High Performance Layout Analysis of Arabic and Urdu Document Images

Syed Saqib Bukhari[1], Faisal Shafait[2], and Thomas M. Breuel[1]
[1]*Technical University of Kaiserslautern, Germany*
[2]*German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany*
bukhari@informatik.uni-kl.de, faisal.shafait@dfki.de, tmb@informatik.uni-kl.de

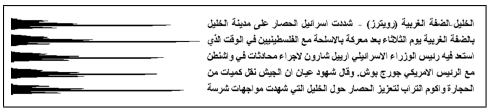
***Abstract*—Text-lines extraction and their reading order determination is an important step in optical character recognition (OCR) systems. Research in OCR of Arabic script documents has primarily focused on character recognition and therefore most of researchers use primitive methods like projection profile analysis for text-line extraction. Although projection methods achieve good accuracy on clean, skew-corrected documents, their performance drops under challenging situations (border noise, skew, complex layouts, . . .). This paper presents a robust layout analysis system for extracting text-lines in reading order from scanned Arabic script document images written in different languages (Arabic, Urdu, Persian) and styles (Naskh, Nastaliq). The presented system is based on a suitable combination of different well established techniques for analyzing Latin script documents that have proven to be robust against different types of document image degradations. Evaluation of the presented system on Arabic and Urdu document image datasets consisting of a variety of complex single- and multi-column layouts achieves high accuracies for text and non-text segmentation, text-line extraction, and reading order determination.**

***Keywords*-Document Layout Analysis, Text-Line Segmentation, Text Image Segmentation, Reading Order Determination**

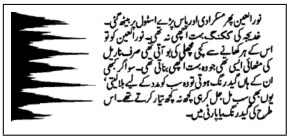
I. Introduction

This paper addresses the problem of layout analysis of machine-printed Arabic script document images. The Arabic script–a cursive script–is used for writing several languages of Asia and Africa, like Arabic, Urdu, Persian, Pashto, Jawi, Nubi, and Juba. Although there are many styles for writing Arabic script, the most widely used styles are Naskh (dominant in Arabic and Pashto) and Nastaliq (standard for Urdu and Persian). An example of Arabic and Urdu texts is shown in Figure 1.

Layout analysis–text and non-text segmentation, text-line extraction, and reading order determination–is a major performance limiting step in large scale document digitization projects. Over the last two decades, several layout analysis algorithms have been proposed in the literature [1], [2] that work for different layouts, scripts and are quite robust to the presence of noise in documents. Research on Arabic OCR has primarily been focused on word recognition [3], and very few approaches have been proposed for layout analysis for machine-printed Arabic script document images. Here, we briefly discuss some state-of-the-art document image layout analysis approaches in connection to Arabic documents.



(a) Sample Arabic Naskh script



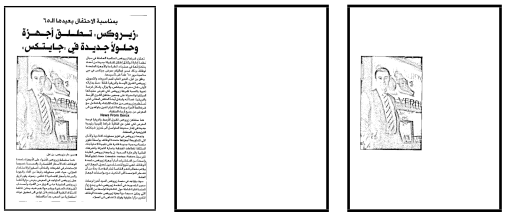
(b) Sample Urdu Nastaliq script

Figure 1. An example of printed Arabic (Naskh) and Urdu (Nastaliq) text. Nastaliq script has many differences to Naskh script, the most important of which from layout analysis point of view are very small inter-line and inter-word spacing, and tall ascenders and descenders that penetrate into adjacent text-lines, which are illustrated here by showing their corresponding projection profiles.

Text and non-text segmentation is an important layout analysis step, which may directly affect the performance of further layout processing tasks such as text-line extraction, and/or character recognition. The performance of classification based text and non-text segmentation approaches [4] heavily depends on training samples, and they can not be directly applied to different scripts. On the other hand, smearing [5] and multiresolution morphology [6], [7] based approaches work on an assumption that non-text elements are bigger than text elements, but these approaches are script independent and can be directly used for Arabic script document images.

Text-line extraction is the backbone of a layout analysis system. Kumar et al. [8] have evaluated the performance of six algorithms for page segmentation on Nastaliq script: the x-y cut [9], the smearing [5], whitespace analysis [10], the constrained text-line finding [11], Docstrum [12], and the Voronoi-diagram based approach [13]. These algorithms work very well in segmenting documents in Latin script as shown in [14]. However, none of these algorithms were able to achieve an accuracy of more than 70% on their test data which had simple book layouts. Shafait et al. [15] adapted RAST [11] for text-line extraction on Urdu script documents. They modified the reading order algorithm in [16] to

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.257



(a) an example image (b) non-text mask [6] (c) non-text mask [7]

Figure 2. Text and non-text segmentation using multiresolution morphology based method [6] and its improved version [7]. The non-text portion in the example image (a) is composed of a large number of small components, that were missed by the original version, but correctly segmented by the improved version. [Note: the left-most image has been taken from the website: http://www.ijma3.org/.]

obtain reading order of Urdu text-lines. More sophisticated approaches for text-line extraction have been presented in the domain of segmenting handwritten Arabic documents [17]. However, the key problem addressed in these approaches is to handle local non-linearity of text-lines.

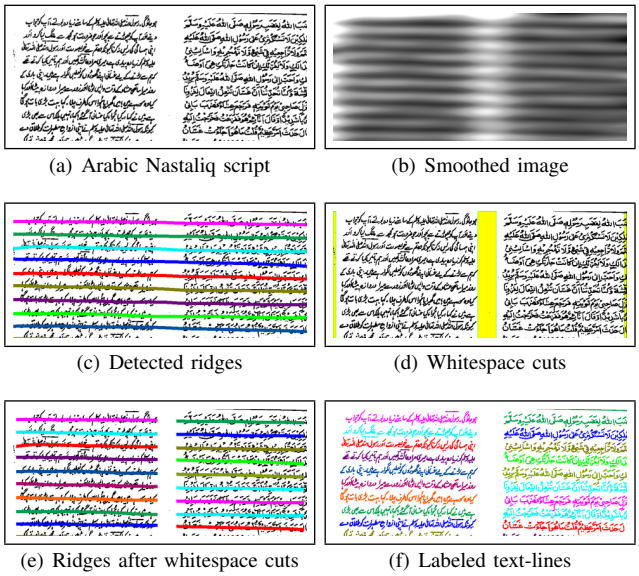
In this paper, we present a high performance layout analysis system for a wide variety of Arabic and Urdu document images that belong to a diverse collection of layout structures such as books, magazines, and newspapers. Our layout analysis system is a suitable combination of robust and well-established text and non-text segmentation, text-line extraction, and reading order determination techniques. First, it performs text and non-text segmentation using multiresolution morphology based method [7]. Then, it extracts text-lines by adapting ridge based text-line finding method [18] for a variety of single- and multi-column layouts. Finally, it determines the reading order of text-lines using topological sorting of extracted text-lines [15]. In this way, our layout analysis system extends Shafait et al. [15] Urdu layout analysis system (text-line extraction in reading order) by incorporating text and non-text segmentation and a better text-line extraction method. To evaluate the performance of the presented layout analysis system for real-world documents, a dataset of Arabic documents is prepared and the already available dataset of Urdu documents [15] is used.

Details of our methods are given in the book chapter [19]. This paper focuses on an extensive experimental evaluation of the presented layout analysis system and its comparison with state-of-the-art techniques.

The rest of this paper is organized as follows. Our layout analysis system for Arabic script document images is described in Section II. Performance evaluation and experimental results are discussed in Section III, followed by a conclusion in Section IV.

II. Arabic Document Layout Analysis

The Arabic document layout analysis system presented here consists of following main steps, text and non-text



(a) Arabic Nastaliq script (b) Smoothed image (c) Detected ridges (d) Whitespace cuts (e) Ridges after whitespace cuts (f) Labeled text-lines

Figure 3. Snapshots of ridge based text-line finding method as described in Section II-B.

segmentation, text-line extraction, and reading order determination. These steps are briefly described here; please refer to [19] for details.

A. Text and Non-Text Segmentation

Bloomberg [6] presented a multiresolution morphology based text and non-text segmentation method. It is a simple and script independent text and non-text segmentation method. It performs well for halftone mask segmentation, for which it was designed, but most of the time fails to accurately segment drawing type non-text elements such as line art, maps etc. We presented an improved multiresolution morphology based text and non-text segmentation algorithm in [7], that can handle halftones as well as drawing type non-text elements. A sample document image and its text and non-text segmentation results for the original and the improved version of multiresolution morphology based methods are shown in Figure 2.

B. Text-line Extraction

There is a large number of script independent text-line extraction methods in literature. Among them, x-y cut [9] is a state-of-the-art method based on project profile analysis that can handle multi-column documents with small inter-line spacing. However, the method fails on document images with skew or noise. We presented a ridge based text-line extraction method for warped camera-captured [18] and handwritten [20] document images. Our ridge based approach is robust to presence of noise, skew and small inter-line spacing, and it can be directly used for Arabic document images. Ridge based text-line finding method

Figure 4. Sample images from Arabic and Urdu documents datasets and their corresponding layout analysis results: the black-pixels represents non-text components, the color coded labeling represent extracted text-lines and the magenta-line shows reading order of segmented text-lines. [top-left] Arabic-English book page; [top-right and bottom-left] Arabic newspapers (these images have been taken from the websites: *http://www.alrostamanigroup.ae* and http://www.mawred.org, respectively; [bottom-right] Urdu poetry image.

consists of two standard and easy to understand image processing algorithms: (i) Gaussian filter bank smoothing and (ii) ridge detection. A brief description of this method is presented below.

To generate a set of filters, the ranges are first defined for the parameters (σ_x, σ_y and θ) of Gaussian filter, either empirically or automatically using document statistics. Then, the set of filters is applied to each pixel value and the correspoding maximum output response is selected for the smoothed image. In order to speedup Gaussian filter bank smoothing, a fast anisotropic Gaussian filter implementation [21], [22] is used. A sample document image and its corresponding smoothed image are shown in Figure 3(a) and 3(b), respectively. After smoothing, text-lines are extracted by detecting ridges from the smoothed image. The detected ridges are shown in Figures 3(c). In case of multi-column document images, a single ridge may cover either a single text-line or multiple corresponding text-lines in different columns. This situation may lead to over-segmentation errors as shown in Figures 3(c). These over-segmentation errors are removed/reduced by cutting the parts of the ridges which lie over white spaces. Here, whitespace cuts are estimated by applying the method as presented in [11]. After cutting, resulting ridges are shown in Figure 3(e), where each ridge covers a single text-line. Labeled text-lines for the input document image of Figure 3(a) is shown in Figure 3(f).

Table I
PERFORMANCE EVALUATION OF THE ORIGINAL MULTIRESOLUTION MORPHOLOGY BASED TEXT AND NON-TEXT SEGMENTATION METHOD [6] AND ITS IMPROVED VERSION [7] FOR ARABIC DATASET (25 DOCUMENTS).

	Original [6]	Improved [7]
text classified as text	99.82%	99.80%
non-text classified as non-text	99.15%	99.60%
segmentation accuracy	99.49%	99.70%

C. Reading Order Determination

A reading order determination method tries to determine the order in which a human will go through different parts of a document. Breuel [16] described a method for determining reading order of Latin script documents. He used topological sorting of text-lines on the basis of predefined ordering criteria for determining their reading order. Reading direction of Arabic text is from right to left, which is opposite to the Latin script. Shafait et al. [15] modified the method in [16] for Urdu (Nastaliq) script. Here, we apply the same modified method for determining reading order of Arabic script document images.

For sample Arabic and Urdu documents, result of the presented layout analysis system (text and non-text segmentation, text-line extraction, and their reading order direction) are shown in Figure 4.

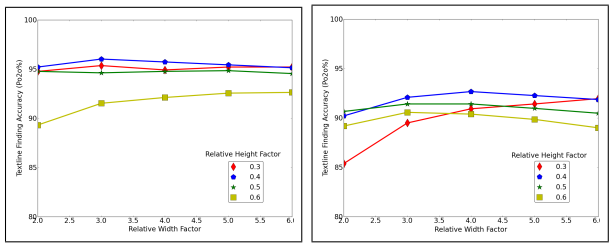
III. PERFORMANCE EVALUATION

For the performance evaluation of the presented layout analysis system, we collected 25 images of Arabic documents, mostly Naskh script, from books, newspapers, and multi-script (English and Arabic) documents. These images contain both text and non-text elements. For this dataset, text and non-text, text-line, and reading order level ground-truths are prepared in color coded pixel form. We have also selected 20 images from Urdu documents dataset [15] (Nastaliq script), which belong to the categories of books, poetries, digests, and magazines. Urdu dataset contains text elements only. Like Arabic dataset, the text-line and reading order level ground-truths are also provided in color coded form in Urdu dataset. Both datasets contain a variety of single- and multi-column layouts as shown in Figure 4, and hence they can be used to evaluate the performance of a layout analysis algorithm for Arabic document images.

Here, the performance evaluation of the presented layout analysis systems is done in three parts. The first part evaluates the performance of text and non-text segmentation (Section III-A), the second part analyzes the errors made in text-line detection (Section III-B), and the third part evaluates the accuracy of reading order (Section III-C).

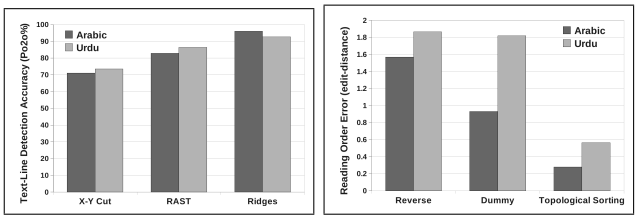
A. Text and Non-Text Segmentation Accuracy

The performance evaluation metrics for text and non-text segmentation accuracy are described in [7]. These metrics



(a) Arabic Dataset (25 documents; 1358 text-lines) (b) Urdu Dataset (20 documents; 2237 text-lines)

Figure 5. Plot against one-to-one segmentation accuracy of our ridge based text-line finding method as described in Section II-B and different values of its free parameters on (a) Arabic documents and (b) Urdu documents.



(a) text-lines accuracy (b) reading order accuracy

Figure 6. (a) Performance evaluation results of one-to-one text-line extraction accuracy of x-y cut [9], adapted RAST [15], and ridge based (Section II-B) text-line finding methods on Arabic and Urdu datasets. (b) Performance evaluation results of reading order determination of topological sorting based [15], the dummy and the reverse reading order determination methods.

evaluates the percentage of non-text pixels classified as non-text, text pixels classified as text, and the average of both is considered as segmentation accuracy. Here, we apply the same metrics for evaluating the performance of multiresolution morphology based text and non-text segmentation method [6] and its improved version [7] on Arabic documents dataset. Performance evaluation results are shown in Table I. Arabic dataset contains only text and halftone elements, and no drawing or any other type of non-text elements. Therefore, both original and improved versions achieved nearly similar and good segmentation accuracy.

B. Text-Line Extraction Accuracy

The performance evaluation metrics for text-line detection accuracy are defined in [14], where a text-line is said to be correctly detected if it does not fall into any of the following types of errors: over-segmentation, under-segmentation, missed text-lines, and false-alarms. From these metrics, we use one-to-one correctly detected text-lines accuracy ($P_{o2o}\%$). For ridge based text-line finding method on Arabic dataset, we achieved a performance gain from 93.89% to 96.02% after text and non-text segmentation. Figure 5 shows the one-to-one text-line finding accuracy of our ridge based text-line finding algorithm for different values of its free parameters for both Arabic documents (after text and non-text segmentation) and Urdu documents datasets. The relative flatness of the curves in Figure 5 indicates that our method is reasonably stable with respect to the free parameters. The performance evaluation results on both Arabic and Urdu datasets of ridge based, adapted RAST [15] and x-y cut [9] text-line finding methods, with optimized values of their free parameters, are shown in Figure 6(a). The ridge based method has achieved above 96% text-line finding accuracy for Arabic dataset and above 92% for Urdu dataset, which are better than the performance of adapted RAST [15] and x-y cut [9] methods on these datasets. X-Y cut method usually fails due to small inter-line gaps and presence of multiple columns. Under these conditions, RAST works better than x-y cut but gives errors for very small inter-line gaps and page curl. Our ridge based method performs better than both x-y cut and RAST under small inter-line gaps, multiple columns and page curl, but fails for very small inter-line gaps.

C. Reading Order Determination Accuracy

A reading order determination algorithm heavily depends on text-line detection accuracy. Here, we apply an edit distance based reading order performance evaluation strategy, such that edit distance is calculated between a detected and the corresponding ground-truth reading orders. We have also defined a dummy and a reverse reading order determination methods for comparison with topological sorting based reading order determination method [15]. The dummy method simply returns the sorted order of text-lines with respect to their baseline positions. The reverse-order method returns a complete reverse reading order with respect to a given ground-truth information. For simple document layouts, dummy method gives better reading order than reverse-order. However, for complex document layouts (like Urdu poetry as shown in Figure 4), both give bad result. The performance evaluation results of topological sorting based [15], dummy, and reverse reading order determination methods are shown in Figure 6(b). Dummy method performs better than reverse-order method for Arabic dataset, but for Urdu dataset both give almost same error, which is also comparatively larger than the errors for Arabic dataset. This also demonstrates that, layouts in Urdu dataset are more challenging than Arabic dataset with respect to reading order. The topological sorting based reading order determination method performs better than both dummy and reverse-order methods. It gives an incorrect reading order if two text-lines from different text columns are merged, because in such a case they are interpreted as a separator. It gives larger error for Urdu dataset than Arabic dataset, because it cannot handle Urdu poetry written in two column format or other likewise layouts, as it is misinterpreted as a two-column text.

IV. Conclusion

In this paper, we have presented a high performance layout analysis system for machine printed, scanned Arabic and Urdu document images, which are composed of a variety of single- and multi-column layouts. The presented layout analysis system is composed of a suitable combination of well-established and robust text and non-text segmentation, text-line extraction, and reading order determination methods. We have evaluated the presented layout analysis system on 25 Arabic and 20 Urdu document images, which are composed of a variety of layouts as shown in Figure 4. For text and non-text segmentation, multiresolution morphology based method [7] is used. We have achieved above 99% text and non-text segmentation accuracy on Arabic dataset. For text-line extraction, ridge based method is used, which is described in (Section II-B). For ridge based method, we have achieved above 96% text-line detection accuracy for Arabic dataset and above 92% for Urdu dataset, which are better than the performance of both x-y cut [9] and adapted RAST [15] based text-line detection methods on these datasets. For determining the reading order of extracted text-lines, we have used topological sorting based reading order determination method [15]. We have achieved better reading order accuracy as compared to the dummy and the reverse reading order determination methods. Altogether, the presented layout analysis system showed good performance for text and non-text segmentation, text-line extraction, and reading order determination on a variety of Arabic and Urdu document images, and it can be used for large scale Arabic and Urdu documents digitization processes.

References

[1] R. Cattoni, T. Coianiz, S. Messelodi, and C. M. Modena, "Geometric layout analysis techniques for document image understanding: a review," IRST, Trento, Italy, Tech. Rep. 9703-09, 1998.

[2] G. Nagy, "Twenty years of document image analysis in PAMI," *IEEE TPAMI*, vol. 22, no. 1, pp. 38–62, 2000.

[3] H. E. Abed and V. Märgner, "ICDAR 2009-Arabic handwriting recognition competition," *Int. Journal on Document Analysis and Recognition*, vol. 14, pp. 3–13, 2011.

[4] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Document image segmentation using discriminative learning over connected components," in *Proc. Workshop on Document Analysis Systems*, Boston, USA, 2010, pp. 183–190.

[5] K. Y. Wong, R. G. Casey, and F. M. Wahl, "Document analysis system," *IBM Journal of Research and Development*, vol. 26, no. 6, pp. 647–656, 1982.

[6] D. S. Bloomberg, "Multiresolution morphological approach to document image analysis," in *Proc. Int. Conf. on Document Analysis and Recognition*, France, 1991, pp. 963–971.

[7] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Improved document image segmentation algorithm using multiresolution morphology," in *Proc. SPIE Document Recognition and Retrieval XVIII*, San Jose, CA, USA, Jan. 2011.

[8] K. S. Kumar, S. Kumar, and C. Jawahar, "On segmentation of documents in complex scripts," in *9th Int. Conf. on Document Analysis and Recognition*, Brazil, Sep. 2007, pp. 1243–1247.

[9] G. Nagy, S. Seth, and M. Viswanathan, "A prototype document image analysis system for technical journals," *Computer*, vol. 7, no. 25, pp. 10–22, 1992.

[10] H. S. Baird, "Background structure in document images," in *Document Image Analysis*, H. Bunke, P. Wang, and H. S. Baird, Eds. World Scientific, Singapore, 1994, pp. 17–34.

[11] T. M. Breuel, "Two geometric algorithms for layout analysis," in *Proc. Workshop on Document Analysis Systems*, Princeton, NY, USA, Aug. 2002, pp. 188–199.

[12] L. O'Gorman, "The document spectrum for page layout analysis," *IEEE TPAMI*, vol. 15, no. 11, pp. 1162–1173, 1993.

[13] K. Kise, A. Sato, and M. Iwata, "Segmentation of page images using the area Voronoi diagram," *Computer Vision and Image Understanding*, vol. 70, no. 3, pp. 370–382, 1998.

[14] F. Shafait, D. Keysers, and T. M. Breuel, "Performance evaluation and benchmarking of six page segmentation algorithms," *IEEE TPAMI*, vol. 30, no. 6, 2008.

[15] F. Shafait, A. Hasan, D. Keysers, and T. M. Breuel, "Layout analysis of Urdu document images," in *10th IEEE Int. Multitopic Conference, INMIC'06*, Islamabad, Pakistan, Dec. 2006.

[16] T. M. Breuel, "High performance document layout analysis," in *Symposium on Document Image Understanding Technology*, Greenbelt, MD, USA, April 2003.

[17] W. Boussellaa, A. Zahour, H. E. Abed, A. Benabdelhafid, and A. M. Alimi, "Unsupervised block covering analysis for text-line segmentation of arabic ancient handwritten document images," in *ICPR*, Istanbul, Turkey, 2010, pp. 1929–1932.

[18] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Ridges based curled textline region detection from grayscale camera-captured document images," in *Int. Conf. on Computer Analysis of Images and Patterns*, Germany, 2009, pp. 173–180.

[19] ——, "Layout analysis of arabic script documents," in *Guide to OCR for Arabic Scripts*. Springer-Verlag, 2011.

[20] ——, "Script-independent handwritten textlines segmentation using active contours," in *Proc. Int. Conf. on Document Analysis and Recognition*, Spain, 2009, pp. 446 – 450.

[21] J. M. Geusebroek, A. W. M. Smeulders, and J. V. D. Weijer, "Fast anisotropic Gauss filtering," *IEEE Trans. on Image Processing*, vol. 12, p. 2003, 2003.

[22] C. H. Lampert and O. Wirjadi, "An optimal nonorthogonal separation of the anisotropic gaussian convolution filter," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3501 –3513, 2006.