2011 International Conference on Document Analysis and Recognition

Enhanced Active Contour Method for Locating Text

Yaakov Navon, Vladimir Kluzner
Document Processing and Management Group
IBM Research – Haifa (previously)
Haifa, Israel
{navonyaakov, vkluzner}@gmail.com

Boaz Ophir
Computer Science Department
Technion – Israel Institute of Technology
Haifa, Israel
boazo@cs.technion.ac.il

Abstract—**Document analysis of images photographed by camera-equipped mobile phones is a growing challenge. These photos are often poor-quality compound images, composed of various objects and text; this makes automatic analysis complicated, thereby limiting the usefulness of the images. Existing image processing techniques do not manage to clearly decipher the text in such pictures. We developed a method for precisely locating text in complex scene images that can then be further processed by OCR systems. A text kernel operator roughly locates the text in an image. This information then serves as an initialization for the active contour method. This technique enhances the convergence process of the active contour and significantly speeds up the overall process. Moreover, our initialization settings enable systems to easily distinguish between the inside and outside parts of contours. Our experimental results show a significant improvement in the ability to locate and preprocess text.**

Keywords-active contours; document analysis; formatting; style; styling;

I. INTRODUCTION

The emergence of new high-end mobile camera phones coupled with the ubiquitous use of mobile technology is changing the way we view document analysis. The vast numbers of images taken everyday is giving rise to many new document indexing applications [1-2]. However, the usefulness of these images is strongly limited by our ability to automatically analyze and tag them.

Images taken by camera phones, even of documents, are rarely of the quality provided by well-calibrated devices, such as flatbed scanners. Furthermore, these images are often not of simple traditional documents, but are compound images including various objects and backgrounds as well as text. The text itself may be of various sizes, fonts, colors, and styles on textured and changing backgrounds. All these factors make an automatic readout process significantly more complicated than traditional document analysis.

In some sense, binarization methods [3–5] can be considered as text detection techniques. However, they have limited precision when processing complex scene images. Region-based and connected component (CC) methods [6] are based on the assumption that text pixel features differ from those of non-text, i.e., background. They require training sets for classifiers and prior information of text position and scale to achieve accurate results. Hybrid methods [7] that combine region-based methods, CCs, and layout analysis methods show improvements. Recently Du et. al. [8] and Li et. al. [9] presented text line segmentation for handwritten documents using the Mumford-Shah model [10] and Chan-Vese piecewise approximation [11], respectively.

In this paper, we propose a new method that robustly locates text of many different styles in an image. The process is fast, yet preserves the quality of the text. Text pixels are roughly located using a text kernel operator, and the results are used for initialization to the active contour method. Our initialization path easily overcomes the active contour restriction to distinguish between the inner and outer parts of the resulting contours and, in our case, between the text and the background.

II. THE TEXT KERNEL OPERATOR

Let $P(x, y)$ denote pixel intensity or color vectors at the x and y point coordinates and let w be the dominant stroke width. Let t be the contrast or color difference to pixels in the immediate neighborhood of the stroke width size. Pixels on strokes in an image can easily be set by applying an operator that checks for contrast along several directions, as in the following operator:

$$\begin{gathered} P(x-w, y) - P(x,y) > t \quad \text{AND} \quad P(x+w, y) - P(x,y) > t \\ \text{OR} \\ P(x, y-w) - P(x,y) > t \quad \text{AND} \quad P(x, y+w) - P(x,y) > t \\ \text{OR} \\ P(x+d, y+d) - P(x,y) > t \quad \text{AND} \quad P(x-d, y-d) - P(x,y) > t \\ \text{OR} \\ P(x-d, y+d) - P(x,y) > t \quad \text{AND} \quad P(x+d, y-d) - P(x,y) > t \end{gathered} ,$$

where d is $w\sqrt{2}$.

The accuracy of the stroke width in this operator is not too important since the strokes of text are well "surrounded" with background; any stroke width greater than the actual stroke width seems to be suitable.

One method of estimating the t parameter is presented by Navon et al. [12]. When text contrast varies across an image, the above method can be applied locally. Practically, t can be set to a minimal contrast to grab lower contrasted pixels.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.53

IEEE computer society

III. THE ACTIVE CONTOUR METHOD

The Chan-Vese model [11] is a curve evolution implementation of the Mumford-Shah model [10]. In this model, segmentation is done by curve evolution driven to minimize the following energy term:

$$\begin{aligned} F(C) &= \mu \cdot Length(C) \\ &+ \lambda_1 \int_{Inside(C)} \left| I(x,y) - c_1 \right|^2 dxdy \\ &+ \lambda_2 \int_{Outside(C)} \left| I(x,y) - c_2 \right|^2 dxdy \end{aligned} .$$

In the term, I is the image, C is the evolving contour, and c_1 and c_2 are the average pixel values inside and outside C, respectively. μ, λ_1, and λ_2 are fixed parameters.

The second and third terms of the function are measures of inter-class variance in each segment. This criterion is the basis for Otsu's well-known binarization method [3]. The first term adds a geometric constraint forcing a tight-fitting minimal length contour. The curve evolution is performed using a level set formulation of the model [11]. In this formulation, the contour C is represented by the zero level of a function φ, so $C = \{(x,y) : \phi(x,y) = 0\}$. Minimizing the energy with respect to φ (while keeping c_1 and c_2 fixed), we obtain the Euler-Lagrange equation for φ (parameterizing the descent direction by an artificial time step):

$$\begin{cases} \dfrac{\partial \phi}{\partial t} = \delta(\phi)\left[\mu div\left(\dfrac{\nabla \phi}{|\nabla \phi|} \right) - \lambda_1 \left(I - c_1 \right)^2 - \lambda_2 \left(I - c_2 \right)^2 \right] \\ \qquad in \quad (0,\infty) \times \Omega, \\ \varphi(0,x,y) = I(x,y) \quad in \quad \Omega, \\ \dfrac{\delta(\varphi)}{|\nabla \varphi|} \dfrac{\partial \varphi}{\partial \vec{n}} = 0 \quad on \quad \partial\Omega \end{cases} ,$$

where Ω is an image domain, $\vec{n}$ denotes the exterior normal to the boundary $\partial\Omega$, and $\dfrac{\partial \varphi}{\partial \vec{n}}$ denotes the normal derivative of φ at the boundary.

When using the active contour method, several problems exist in connection with the initialization of the level set, including a difficulty in differing between the inside and outside parts of contours, and the lengthy processing time. In this paper, we suggest a new initialization to partially overcome these problems.

IV. THE NEW APPROACH PROCEDURE

Our approach for text location is composed of the following steps:

- Create a mask for potential text regions
- Initialize level set map
- Apply active contour on mask
- Handle reverse video and non-reverse video texts

A. Potential Text Regions

Text regions are roughly set by applying the text kernel operator on the input image. The parameter *t* can be set to a minimal contrast value to grab low- and high-contrasted text pixels. The stroke width *w* can be estimated [12] or predefined according to the expected input. Optionally, the operator can be applied for several stroke widths, summing the results. Figure 1 shows the results of applying the text operator on a snippet of an old book image. Most of the text pixels are clearly set. The potential text regions include text kernel pixels and pixels in their close vicinity, as depicted at the bottom of Figure 1.

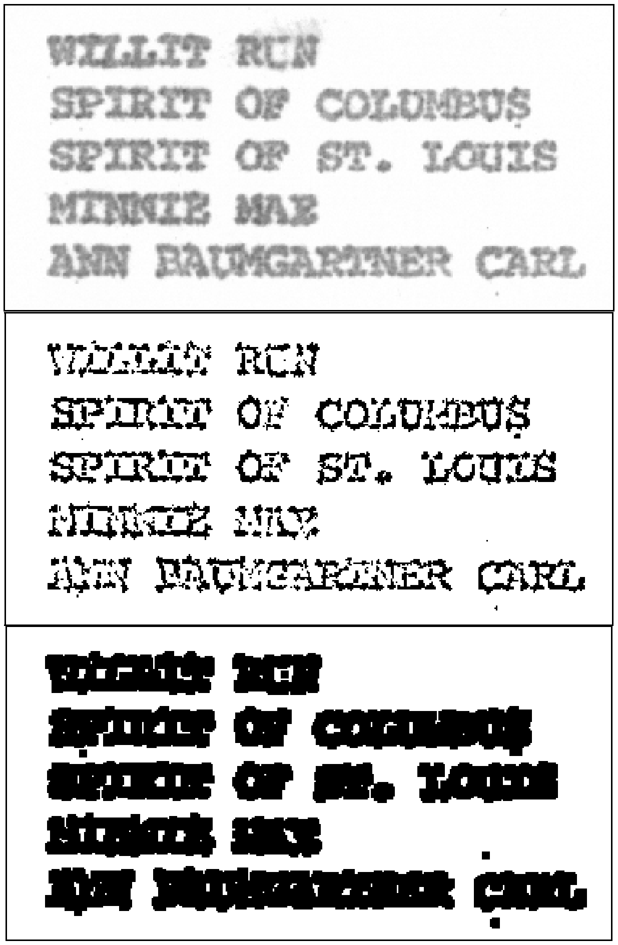

Figure 1. Top – original image; middle – text kernel with *t=30* and *w=4*; bottom – potential text region mask.

B. Initialization of Level Set

The initialization of the level set is required to carry out the time step integration [11]. When no *a priori* information is given on the input image, the initialization of the level set map is set arbitrarily, e.g., using a set of cones spaced over the image, Figure 2 (top) depicts an example of such an initialization.

When *a priori* information on the text regions exists, as shown in the previous section, it can be used for setting the initial level set. This setting is expected to be pretty close to the results after convergence. We choose the following setting:

$$\varphi(0,x,y)=\begin{cases}+1, & (x,y)\in \quad mask\\ -1, & (x,y)\notin \quad mask\end{cases}.$$

Figure 2 (bottom) depicts an example of initialization based on potential text regions.

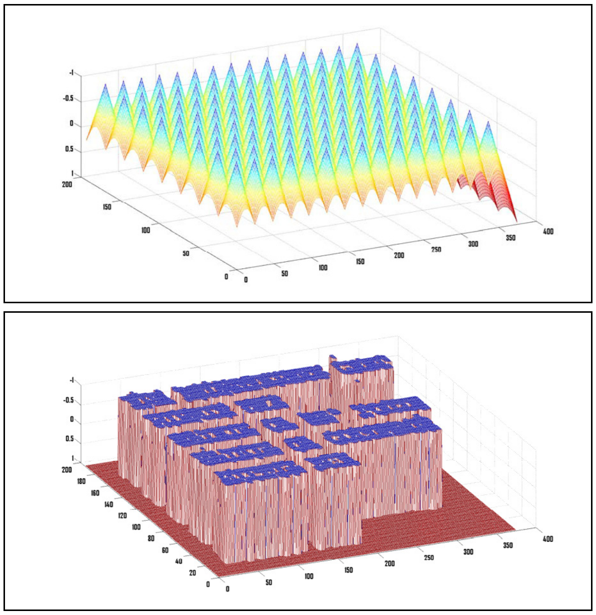

Figure 2. Top – an arbitrary initialization of level set (cones); bottom – initialization based on potential text region mask (set -1.0 in mask, 1.0 out of mask).

C. *Applying Active Contour on Mask*

In our approach, the curve evolution is performed using a level set formulation of the model [11], but only on the potential text regions mask space, and not over the entire image space.

Since text occupies less than 20% to 30% of the image, this path significantly reduces the computational load and converges after a small number of iterations.

Figure 3 depicts the results of the active contour when arbitrarily initialized (Figure 3 top) and when initialized through a potential text mask (Figure 3 bottom), respectively. The quality from our approach is clearly better and is achieved at one-tenth the time-step iterations. Moreover, the inner and the outer parts of contours, as resulting from the active contour process, can be easily mapped to text and non-text areas through the Heaviside signs values [11]. There is no need for post-analysis to distinguish between them. Figure 4 shows an enlarged display of the words "ANN" and "OF" achieved from the arbitrary initialized active contour and the active contour initialized through a potential text mask on top and bottom, respectively.

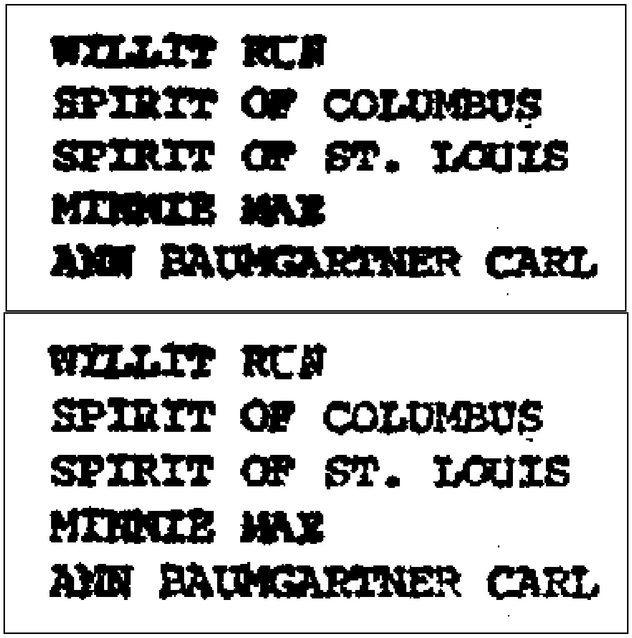

Figure 3. Top – results of arbitrary initialization of active contour after 100 time steps; bottom – results of active contour applied only on potential text region mask after 12 time steps.

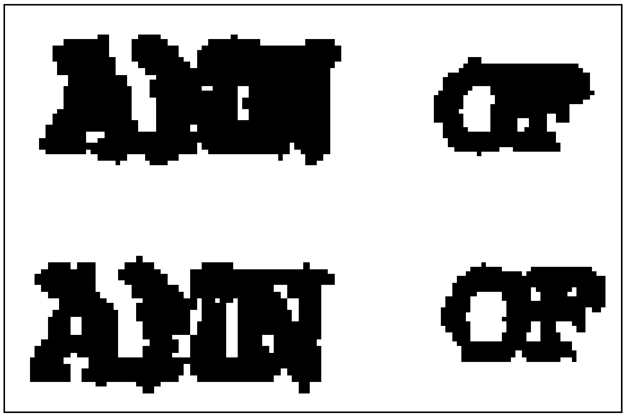

Figure 4. Close up on two words from Figure 3. Top – results of arbitrary initialization of active contour; bottom – results of active contour applied only on potential text region. Note the character separation qualities of our solution.

D. *Reverse Video and Non-reverse Video Text*

Images can include text on different background intensities relative to the text; this is generally portrayed as light text on a dark background, known as reverse video, and/or dark text on a light background, known as non-reverse video. Traditional methods are insufficient for properly extracting both text types; a special technique is required to handle this. The text kernel operator distinguishes between the two types of text by choosing the proper inequality sign. When used with the inequality sign ">", non-reverse video text pixels are set, and when used with the "<" sign, reverse video text pixels are set.

However, when applying the operator, for example, to detect reverse video text in regions where text is printed as non-reverse video, inner parts of characters may behave as reverse video text pixels. This artifact is reflected in an active contour image result, as shown in Figure 5. Two CCs are highlighted. The "red" CC ("L" character) is a regular CC, while the "green" CC (inside the "H" character) is an artifact. Such artifacts can easily be filtered out by checking the homogeneity of CC pixels of close contour; the "green" CC would be filtered out.

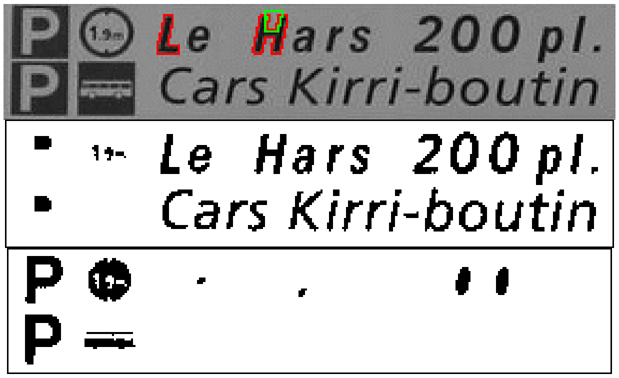

Figure 5. Top - original with close contour of "good" CC (red) and artifact (green) of inner part of "H" character; middle and bottom - non-reverse and reverse video text layers, respectively.

Figure 5 also shows the results when applying our approach to non-reverse and reverse text, respectively, and after filtering out artifacts. There still exist some CCs that required more attention; they can easily be filtered out by OCR means, for example.

V. Experimental Results and Analysis

We evaluated the quality of our approach using human experts. The convergence speed-up is expressed through the number of time steps. The ratio of the number of pixels in a potential text region mask to the total pixels in the entire image, which is needed to process a straightforward case, is indicative of the computational load. The results displayed in Figures 6-8 emphasize the advantages of our results in quality, convergence, and computational load.

Figure 6 shows the results of localizing the non-reverse and reverse video text: 15 time steps with a 0.3 pixel ratio are required to achieve significantly improved text segmentation results.

Eventually, we tested our proposed method as a text segmentation tool, which generally serves as a pre-processing phase for any optical character recognition (OCR) engine. Although the accuracy of commercially available OCR engines has improved to the point that many regard the OCR problem as having been solved, in practice, this statement is far from true. The mean word-level error rates for most OCR engines range roughly from 1–10%, which is unsatisfactory. The digitization process of ancient books constitutes a more serious challenge for every modern OCR technique than standard text. The correct segmentation technique may cardinally improve the OCR accuracy.

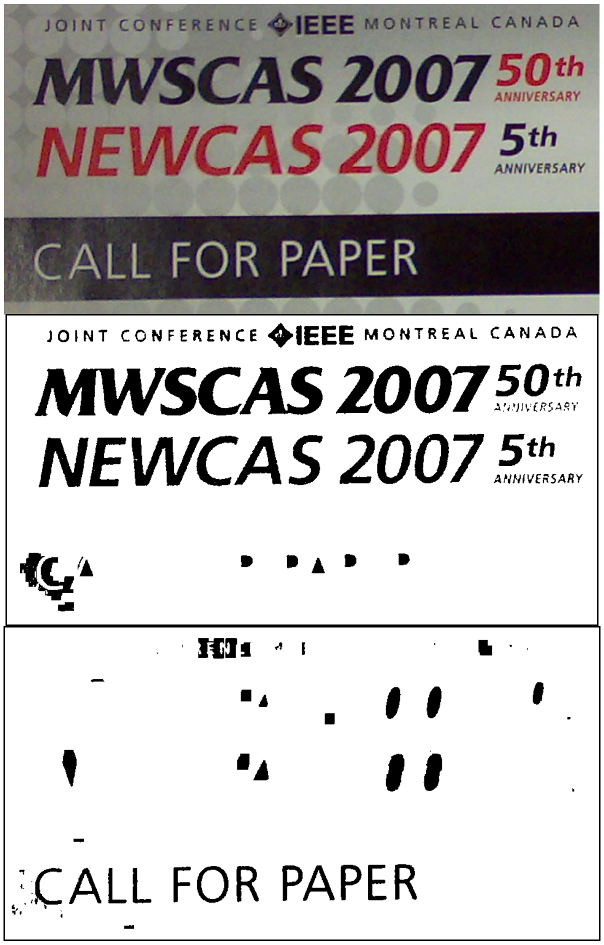
JOINT CONFERENCE IEEE MONTREAL CANADA
MWSCAS 2007 50th ANNIVERSARY
NEWCAS 2007 5th ANNIVERSARY
CALL FOR PAPER
JOINT CONFERENCE IEEE MONTREAL CANADA
MWSCAS 2007 50th ANNIVERSARY
NEWCAS 2007 5th ANNIVERSARY
CALL FOR PAPER

Figure 6. Top - original image; middle - non-reverse layer; bottom - reverse video layer. 15 time steps, pixel ratio 0.3, small details preserved.

The segmentation results of Old Gothic text from an 18th century book using the active contour method are presented in Figure 7. The bottom-right image shows the sufficient result achieved after 21 iterations.

Figure 8 shows good text segmentation results from the same book, when the image contains the bleed-through phenomenon.

We also compared the binarization part of our method with the well-known Niblack's method [4] (see rows 3-4 in Figure 8). Due to space restrictions, we demonstrate the comparison visually, without OCR comparison. We believe a visual evaluation is sufficient for a qualitative estimation of our method.

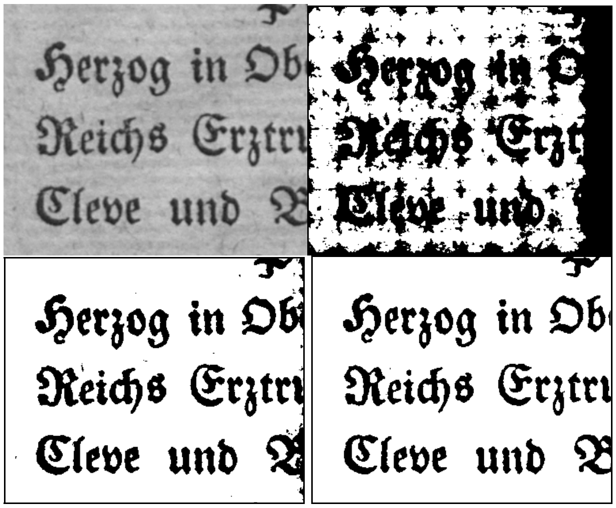

Figure 7. Top left - original image; top right and bottom left - arbitrarily initialized active contour after 21 and 200 time steps, respectively; bottom right - our approach (after 21 time steps, pixel ratio 0.49).

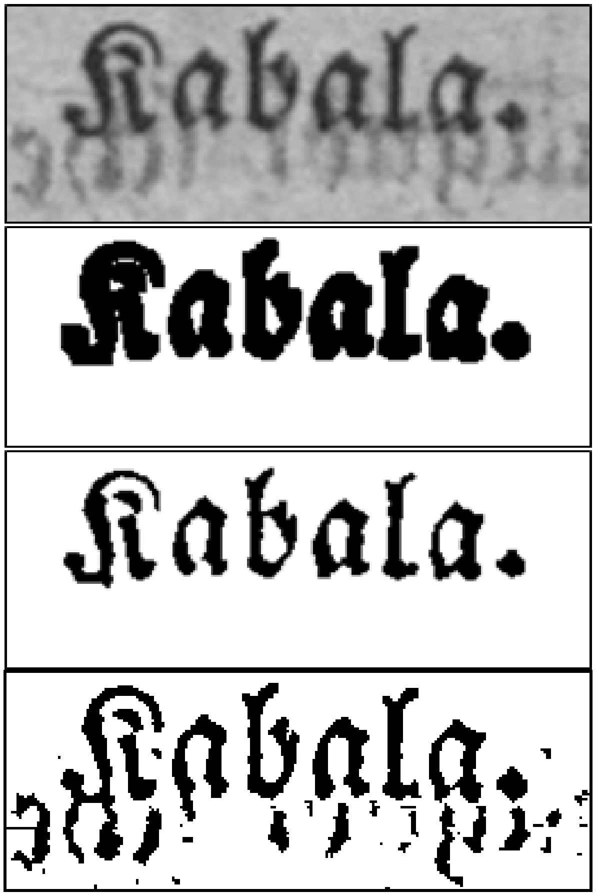

Figure 8. Top - original image; second row – mask image; third row – result of our method; bottom – result of Niblack's method.

VI. Conclusions

The proposed method combines the text kernel operator and active contour method to locate textual areas in images. With the proposed level set initialization scheme, we overcome some of the problems raised by the traditional active contour methods. We easily differ between texts printed on light and dark backgrounds, speed up the convergence, and reduce computational load while achieving high quality results.

References

[1] K. Jung, K. I. Kim, and A. K. Jain. Text Information extraction in images and video: A survey, *Pattern Recognition*, 37(5):977–997, 2004.

[2] J. Liang, D. Doermann, and L. Huiping. Camera-based analysis of text and documents: A survey, *IJDAR*, 7:84–104, 2005.

[3] N. Otsu. A threshold selection method from grey level histogram, *IEEE Trans SMC*, 9, pp. 62-66, 1979.

[4] W. Niblack. *An Introduction to Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, pp. 115-116, 1986.

[5] Y. Navon, Layer-based binarization for textual images, *19th International Conference on Pattern Recognition* (ICPR 2008), December 8-11, 2008.

[6] J. Zhang and R. Kasturi. Extraction of text objects in video documents: Recent progress, in *Proceeding of the Eighth IAPR Workshop on Document Analysis Systems* (DAS'08), pages 1–13, Nara, Japan, 2008.

[7] Yi-Feng Pan, Xinwen Hou, Cheng-Lin Liu. Text Localization in Natural Scene Images based on Conditional Random Field, *10th International Conference on Document Analysis and Recognition*, Barcelona, Spain, 2009.

[8] X. Du, W. Pan, and T. D. Bui, Text Line Segmentation in Handwritten Documents Using Mumford-Shah Model, *Pattern Recognition*, vol. 42, pp. 3136-3145, 2009.

[9] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger, Script-Independent Text Line Segmentation in Freestyle Handwritten Documents, *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, no. 8, pp. 1313 – 1329, 2008.

[10] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems, *Comm. Pure Appl. Math.*, **42** (1989) 577-685.

[11] T. Chan and L. Vese, An active contour model without edges, in *Proceeding of the Second International Conference on Scale-Space Theories in Computer Vision* (Scale-Space '99), Corfu, Greece, September 26–27, 1999.

[12] Y. Navon, A. Heilper, and E. Walach. "Method for OCR Oriented Image Binarization", European Patent No. 98480038.3-2201, November 10, 1998.