

A New Method for Curvilinear Text line Extraction and Straightening of Arabic Handwritten Text

Ayman Al Dmour¹, Ibrahim El rube², and Laiali Almazaydeh¹

¹Faculty of Information Technology, Al-Hussein Bin Talal University, Jordan

²Department of Computer Engineering, Taif University, KSA

Abstract: Line extraction is a critical step from one of the main subtasks of Document Image Analysis, which is layout analysis. This paper presents a new method for curvilinear text line extraction and straightening in Arabic handwritten documents. The proposed method is based on a strategy that consists of two distinct steps. First, text line is extracted based on morphological dilation operation. Secondly, the extracted text line is straighten in two sub-steps: Course tuning of text line orientation based on Hough transform, then fine tuning based on centroid alignment of the connected component that forms the text line. The proposed approach has been extensively experimented on samples from the benchmark datasets of KFUPM Handwritten Arabic Text (KHATT) and Arabic Handwriting DataBase (AHDB). Experimental results show that, the proposed method is capable of detecting and straightening curvilinear text lines even on challenging Arabic handwritten documents.

Keywords: Document image analysis, arabic handwriting, text line extraction, hough transform.

Received January 14, 2016; accepted May 11, 2016

1. Introduction

As the paper document is considered as an important and comfortable media for human to deal with, the ultimate goal would be to integrate the paper documents into our computerized world in order to ensure that the paper documents would be readable like other computer media as optical and magnetic disks [24].

Document Image Analysis (DIA) comes from the advantage of increased advancements in hardware technology either in lower cost or faster speed, to develop a complete system to deal with a huge amount of paper documents flow for more efficient methods of digitizing and indexing documents [31].

The objective of DIA is to extract and recognize both text and graphics components from document images. Therefore, the categories of DIA are divided into Textual processing and Graphics processing. The main subtasks here are:

1. Recognizing the text in a scanned document by Optical Character Recognition (OCR).
2. Determining the skew, finding text lines and paragraphs of the document by Page Layout Analysis.
3. Delimiting straight lines between text sections by Line processing.
4. Filling regions by Region and Symbol Processing. After the application of these DIA techniques, the result will be much more organized semantic description of the document [15].

One of the first and the most major components in pre-processing field of document analysis is the automatic

text line extraction. The objective of this step is the extraction of the text blocks into appropriate text line, to make it possible to prepare the text line for further processing such as word recognition in different OCR systems [13].

Compared to machine printed documents, line extraction in hand printed documents still present a major challenge because of irregular layout, curved and multi-skewed lines, touching and overlapping between neighbouring lines, and no well-defined baselines [16, 29].

Many studies proposed handwritten recognition methods to English language but few have been explored for Arabic language [4]. Arabic is the official language for more than 280 million peoples in 22 countries [19].

The Arabic language needs specifically designed techniques for handwritten recognition especially for line extraction, since this language contains specific characteristics such as diacritical points to distinguish some Arabic characters having the same basic shape with different forms according to its position in the word, and specials marks to modify the character accent, example (م - م - م). These diacritical symbols could reflect extra lines which make the separation of text lines difficult due to the filled space and touching components [5, 31]. In addition, the highly natural cursiveness of character segments in Arabic script greatly complicates the text-line extraction. Furthermore, for Arabic, irregular layout, variable character sizes, varying skew and fluctuating within a text line and the lack of a well-defined baseline also contribute to complicating the extraction of Arabic handwritten lines [17].

Handwritten Arabic document image with different skew in different text-lines is shown in Figure 1.

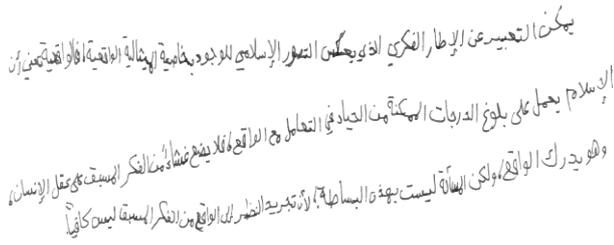


Figure 1. Sample of handwritten Arabic document image which includes text lines with varying skew and fluctuation [20].

To summarize, new handwritten Arabic text lines extraction method is needed in order to have a major benefit of the extraction on DIA outcomes. In this regard, we propose a method for curvilinear text line extraction and straightening of Arabic handwritten documents. The proposed method consists of the following several steps. First, pixel pre-processing is operated in order to simplify and improve the image to make it ready for the next step by thresholding, binarization and noise removing. Then, the document skew is estimated and corrected using Hough transform. Afterwards, the text lines are extracted.

Finally, the coarse and fine tuning is employed in each text line for straightening the line.

The rest of this paper is organized as follows. In section 2, we glance at a variety of text line extraction and straightening methods. Section 3 contains an overview of the proposed approach, and details on the extraction of Arabic handwritten documents into text lines. In section 4, we present the experimental results of our approach. Finally, section 5 concludes this paper regarding the potential usefulness of our approach, and highlights some directions for a possible extension of the proposed approach for future research.

2. Related Work

Over the past few years, several works have been done for the text line extraction for Latin and non-Latin scripts, but, until recently, only few of the related research has focused on the line extraction of handwritten Arabic text because of cursiveness nature of Arabic script.

Concerning the related approaches, they can be roughly divided into two main classes: bottom up and top down. Bottom up approaches are based on pixels processing of the image document. Then, these connected pixels are clustered into bigger elements such as words and lines [25].

In this regards, K_NN is considered as one of the clustering based methods. Zahour *et al.* [30] developed such a method for line extraction suited for Arabic scripts. The document is divided into columns of fixed width. Then, within each column: diacritics, words, and overlapped words between successive lines are considered and represented with the entire three

different size columns. Finally, the close blocks are reassembled to form lines. From this method, an improved version is developed in [9], to deal with multi touching text lines. Also, in [17] a graph based method is employed to extract text lines from Arabic scripts. After computing the orientation of each component, the similarities between closed components are computed to determine the shortest path for text line extraction. The methods in [1, 26] proposed bottom up methods based on connected components clustering.

However, these methods can handle overlapping components, but they need much improvement to handle multi touching components.

On other hand, the top down approach starts from the whole image and iteratively separate it into blocks and the smaller blocks into lines [25]. Projection based methods [6] is one of the top down algorithms that attempts to divide the image into text lines based on the assumption that the text lines are straight and the gap between two neighbouring lines is significant. It is clearly that the approach still not general enough to deal with curvilinear text lines.

Zheng *et al.* [32] authors presented a document model based method to detect broken lines in noisy textual documents. The authors use directional single-connected chain, to extract the line segments. Then, they represent the line model with three parameters: the skew angle, the vertical line gap, and the vertical translation. However, the model is limited to deal with a straight line.

New Hybrid methods have yielded promising results in the field of text line extraction. Here are few examples of hybrid methods, which combine bottom up and top down algorithms: the parametric snake in [10, 11], and the level set method in [19].

The parametric snake: In the formulation of [10, 11], the initial snake position was defined as the central line of text lines image. Then, the energy minimization mechanism is applied to extract text lines.

In [18] proposed level set method based on a Gaussian filter to estimate the pixel density which helps to find text lines.

Recently, in [22] authors have proposed a new painting algorithm to trace and aligning baseline in Persian handwritten text. However, because of cursiveness of Persian handwritten text line, sub words were misaligned with respect to their actual baseline in the sample text lines.

Based on the mentioned related works in this area, our contribution in this work is developing a new approach that is used to correct text lines cursiveness and orientation, in order to have a major benefit of the straightening text lines of Arabic Handwritten documents. The main benefits here are:

1. Improving text recognition when text is oriented upside- down or sideways.

2. Simplifying page layout analysis, finding text lines and text columns of the document.
3. Improving document visual appearance, by removing skew, the document will be much more organized [8].

3. Methodology

Generally, three types of skew can exist within documents [28]: a global skew, when all text lines possess the same orientation; a multiple skew, when certain text lines possess a different orientation than the others in different blocks of the page; and a non-uniform text line skew, when the orientation varying along the same line, and this type is the most challenging one, which is the purpose of this study.

The framework of the overall methodology used in this study is shown in Figure 2.

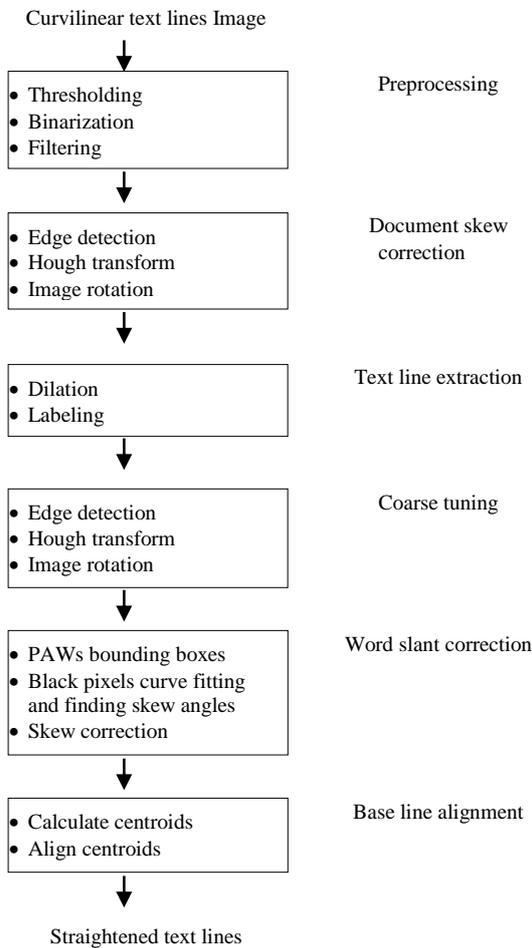


Figure 2. Proposed frame work.

- *Step 1:* The input document image is prepared for the next step by thresholding, binarization and filtering to remove the undesired components.
- *Step 2:* The whole document skew is corrected using Hough transform.
- *Step 3:* Line extraction process begins.

Finally, the coarse and fine tuning is employed in each text line for straightening the line.

In the following; we explain the main steps in more details, along with an example of input Arabic document image (Figure 1), to follow change at each step.

3.1. Pre-Processing

The approach starts by pre-processing the input document using these sub steps [14, 27]:

1. *Thresholding.* By selecting a threshold value of an image with many gray levels we can successfully separate the background/foreground of the image. The threshold is calculated to be the sum of the image mean and standard deviation. Then the threshold is divided by 255.
2. *Binarization.* Which is achieved by applying thresholding on the input image document to get binary image. Using the appropriate calculated threshold, we represent the input image with only two possible values: 1 as the image background with white pixels (paper) and 0 as the image foreground with black pixels (ink).
3. *Filtering.* A 3×3 mask filtering is applied to the document to remove any stray noise, which are random and isolated pixels of white and black pixels. An example of the result of initial pre-processing is shown in Figure 3 with greater quality effect.

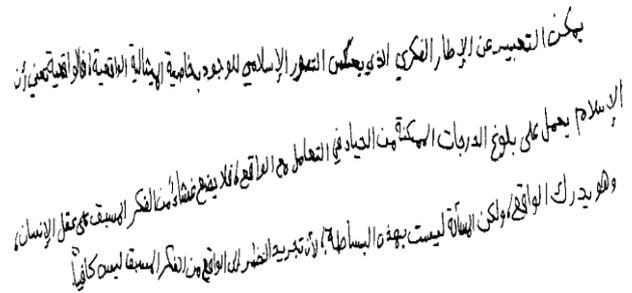


Figure 3. Initial preprocessing image result.

3.2. Document Skew Correction

When a document is scanned, it gets skewed. Therefore, skew correction is necessary for aligning document image horizontally. The following are the sub steps which have applied in order to correct document skew:

- *Edge detection:* the aim of the edge detection step is to locate the text image using processed binary image by employing canny edge detection algorithm [12], which considers as the optimal edge detector. Figure 4 show edge detection result.

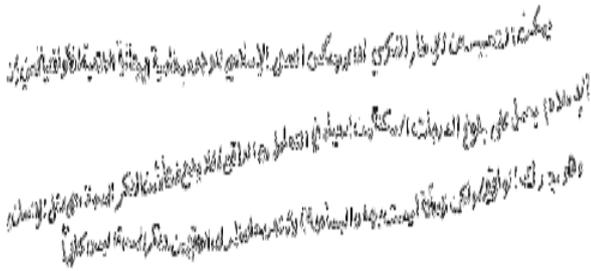


Figure 4. Edge detection result.

- **Hough transform:** hough transform can be applied for skew estimation for handwritten documents. The Hough transform is a global transformation that is used to map points in spatial space (x, y) to a parametric space in (ρ, Θ) via the parametric representation of a line: $\rho = x \cos\Theta + y \sin\Theta$

The variable (ρ) is the distance from the origin to the line along a vector vertical to the line. In addition, the measured skew angle (Θ) is the angle between the x-axis and this vector, which is given as we have applied Hough transform for document skew correction.

Then, each pixel is accumulated on a beam of lines defined in polar coordinates [7]. Thus, Hough transform function automatically detects lines in the document.

- **Image rotation:** obtained by rotating the skewed image with a skew angle (Θ). Figure 5 show result of rotation process using Hough transform.

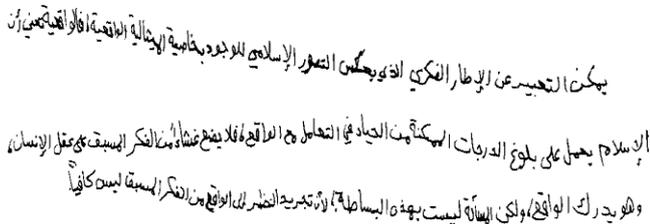


Figure 5. Rotation step result.

Theta, is the angle that gives the highest number of co-linear pixels, i.e., the baseline of the text.

3.3. Text Line Extraction

- **Dilation.** The input document is dilated using a rectangular structuring element. Dilation is a morphological operation whose effect is to thicken or broaden objects in a binary image. The extent of this growing and thickening is controlled by adjusting the size and shape of the structuring element [21].

The resultant image with large connected components along their longer dimension is shown in Figure 6, where each text line is shown as one boundary.

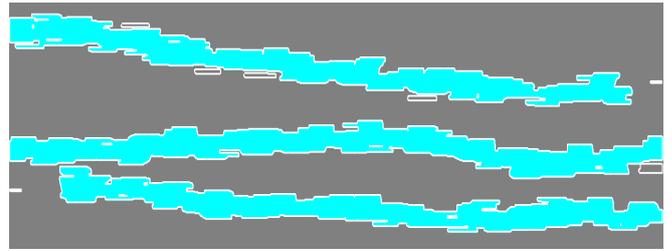


Figure 6. Dilated image.

- **Labelling.** The region boundaries is traced in the binary image. Each connected component is labelled with a specific number.

3.4. Coarse Tuning

As we have applied in step 2 Edge detection, Hough transform and rotation for document skew, the same process is applied here for skew correction concerns the orientation of the selected text line. Figures 7-a and 7-b show the effects of applying Edge detection, Hough transform and rotation on the extracted second text line from the input document.

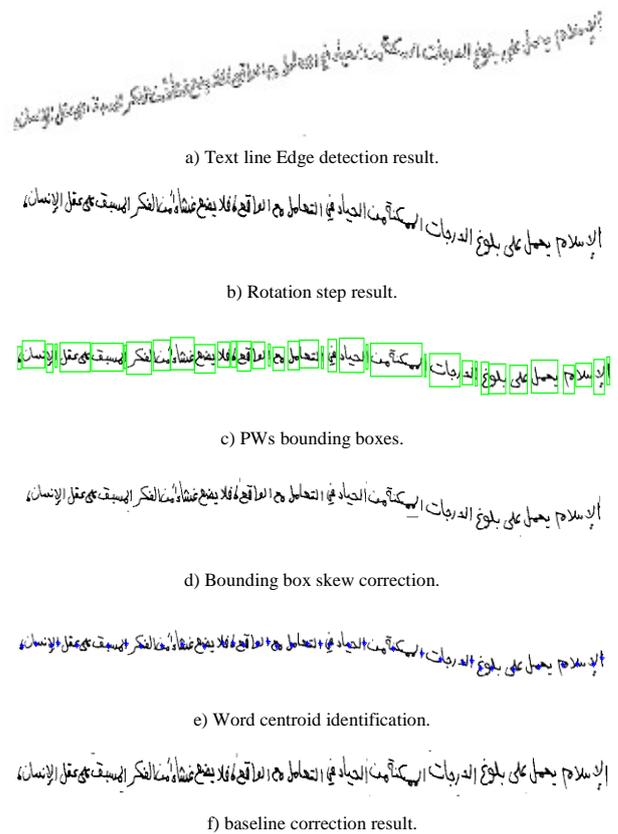


Figure 7. Text line coarse and fine tuning steps.

3.5. Fine Tuning

Each extracted text line which is the outcome of the coarse tuning step, can be fine-tuned for aligning the words inside each text line with respect to the actual base line. Fine tuning follows the process of word slant correction and baseline alignment. In the following, we illustrate each applied sub step:

1. Word slant correction. This step is based on finding centroids of selected Partial Words (PWs) in the image using connected component analysis to estimate optimum skew angles [23]. Therefore, the set of PWs are identified using bounding boxes. Secondly, for curve fitting, the black pixels inside each bounding box are used to find best fit line, then, the skew angle of the fitted line made with x-axis is calculated. Finally, bounding boxes are rotated using calculated angle; the result is shown in Figures 7-c and 7-d.
2. Baseline alignment. For each bounding box, the centroid which represents the center of bounding box mass region and its X, Y coordinates is calculated. Then, a bounding box is adjusted according to the calculated mean centroid. Figures 7-e and 7-f show the effects of baseline correction.

4. Experimental Results

4.1. Database

The database of Arabic handwritten documents used is available from KFUPM Handwritten Arabic Text (KHATT) in [20] and Arabic Handwriting DataBase (AHDB) in [3].

KHATT offers free access to the comprehensive Arabic Handwritten Text Database, which we used to assess and validate our approach. KHATT contains 1000 forms written by distinct writers from different countries. This database provides benchmark data that can be used by researchers to enhance their used mechanisms and tools in the challenging handwritten-related field such as text recognition, writer identification, segmentation, etc., [20].

AHDB contains collection of Arabic words, sentences and pages written by a hundred distinct writers. Arabic handwritten words and text used in different domains such as bank cheques are collected to provide organized database. In addition, a pre-processing class, which contains useful operations to organize the information to facilitate word recognition carried out on AHDB [3].

4.2. Performance Evaluation

We evaluated the effectiveness of our algorithm on KHATT and AHDB databases using different documents available in these databases. The algorithm was implemented using MATLAB (R 2010). All experiments have been run on a machine with 2.1 GHz, 2GB RAM, Windows 7 Operating System.

We can evaluate the performance of our proposed method based on the experimental results of Horizontal Projection Profile (HPP). HPP is a simple histogram accumulating black pixels along every horizontal row of the image. With HPP plot, you can see peaks represent text baseline positions [2].

A sample input image of Arabic handwritten text lines with its straightening text lines are shown in Figures 8-a and 8-b.



Figure 8. A sample output from the proposed technique.

Arabic text-lines are assumed to have been placed with a reference imaginary line called baseline, therefore, the HPP are calculated to reveal the relevance between the position of peak value and the corresponding baseline in the Arabic handwritten text line.

Figure 9 demonstrates the results of HPP in the original image, and image after fine tuning of baseline alignment. Substantial improvements can be observed by aligning the words in image of text line with respect to the actual base line. However, there is only one long and sharp peak in HPP of Arabic text line, which aligns with the position of baseline. Therefore, this characteristic of alignment between position of actual baseline and the peak value in the HPP could be used in Arabic document analysis [22].

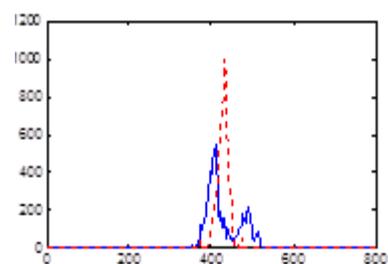


Figure 9. Comparison of HPP of a textline before and after straightening.

From such results of HPP, we can see that our technique succeed in identifying a skew angle when text lines deviate from the x-axis, which belongs to the last group of document skew. Moreover, the approach we developed can be used as a basis for future development of a tool for Arabic handwritten recognition system.

Table 1 provides a summary of the literature research with respect to approach, technique used, and more extra method abilities to deal with line type. From such results, we can see that our approach

outperformed others works by extracting curvilinear text line and straightening the baseline in Arabic Handwritten text-lines.

Table 1. Comparing proposed methods with well known methods in the literature.

Approach	Category	Authors	Straight	Oriented	Cursive	Overlapping	Multi-oriented	curvilinear
Top down	Projection based	Bennasri <i>et al.</i> [6]	✓	✓				
		proposed	✓	✓	✓		✓	✓
	Document Model	Zheng <i>et al.</i> [32]	✓					
Bottom up	K_NN	Zahour <i>et al.</i> [31]	✓	✓	✓			
		Boussellaa <i>et al.</i> [9]	✓	✓	✓			
		Kumar <i>et al.</i> [17]	✓	✓				
	Repulsive-attractive	Oztop <i>et al.</i> [26]	✓	✓	✓	✓		
	Spanning Tree	Abuhaiba <i>et al.</i> [1]				✓		
Hybrid	Snake	Bukhari <i>et al.</i> [10]	✓	✓	✓		✓	
	Level set	Li <i>et al.</i> [18]	✓	✓	✓	✓	✓	

5. Conclusions and Future Works

In this paper, we proposed a new method for curvilinear text line extraction and straightening in Arabic handwritten documents. We first convert a gray scale image to a binary image. Then, the skew of document and text lines are corrected. Afterwards, the text lines are extracted, and finally, correcting baseline by coarse and fine tuning. From the experimental results, we conclude that our algorithm robust to curviness of handwritten Arabic text.

As a future work, we plan to incorporate this work into Arabic handwritten text recognition system that involve efficient operations like line segmentation, word extraction, word recognition and word spotting.

References

[1] Abuhaiba I., Datta S., and Holt M., "Line Extraction and Stroke Ordering of Text Pages," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, pp. 390-393, 1995.

[2] Al-Dmour A. and Fraij F., "Segmenting Arabic Handwritten Documents into Text Lines and Words," *International Journal of Advancements in Computing Technology*, vol. 6, no. 3, pp. 109-119, 2014.

[3] Al-Ma'adeed S., Elliman D., and Higgins C., "A Data Base for Arabic Handwritten Text Recognition Research," *The international Arab Journal of Information Technology*, vol. 1, no. 1, pp. 117-121, 2004.

[4] Al-Nashashibi M., Neagu D., and Yaghi A., "An Improved Root Extraction Technique for Arabic Words," in *Proceedings of 2nd International Conference on Computer Technology and Development*, Cairo, pp. 264-269, 2010.

[5] Al-Rashdi S. and Arockiasamy S., "Adopting Quadrilateral Arabic Roots in Search Engine of E-library System," *International Journal of Recent Research in Social Science and Humanities*, vol. 1, no. 1, pp. 47-53, 2014.

[6] Bennasri A., Zahour A., and Taconet B., "Extraction Des Lignes D'un Texte Manuscrit Arabe," in *Proceedings of Vision Interface'99*, Canada, pp. 42-48, 1999.

[7] Bhowmik T., Roy A., and Roy U., "Character Segmentation for Handwritten Bangla Words Using Artificial Neural Network," in *Proceedings of International Workshop on Neural Networks and Learning in Document Analysis and Recognition*, Seoul, pp. 28-32, 2005.

[8] Bloomberg D., Kopec G., and Dasari L., "Measuring Document Image Skew and Orientation," in *Proceeding of IS & T/SPIE EI'95 Conference*, San Jose, pp. 302-316, 1995.

[9] Boussellaa W., Zahour A., Elabed H., Benabdelhafid A., and Alimi A., "Unsupervised Block Covering Analysis for Text-Line Segmentation of Arabic Ancient Handwritten Document Images," in *Proceedings of International Conference on Pattern Recognition*, Istanbul, pp. 1929-1932, 2010.

[10] Bukhari S., Shafait F., and Breuel T., "Segmentation of Curled Textlines using Active Contours," in *Proceedings of the 8th IAPR Workshop on Document Analysis Systems*, Nara, pp. 270-277, 2008.

[11] Bukhari S., Shafait F., and Breuel T., "Performance Evaluation of Curled Textlines Segmentation Algorithms," in *Proceedings of 9th IAPR Workshop on Document Analysis Systems, DAS'10*, Boston, 2010.

[12] Canny J., "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679-698, 1986.

[13] Fethi G., Mondher M., Snoussi M., and Margner V., "Segmentation of Handwritten and Printed Arabic Documents," in *Proceedings of Workshop on Signal and Document Processing*, Hammamet, pp. 1-5, 2012.

[14] Gonzalez R., Woods R., and Eddins S., *Digital Image Processing Using Matlab*, Reading, MA: Addison-Wesley, 2004.

[15] Kasturi R., O'Gorman L., and Govindaraju V., "Document Image Analysis: A Primer," *Sadhana*, vol. 27, no. 1, pp. 3-22, 2002.

- [16] Khayyat M., Lam L., Suen C., Yin F., and Liu C., "Arabic Handwritten Text Line Extraction by Applying an Adaptive Mask to Morphological Dilation," in *Proceedings of 10th IAPR International Workshop on Document Analysis Systems*, Gold Cost, pp. 100-104, 2012.
- [17] Kumar J., Abd-Almageed W., Kang L., and Doermann D., "Handwritten Arabic Text Line Segmentation using Affinity Propagation," in *Proceeding of the 9th IAPR International Workshop on Document Analysis Systems*, Boston, pp. 135-142, 2010.
- [18] Li Y., Zheng Y., Doermann D., and Jaeger S., "Script- Independent Text Line Segmentation in Freestyle Handwritten Documents," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1313-1329, 2008.
- [19] Lutf M., You X., and Li H., "Offline Arabic Handwriting Identification Using Language Diacritics," in *Proceedings of International Conference on Pattern Recognition*, Istanbul, pp. 1912-1915, 2010.
- [20] Mahmoud S., Ahmad I., Al-Khatib W., Alshayeb M., TanvirParvez M., Märgner V., and Fink G., "KHATT: An Open Arabic Offline Handwritten Text Database," *Pattern Recognition*, vol. 47, no. 3, pp. 1096-1112, 2014.
- [21] Marqoues O., *Practical Image and Video Processing Using MATLAB*, Wiley, 2011.
- [22] Nagabhusan P. and Alaei A., "Tracing and Straightening the Baseline in Handwritten Persian/Arabic Text-line: A New Approach Based on Painting-Technique," *International Journal on Computer Science and Engineering*, vol. 2, no. 4, pp. 907-916, 2010.
- [23] Nandini N., Murthy K., and Kumar H., "Estimation of Skew Angle in Binary Document Images using Hough Transform," *World Academy of Science, Engineering and Technology*, vol. 2, no. 6, pp. 44-49, 2008.
- [24] O'Gorman L. and Kasturi R., *Document Image Analysis*, Computer Society Executive Briefing, 2009.
- [25] Ouwayed N. and Belaid A., "A General Approach for Multi-oriented Text Line Extraction of Handwritten Documents," *International Journal on Document Analysis and Recognition*, vol. 15, no. 4, pp. 1-18, 2011.
- [26] Oztop E., Mulayim A., Atalay V., and Yarman Vural F., "Repulsive Attractive Network for Baseline Extraction on Document Images," *Signal Processing*, vol. 75, no. 1, pp. 1-10, 1999.
- [27] Peake G. and Tan T., "Script and Language Identification from Document Images," in *Proceedings of the British Machine Vision Conference (BMVC97)*, Essex, pp. 610-619, 1997.
- [28] Razak Z., Zulkiflee k., Idris M., and Yaacob M., "Off-Line Handwriting Text Line Segmentation: A Review," *International Journal of Computer Science and Network Security*, vol. 8, no. 7, pp. 12-20, 2008.
- [29] Yi L., Zheng Y., Doermann D., and Jaeger S., "Script-Independent Text Line Segmentation in Freestyle Handwritten Documents," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1313-1329, 2008.
- [30] Zahour A., Taconet B., and Ramdane S., "Contribution `a la Segmentation De Textes Manuscrits Anciens," in *Proceedings of Confrence Internationale Francophone Sur l'Ecrit et le Document*, La Rochelle, 2004.
- [31] Zahour A., Taconet B., Likforman L. and Boussella W., "Overlapping and Multi-Touching Text-Line Segmentation by Block Covering analysis," *Pattern Analysis and Applications*, vol. 12, no. 4, pp. 335 -351, 2009.
- [32] Zheng Y., Li H., and Doermann D., "A Model-Based Line Detection Algorithm in Documents," in *Proceedings of 7th International Conference on Document Analysis and Recognition*, Edinburgh pp. 44-48, 2003.



Ayman Al-Dmour received his BSc in Electronic - Communication Engineering in 1994 from Jordan University of Science and Technology, Irbid, Jordan. He pursued his MSc and PhD in 2003 and 2006, respectively, both in Computer Information Systems in the Arab Academy for Banking and Financial Sciences, Amman, Jordan. At Al-Hussein Bin Talal University (AHU), he has led the Department of Computer Information Systems, the Computer and Information Technology Center and the College of Information Technology. His research interests are in Arabic language processing, data compression and computer education.



Ibrahim El rube' received his M.Sc. degree in Computer Engineering from Arab Academy for Science and Technology, Egypt in 1999 and his Ph.D. in Systems Design Engineering from the University of Waterloo in 2005. Currently, he is working as associate professor at the Computer Engineering Department in Taif University, Taif-KSA. His research interests include pattern recognition and image processing.



Laili Almazaydeh is an assistant professor of Software Engineering at Al-Hussein Bin Talal University (AHU) in Jordan. She received a B.S. in Computer Science from Al-Hussein Bin Talal University and an M.S. in Computer Information Systems from The Arab Academy for Banking and Financial in 2003 and 2007, respectively. She received her Ph.D. in Computer Science and Engineering at the University of Bridgeport in 2013, USA. Her research interests involve the wireless sensor networks, image processing and human computer interaction.