $F-Measure$. This means that the proposed method produces a higher precision and preserves the text stroke contour better.

The proposed document binarization method has a few limitations, the proposed method can deal with the ink-bleeding as illustrated in Fig. 7 when the back-side text strokes are much weaker compared with the front-side text. But when the back-side text strokes are as dark as or even darker than the front-side text strokes, the proposed method cannot classify the two types of character strokes correctly. We will study this issue in our future works.

IV. CONCLUSION AND FUTURE PROSPECTS

Document image binarization is an important basic task needed in most document analysis systems. The quality of binarization result affects to subsequent processing by offering pre-segmented objects in precise form (object/non-object). In this paper we proposed a new simple technique to document image binarization, using active contours. Our techniques based on deforming an initial contour C_0, extracted before by using Canny edge detector, towards the boundary of the object to be detected (text). And, to solve the problem of the active contours propagation in the degraded regions around the text, it applies to the input document image a mask which is obtained using Niblack's thresholding.

The proposed method has been tested on the dataset that is used in the recent DIBCO contests. Experiments show that the proposed method outperforms several other document binarization methods in term of the $F-measure$.

As a prospect for the future, improvement to the driving forces (energy terms) to make them more adaptable to the variations on the degraded document images will be considered.

V. REFERENCES

[1] N. Otsu, "A thresholding selection method from gray-level histogram," IEEE Transactions on Systems, Man, and Cybernetics, vol.9, no. 1, pp.62–66, January 1979.
[2] J.N. Kapur, P.K. Sahoo, and A.K.C. Wong, "A new method forgray-level picture thresholding using the entropy of the histogram," Graphics and Image Processing, vol. 29, pp. 273-285, 1985.
[3] J. Kittler and J. Illingworth, "Minimum error thresholding," Pattern Recognition, vol. 19, no. 1, pp. 41-47, 1986.
[4] W. Niblack, "An introduction to digital image processing," Prentice Hall, Englewood Cliffs, July 1986.
[5] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," Pattern Recognition, vol. 33, no. 2, pp. 225-236, 2000.
[6] J. Bernsen, "Dynamic thresholding of grey-level images," in: Proceedings of the Eighth International Conference on Pattern Recognition, Paris, France, pp. 1251–1255, October 1986.
[7] C. Wolf and J.M. Jolion, "Extraction and Recognition of Artificial Text in Multimedia Documents," Pattern Analysis and Applications, vol. 6, no. 4, pp. 309-326, 2003.
[8] M.L. Feng and Y.P. Tan, "Contrast adaptive binarization of low quality document images," IEICE Electronics Express, vol. 1, no. 16, pp. 501-506, November 2004.
[9] I.K. Kim, D.W. Jung, and R.H. Park, "Document image binarization based on topographic analysis using a water flow model," Pattern Recognition, vol. 35, no. 1, pp. 265–277, 2002.
[10] B. Gatos, I. Pratikakis, and S.J. Perantonis, "Adaptive degraded document image binarization," Pattern Recognition, vol. 39, no. 3, pp. 317–327, 2006.
[11] S. Lu, B. Su, and C.L. Tan, " Document image binarization using background estimation and stroke edges," International Journal on Document Analysis and Recognition, vol. 13, no. 4, pp. 303–314, October 2010.
[12] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "A combined approach for the binarization of handwritten document images," Pattern Recognition Letters - Special Issue on Frontiers in Handwriting Processing, DOI 10.1016/j.patrec.2012.09.026, In Press, October 2012.
[13] R.F. Moghaddam and M. Cheriet, "RSLDI: restoration of singlesided low-quality document images," Pattern Recognition, vol. 42, no. 12, pp. 3355–3364, December 2009.
[14] Q. Chen, Q. Sun, P.A. Heng, and D. Xia, "A double-threshold image binarizationmethod based on edge detector," Pattern Recognition, vol. 41, no. 4, pp. 1254–1267, April 2008.
[15] B. Su, S. Lu, C.L. Tan, "Binarization of historical handwritten document images using local maximum and minimum filter," International Workshop on Document Analysis Systems, pp. 159-165, June 2010.
[16] D. Rivest-Hénault, R. Farrahi Moghaddam, M. Cheriet, "A local linear level set method for the binarization of degraded historical document images," International Journal on Document Analysis and Recognition, vol. 15, pp. 101-124, April 2011.
[17] R. Kimmel, "Fast Edge Integration," Scale Space and PDE methods in image analysis and processing, June 2008.
[18] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active Contour Models," International Journal of Computer Vision, vol. 1, no. 4, pp. 321-331, 1988.
[19] V. Caselles, F. Catte, T. Coll, and F. Dibos, "A geometric model for active contours," Numerische Mathematik, vol. 66, pp. 1-31, 1993.
[20] R. Malladi, J.A. Sethian, and B.C. Vemuri, "Shape modeling with front propagation: A level set approach," IEEE Trans. on PAMI, vol. 17, pp. 158-175, 1995.
[21] S. J. Osher and J. A. Sethian, "Fronts propagation with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations," Journal of Computational Physics, vol. 79, pp. 12-49, 1988.
[22] T.F. Chan, L.A. Vese, "Active Contour Without Edges," IEEE transactions on image processing, vol. 10, no. 2, pp. 266-277, February 2001.
[23] J.F. Canny, "A computational approach to edge detection," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 8, no. 6, pp. 679-698, November 1986.
[24] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 document image binarization contest (DIBCO 2009)," In International Conference on Document Analysis and Recognition, pp. 1375–1382, July 2009.

(1) It placed 1st in DIBCO'09, (2) It placed 2ed in DIBCO'09, (3) It placed 3rd in DIBCO'09