A Zone Classification Approach for Arabic Documents using Hybrid Features

Amany M.Hesham, Sherif Abdou, Amr Badr
Faculty of Computers and Information
Cairo University
Cairo, Egypt

Mohsen Rashwan
Faculty of Engineering
Cairo University
Cairo, Egypt

Hassanin M.Al-Barhamtoshy
Computing and Information Technology
King Abdulaziz University (KAU)
Saudi Arabia

***Abstract*—Zone segmentation and classification is an important step in document layout analysis. It decomposes a given scanned document into zones. Zones need to be classified into text and non-text, so that only text zones are provided to a recognition engine. This eliminates garbage output resulting from sending non-text zones to the engine. This paper proposes a framework for zone segmentation and classification. Zones are segmented using morphological operation and connected component analysis. Features are then extracted from each zone for the purpose of classification into text and non-text. Features are hybrid between texture-based and connected component based features. Effective features are selected using genetic algorithm. Selected features are fed into a linear SVM classifier for zone classification. System evaluation shows that the proposed zone classification works well on multi-font and multi-size documents with a variety of layouts even on historical documents.**

***Keywords*—segmentation; layout analysis; texture features; connected component analysis; Arabic script; genetic algorithms**

I. INTRODUCTION

Layout analysis is the process of extracting text lines from a document image and identifying their reading order. It is a major step in converting documents into electronic text [1]. The first major step in document layout analysis is segmentation. In this step a documents are divided into distinct geometrical regions where each is classified as text or non-text region. Text regions contain text only while non-text regions may contain images, graphs, drawings, etc. as shown in Fig. 1. Text regions are sent to character recognition engine [2] which converts text images into digital text while there is no need to send images to recognition engine as it will give garbage output.

(a) (b) (c)

Fig. 1. Examples for regions in Arabic documents (a) text region (b) non-text (image) region (c) non-text (drawing) region

Our focus is on Arabic script documents. Arabic script languages –such as Arabic, Kurdish, Urdu, Persian etc.- is the second most used script after Latin script. Arabic script segmentation is a challenging problem as it is written in many different styles. In addition, Arabic script is cursive where letters are connected together forming ligatures.

In literature, many algorithms were presented for document images segmentation. Kumar et al. [3] evaluated the performance of six page segmentation algorithms on different scripts. The evaluated algorithms are X-Y cut, whitespace analysis, constrained text-line finding, Docstrum, Voronoi diagram and smearing algorithms. X-Y cut algorithm [4] is a recursive top-down algorithm. It is a tree based algorithm where the whole document image represents the root of the tree. A document is then decomposed recursively into rectangular blocks, until the document can't further be split, which forms tree nodes. Whitespace analysis algorithm [5] analyzes the structure of background in a document image. It extracts the maximal whitespace rectangles, to be merged subsequently. Merged rectangles cover the background and thus isolate text blocks. Constrained text-line detection algorithm [6] is also a top down algorithm. It performs a two-step text lines extraction algorithm. First, it extracts maximal empty rectangles covering the whitespace background as in whitespace analysis algorithm. Then, the extracted rectangles are considered obstacles that are used in text line detection step. Document spectrum, or docstrum, algorithm [7] is a bottom up algorithm. It works on connecting document components using the nearest neighbor clustering algorithm forming the regions. Voronoi-diagram based algorithm [8] is also a bottom-up algorithm. It extracts sample points from the boundaries of connected components. Voronoi cells are extracted surrounding the sample points. A decision is taken on the edges whether to be removed or not based on two features, distance and ratio of cell areas. The predefined algorithms are evaluated on different scripts including Arabic scripts particularly on Urdu documents. They give poor segmentation results on complex script documents such as Arabic scripts. As based on the algorithms evaluation on Urdu documents, maximum accuracy achieved was 71.5%. Another evaluation was done by Shafait et al. [9] on the same six algorithms but only on Latin English scripts. Their algorithms evaluation shows that constrained text-line finding, Docstrum and Voronoi give better results than X-Y cut, whitespace analysis and smearing algorithm.

Moreover, other approaches are used like multiresolution morphological operations as in Bloomberg algorithm [10]. Bloomberg algorithm performs well on text and halftone document segmentation. But except for halftones, it gives poor results on any other non-text like drawings, graphs, etc. Bukharia et al. [11] proposed an improvement on Bloomberg algorithm to be generalized. A hybrid between Voronoi and Docstrum is also proposed in [12].

On the other hand there are other text segmentation approaches based on classification of segmented text zones. Pixel based approaches [13] which work on classifying each pixel in a document into text or non-text. Connected component analysis based approaches [14] where each extracted component is classified into a text or non-text class. Edge based [15] approaches where edged images are portioned into blocks and each block is classified to a class. Texture analysis based [16] approaches which make use of the regular periodic texture property in text regions and irregular textures in non-text zones. In this work we propose a new approach for zone classification in Arabic documents. The proposed algorithm applies text and non-text segmentation using dilation morphological operation and connected component analysis. Features extracted from each segmented zone are then used in classifying text and non-text regions. Those features combine texture and connected component based features.

The rest of this paper is organized as follows. Page segmentation of Arabic script document images is described in Section II. Starting by preprocessing then followed by zone segmentation. Extracted zones classification is discussed in Section III. Performance evaluation and experimental results are discussed in Section IV. Results analysis and final conclusions are included in Section V.

II. Page Segmentation

A. Preprocessing

Collected document images are required to be in the same format for any further processing. Since, collected document images may be colored or grey scaled; images need to be unified in bi-level representation –black, white-. This binarization step is the main pre-processing step. In [17] an adaptive local thresholding method is used for image binarization. Image is then cleaned using a 5×5 median filter for noise removal. Marginal noise and document borders are also removed. Skewed images are corrected using Radon transform [18] to be aligned correctly for further processing. The result from preprocessing phase is a unified, cleaned and aligned binary image with text and non-text components only.

B. Zone Segmentation

This step works on segmenting binary image into a set of zones. Segmentation algorithm used in this step is based on multiresolution morphological operations, dilation operation. Connected component analysis is applied to extract mean height (μ_h) and mean width (μ_w) of the extracted connected components. Image is dilated using a rectangular structuring element with size relative to μ_h and μ_w. The resulting image contains each zone components connected together. Each zone is surrounded by a rectangle as shown in Fig. 2.

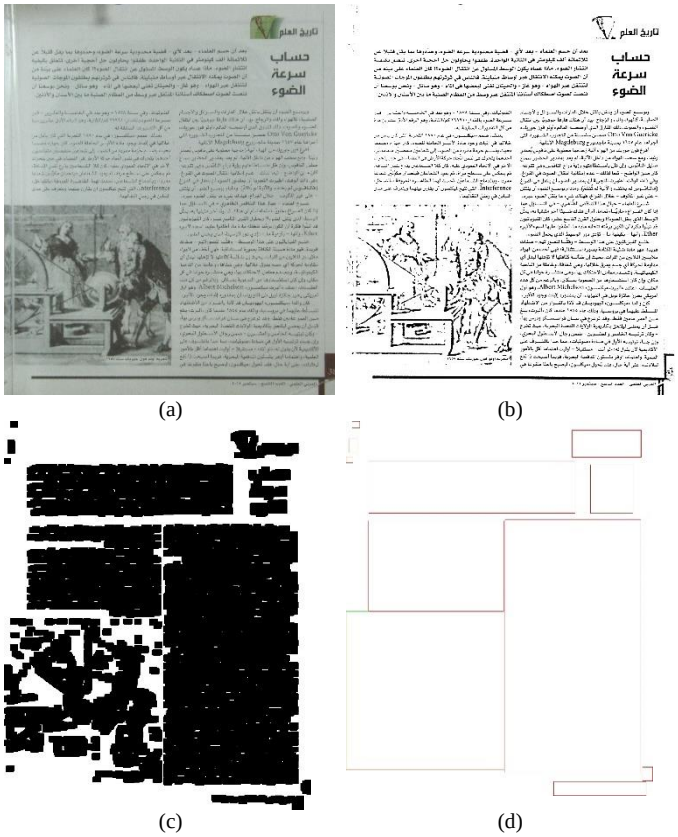


Fig. 2. Given (a) Original book page image (b) binarized and deskwed image using Sauvola algorithm (c) dilated image (d) image segmented into zones

III. Zone Classification

Classifying each zone into text and non-text requires extraction of a set of features. Those features are a mixture of connected component and textural based features. Textural features are features which contains spatial distribution of an image pixel intensities. Given the regular periodicity of the intensity distribution in text regions unlike non-text regions [16], textural features are good choice in zone classification. Textural features are extracted from sub-blocks where each sub-block represents part of a component e.g. letter or word as shown in Fig 3. Connected components based features are added to represent the whole zone characteristics. It was found that letters and words have common features e.g. height of connected components are near, while non-text regions have random structure e.g. big scattered objects and small noise.

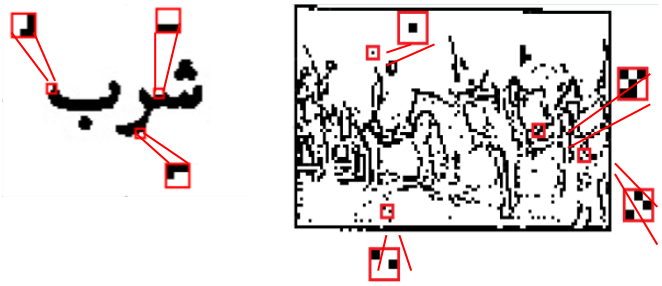


Fig. 3. Example for textural feature extraction for (a) text (b) non-text region

b_8	b_7	b_6
b_5	b_4	b_3
b_2	b_1	b_0

Fig. 4. Order of pixels in DSE block

C. Textural based features

Feature used in the proposed system is the document structure element DSE [19]. It is based on a 3x3 blocks. Pixels distribution in a given image block is shown in Fig. 4. There exist 2^9 = 512 different DSEs range from 0 to 511. For each pixel a 3x3 block, with this pixel as center b_4, is converted into integer which is called document structure element characteristic number DSECN [19]. DSECN is calculated using the following formula:

$$DSECN_{(x,y)} = \sum_{i=0}^{8} b_i 2^i \quad (1)$$

where b_i = *{0,1}*; 0 for black pixels and 1 for white pixels.

Features are calculated by counting the frequency of each of the 512 blocks. Some of the features are irrelevant which causes conflict in the classification between classes -text and non-text-. It is infeasible to extract all 2^{512} different possible combination between features of the DSEs to find the relevant and most effective features. As a result, a feature selection algorithm is used for dimensionality reduction without reducing accuracy. Only subset of features, which hold sufficient information used in discriminating classes, are selected. There exist many feature selection algorithms in literature, some apply linear transformations on the extracted feature vectors like principle component analysis (PCA) [20]. Others use evolution-based optimization techniques like Genetic algorithms (GA) [20] which focus on feedback result from the classifier to the feature selector.

In the proposed system GA is applied. Binary chromosome is used to represent the 512 features. Fitness used for the selection process is the accuracy of the model generated from a linear kernel support vector machine [21] (SVM) classifier. Input to the SVM is a subset of selected features and the output is the accuracy of the input features on a training set. The subset of features giving the highest accuracy is considered to hold the most effective subset of features. GA results in 223 features which is almost half the original set. Accuracy of the classification process using the selected features is 97.83% on cross validation data. Experimenting the whole 512 features and the 223 selected features gives equal results.

D. Connected Component based features

As previously explained, features are mixture of textural and connected component based features. Connected component analysis is applied and some statistical features are extracted to enhance the accuracy of the classification process. Features extracted from each zone are:

1) *Aspect ratio between width and height*
2) *Percentage of foreground to total zone pixels*
3) *Normalized maximum components height*
4) *Normalized maximum components width*
5) *Normalized mean width of the components*
6) *Normalized Standard deviation for width of the components*
7) *Normalized mean height of the components*
8) *Normalized standard deviation for height of the components*
9) *Mean for aspect ratio of the components*
10) *Standard deviation for aspect ratio of the components*

Features are normalized by dividing its value by the document image value, e.g. width, height. Adding the 10 connected component based features to the selected textural features improves the accuracy of zones classification to reach 98.54% on cross validation data. This means that connected component features increased the classification by 0.71%.

IV. System Evaluation

Evaluation of the current system requires collecting document images because of the absence of standard Arabic document dataset. Our Data set was 85 scanned documents collected from different sources. Documents are scanned using 12 Mega pixel camera. Images are collected from 3 main sources – books, magazines and historical documents- with different layouts e.g. single and multi-column documents. Images are as follow, 60 images from books, 10 images from historical documents and 15 images from magazines. Evaluation of the proposed zone classification algorithm is shown in table I. Results are shown for each set separately. In Books set, 1400 segmented and tested zones, 650 text zones and 750 non-text zones. In magazine set, 350 segmented and tested zones, 210 text zones and 140 non-text zones. Finally, historical documents set, 150 segmented and tested zones, 40 text zones and 110 non-text zones. Accuracy is calculated from zone classified as text or non-text using precision and recall. The precision is the ratio of the number of text zones to the total number of zones classified as text. The recall is the ratio of the number of text zones to the total text zones.

$$recall = \frac{\#(text\ zones\ correctly\ recognized)}{\#(actual\ text\ zones)} \quad (2)$$

$$precision = \frac{\#(text\ zones\ correctly\ recognized)}{\#(zones\ recognized\ as\ text)} \quad (3)$$

Table I shows that books and magazines precision and recall are near. However, in historical documents the values are far. The high difference occurs as a result of high degradation in old historical documents. Low precision value results from big amount of noisy non-text zones detected as text. However, recall value is high which shows that only few text-zones are detected as non-text. Figure 5 shows some examples for documents sent to the system.

Our approach is also evaluated against the preprocessing components of commercial state of art Arabic OCR systems, RDI Clever Page [22] and Sakhr Arabic OCR [23]. RDI Clever Page is based on connected component analysis and whitespace analysis. Evaluation is applied on a set of 109 books and magazines images. Results in table II show that our proposed classification approach has high recall value; which means that few text zones were missed. RDI system results in higher precision value but with smaller recall value compared

to our system. This means that RDI system misclassifies many text zones as non-text than our system which may result in losing important information from original document. On the other hand, our system overpasses the values of Sakhr system in both precision and recall.

V. CONCLUSION

In this paper we proposed a framework for documents layout analysis. It consists of two main phases 1) segmentation and 2) classification. Collected documents are binarized using an adaptive global binarization method, Sauvola. Images are then cleaned and skewed for further processing. Binarized images are segmented using morphological dilation operation into zones. Features are extracted for classification of segmented zones. Features used are combination between connected components and textural based features. Textural features are reduced using genetic algorithm and added to connected component features. Each zone is classified into text or non-text by sending features to a linear SVM classifier. Only text zones are sent for character recognition engine OCR to avoid garbage resulting from sending non-text regions. Evaluation results show that the proposed approach managed to achieve high accuracy for text zones classification in Arabic documents. Comparing those results with state of art commercial Arabic OCR systems it achieved the best recall while saving high accuracy.

TABLE I. EVALUATION OF ZONE CLASSIFICATION

Document type	Precision	Recall
Books	97.5%	96.7%
Magazines	94.5%	94.5%
Historical Documents	77%	95.2%
Total	**95.7%**	**96.1%**

TABLE II. COMPARISON BETWEEN PROPOSED SYSTEM AND OTHER SYSTEMS

Methods	Precision	Recall
Proposed system	93.5%	99.1%
RDI	97.9%	94.1%
Sakhr	91.6%	95.7%

ACKNOWLEDGEMENT

The teamwork of the "Arabic Printed OCR System" project was funded and supported; by the NSTIP strategic technologies program in the Kingdom of Saudi Arabia- project no. (11-INF-1997-03). In addition, the authors acknowledge with thanks Science and Technology Unit, King Abdulaziz University for technical support.

REFERENCES

[1] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Layout analysis of Arabic script documents," in Guide to OCR for Arabic Scripts, V. Märgner and H. El Abed, Eds. London: Springer London, 2012, pp. 35 – 53.

[2] M. Attia, M. a a Rashwan, and M. S. M. El-Mahallawy, "Autonomously normalized horizontal differentials as features for HMM-based omni font-written OCR systems for cursively scripted languages," in International Conference on Signal and Image Processing Applications, 2009, pp. 185–190.

[3] K. S. S. Kumar, S. Kumar, and C. V. Jawahar, "On segmentation of documents in complex scripts," in International Conference on Document Analysis and Recognition (ICDAR 2007), 2007, pp. 1243–1247.

[4] R. M. Haralick and I. T. Phillips, "Recursive X-Y cut using bounding boxes of connected components," in International Conference on Document Analysis and Recognition, 1995, vol. 2, pp. 952–955.

[5] H. Baird, "Background structure in document images," Int. J. Pattern Recognit. Artif. Intell., vol. 8, pp. 1013 – 1030, 1994.

[6] T. Breuel, "Two geometric algorithms for layout analysis," Doc. Anal. Syst. v, pp. 188 – 199, 2002.

[7] L. O'Gorman, "The Document Spectrum for Page Layout Analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 11, pp. 1162–1173, 1993.

[8] K. Kise, A. Sato, and M. Iwata, "Segmentation of Page Images Using the Area Voronoi Diagram," Comput. Vis. Image Underst., vol. 70, no. 3, pp. 370–382, Jun. 1998.

[9] F. Shafait, D. Keysers, and T. Breuel, "Performance comparison of six algorithms for page segmentation," Doc. Anal. Syst. VII, pp. 368 – 379, 2006.

[10] D. S. Bloomberg, "Multiresolution morphological approach to document image analysis," in International Conference on Document Analysis and Recognition, 1991, pp. 1–12.

[11] S. Bukhari, "Improved document image segmentation algorithm using multiresolution morphology," IS&T/SPIE Electron. Imaging, 2011.

[12] M. Agrawal and D. Doermann, "Voronoi++: A dynamic page segmentation approach based on voronoi and docstrum features," in International Conference on Document Analysis and Recognition, 2009, pp. 1011 – 1015.

[13] M. a. Moll, H. S. Baird, and C. An, "Truthing for Pixel-Accurate Segmentation," in International Workshop on Document Analysis Systems, 2008, pp. 379–385.

[14] S. S. Bukhari, M. Ibrahim, F. Shafait, and T. M. Breuel, "Document Image Segmentation using Discriminative Learning over Connected Components," in International Workshop on Document Analysis Systems, 2010, pp. 183 – 190.

[15] M. Pietikäinen and O. Okun, "Edge-Based Method for Text Detection from Complex Document Images," in International Conference on Document Analysis and Recognition, 2001, pp. 286–291.

[16] O. Okun and M. Pietikäinen, "A survey of texture-based methods for document layout analysis," in Texture analysis in machine vision, World Scientific, 1999, pp. 165–177.

[17] F. Shafait, D. Keysers, and T. M. Breuel, "Efficient implementation of local adaptive thresholding techniques using integral images," Doc. Recognit. Retr. XV, Proc. SPIE, vol. 6815, p. 681510, 2008.

[18] J. Dong, D. Ponson, A. Krzyżak, and C. Y. Suen, "Cursive word skew/slant corrections based on Radon transform," in 8th International Conference on Document Analysis and Recognition, 2005, pp. 478 – 483.

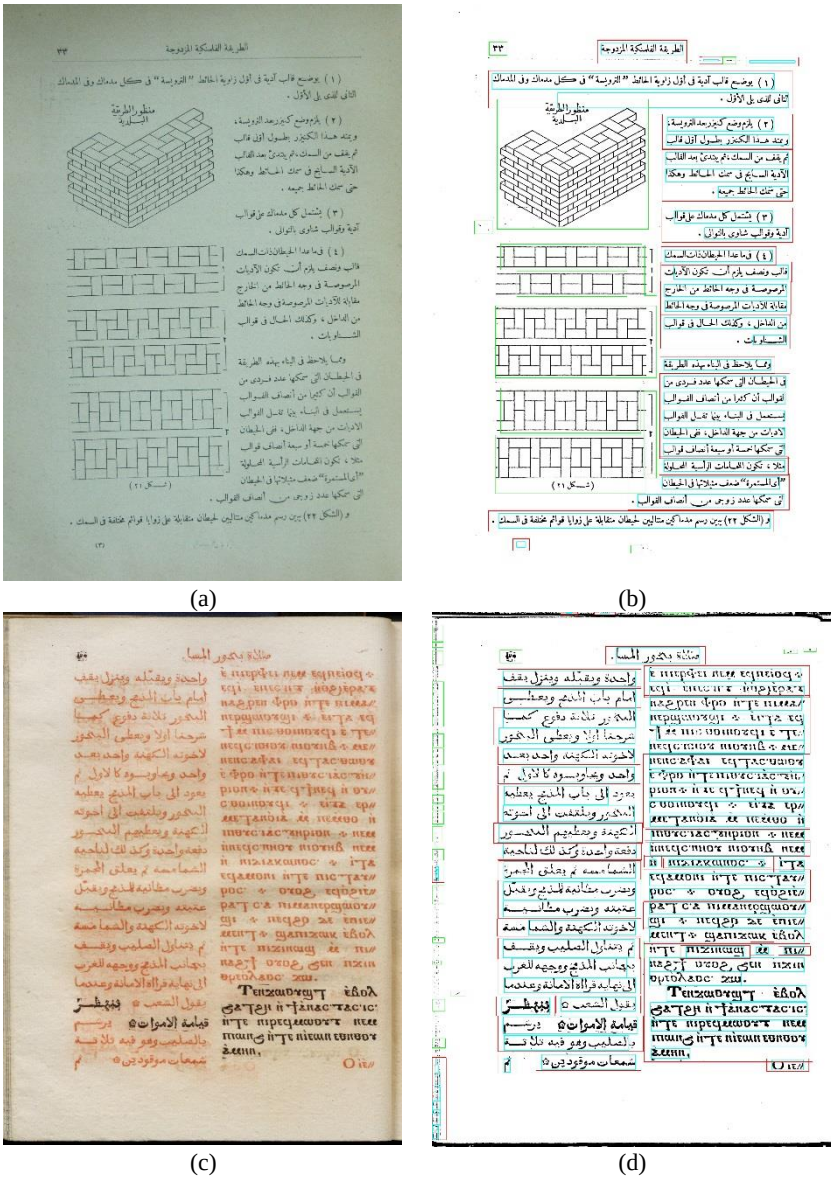
[19] C. Strouthopoulos and N. Papamarkos, "Text identification for document image analysis using a neural network," Image Vis. Comput., vol. 16, pp. 879–896, Aug. 1998.

[20] M. Raymer and W. Punch, "Dimensionality reduction using genetic algorithms," IEEE Trans. Evol. Comput., vol. 4, no. 2, pp. 164–171, 2000.

[21] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, pp. 27 – 66, 2011.

[22] RDI, "Clever Page - Arabic omni font-written OCR," http://www.rdi-eg.com/projects/OCR.htm.

[23] Sakhr, "OCR (Optical Character Recognition)," http://www.sakhr.com/index.php/en/solutions/ocr.



(a) (b) (c) (d)

Fig. 5. Examples for applying algorithm on different documents