Doctoral School of Computer Science
University of Szeged
Institute of Informatics

THE APPLICATION OF NEW METHODS FOR OFFLINE RECOGNITION IN PRINTED ARABIC DOCUMENTS

Summary of the Ph.D. thesis

by

Hassina Bouressace

Supervior: Dr.János Csirik
Professor emeritus

Szeged, 2020

1 Introduction

Nowadays, the raw image can be replaced by a set of structured information exploitable by the machine, and millions of stored paper volumes can be replaced by computer files in XML format. These types of document analysis systems should cut the number of misfiled, misshelved and lost files and will increase automated sorting, automated text/non-text recognition and make it quicker and more accurate especially for Arabic documents, where great effort is still required to attain the performance for English documents. These objectives serve as a motivation for exploring prospective solutions for Arabic document image analysis and Arabic writing analysis.

Figure 1: Examples of Arabic document pages

The analysis of document layouts (DLA) is an important step in many areas especially for an OCR system where its input is a printed or handwritten text document without any graphic elements, hence if any document has a non-text element, the system will not give the desired perfect solution. For this, DLA is a necessary step before OCR, which is used for extracting and recognizing all the existing elements on a page, either as text or non-text elements and specifying each element according to its features. This is why extracting information from a document means the extraction of all the graphical-textual elements present on a document page (see Figure 1). Each category has various types and sizes as well because each document can have a different form and structure (e.g. a newspaper page or a magazine page). Such a combination complicates the segmentation process for homogeneous regions, and it is made worse if the document was acquired under different imaging conditions (e.g. with noise, uneven illumination, skew, perspective distortion, motion blur). Several algorithms and methods have recently been developed for DLA for English and other languages, but with documents in Arabic, many challenges still have to be overcome.

Natural language processing is the basic part of most language searches of the world, where the programs are generated in such a way that they can readily comprehend

and manipulate human language text. Segmenting Arabic sentences is a crucial step for Arabic recognition as it is used in many natural language processing technologies such as parsing, machine translation, and research. Many recent studies have addressed segmentation problems in the Arabic language, and many issues still have to be overcome due to the great variety of handwriting styles.
The main objective of this study is to develop new methods and techniques for Arabic document layout analysis and Arabic handwritten text segmentation by improving the results of these tasks and making them more general than before. For this, we present a summary of several approaches that seek to overcome the main problems encountered here.
Here, we give a brief overview of the most important results, which are divided into three major sections. In each section we highlight the main problems, then provide a proposed solution to tackle it, describing an Arabic document database in Section 2. After, we outline our proposal methods for Arabic document layout analysis in Section 3. In Section 4, we present our improved approach for Arabic word detection.
Most of the ideas, tables and figures that have appeared in the dissertation are covered by the following publications:
- Section 2 presents our Arabic document database with its statistics. Publication (1);
- Subsection 3.1 presents our printed Arabic newspaper recognition approach with its results. Publication (2);
- Subsection 3.2 presents our title/subtitle detection in printed Arabic newspapers approach with its results. Publication (3);
- Subsection 3.3 presents our smartphone-captured Arabic newspaper analysis approach with its results. Publication (4);
- Section 4 presents our Arabic handwritten word detection approach with its results. Publication (5);

2 Arabic Document Database

Due to the interest in transforming the documents into digital images over the past few decades, many algorithms, systems, and databases have been created and developed. However, this abundance does not cover all the languages such as the Arabic language. This motivated us to create a new database for camera-captured document images.
We sought to create a database of Arabic newspapers that would include a wide range of structures, pictures, sizes, fonts and certain images of varying smartphone-captured quality. Therfore, we selected a set of document images taken from ten different newspapers (Alriad, Alsharek, Alhadaf, Alnahar, Alshourouk, Alshorouk-almisri, Alsharek al-Awsat, Aswaq Qatar, Alayam, Akhersaa), composed of 200

newspaper pages; 705 articles, where each article can appear on more than one image (see Table 1), and 14 fonts written in three writing styles (normal, italic and bold). This set of document pages was printed with an HP laser printer, and the pages were scanned at 300 dpi resolution in grayscale format with an HP scanner.

Table 1: Statistics of the image sets used for database production.

Newspaper	**One article**	**Two articles**	**Three articles**	**More than Three**
Alriad	41 images	15 images	6 images	13 images
Alsharek	36 images	8 images	5 images	14 images
Alhadaf	28 images	6 images	4 images	9 images
Alnahar	46 images	18 images	7 images	17 images
Alshourouk	74 images	21 images	10 images	15 images
Alshorouk almisri	46 images	12 images	8 images	13 images
Alsharek al-Awsat	35 images	11 images	7 images	15 images
Aswaq Qatar	36 images	13 images	6 images	14 images
Alayam	37 images	15 images	3 images	13 images
Akhersaa	79 images	26 images	10 images	19 images
Total	**458 images**	**145 images**	**66 images**	**141 images**

Using both the printed and screen versions, we created a dataset for a smartphone-captured Arabic text database. Afterwards, we used 810 documents (page parts) to capture 2954 images with mobile phones for our text image dataset. The images were then stored in JPG format. The page images were divided into article-blocks with manual segmentation that contained one article, several articles or all the articles on one page. Each page has a unique structure due to the diversity of its contents, which appear in many article-shapes, image-shapes, titles-sizes, titles-fonts, text-sizes, and text-fonts.

We created a new smartphone-captured database by capturing the document images for the PATD (Printed Arabic Text Database) scanned database in a reproducible and controlled environment. As a result, most of the procedures were performed manually in realistic environments with different lighting conditions and various geometric distortions of the images. We used three smartphones with a focus-select feature of the camera hardware to generate a series of images with focal blur depending on the variation in focal distance. The focus distance was decided at random for each image. The location of capture was varied and the lighting conditions varied according to the time of day (morning, noon, evening).

The images that were taken by the smartphone cameras were sorted into two types of distortions, namely single and multiple distortions. For a single distortion, we had different lighting conditions, out-of-focus blur and motion blur. The distorted images were captured at different positions, distances and times of the day. A ground truth data file was then created for each image of our two datasets (the smartphone-captured Arabic printed document dataset and a scanned Arabic printed document dataset), where each file provides;

• A reproduction of the text in a document using the Free Online OCR program in a PDF format of the newspaper.

• The types of distortion in each document and the ID of a captured document.

Thesis 1: We created a new database (with ground truth data) for analysis and recognition purposes owing to the need for a database that covers Arabic magazines and newspapers. The advantage of having such a document database is that it permits us to make a reasonable comparison between old and new results, and because the evaluation of studies recently published were based on random choices, which may different, and this leads to vague estimation.

3 Arabic Document Layout Analysis

Three approaches for Arabic document layout analysis purposes are presented:

3.1 Printed Arabic Newspaper

Here, we present a system for recognizing the logical structure (hierarchical organization) of Arabic newspaper pages. We start by extracting the physical structure, which is essentially based on the RLSA (run length smoothing algorithm) [11], Projection Profile analysis, and CC (connected component) labeling [4, 9, 16]. Logical structure extraction is then performed based on certain rules of sizes and positions of the physical elements extracted earlier, and also on a priori knowledge of certain properties of logical entities (titles, figures, authors, captions, etc.). Lastly, the hierarchical organization of the document is represented as an XML/DTD file generated automatically.
We chose the daily newspaper called Echorouk [7] for our test corpus. The pages of this newspaper have a great variety of possible structures and this makes their treatment and analysis quite difficult.
We commence with an upward segmentation that starts from the pixels of the image and merges them into CCs. Then the CC features are used for text/non-text segmentation. After, with article detection, we use mixed segmentation based on an analysis using Projection Profile, the RLSA algorithm, and the labeling of CCs. Lastly, we apply a descending segmentation to divide the articles of the page into blocks, the blocks into lines and lines into words.
The logical labels of the various elements of the test images were established manually for each step of the newspaper recognition process. After, we used our system to all test images to label them automatically.

The automatic labeling results of each image were compared with the actual labels (manually set) to determine the recognition rate. In order to test the generality of our system, we attempted to vary the test images, so that they would contain a different number of articles, with different positions, and also contain straight lines, strips and figures.
Table 2 summarizes the average recognition rate for each logical entity. In this ta-

Table 2: Test results.

Label	Recognition score
Page header	99.23%
Footer	89.45%
Figures	90.20%
Black bands	95.70%
Borders	88.55%
Straight horizontal lines	98.87%
Articles	90.03%
Blocks	90.32%
Lines	99.85%
Words	75.08%
Columns	90.28%
Legends	93.10%
Authors	94.16%
Average	91.90%

ble, we see that the system has managed to recognize most of the existing logical entities, which corresponds to a recognition rate of 91.90%.

Table 3: A comparaison with other approach.

Algorithms	Tested on	Extracted label	Recognition score
Proposed method	55 images extracted from (Alshorouk) newspaper pages	13 labels extracted	91.90%
Hadjar and Ingold	50 images extracted from (Alhayat+Annahar) newspaper pages	5 labels extracted	87.93%(Annahar) 93.08% (Alhayat)

When comparing the recognition performance of structures (physical and logical) with that of other studies, it can be seen that the identification and verification results are quite different due to the diversity of the structured pages and the nature of the detected elements. However for evaluation purposes, we chose the one most similar one to ours among the other studies [10], where the detected elements were the same one as in our study (see Table 3).

Thesis 2: We provided an improved approach for the logical/physical labeling of Arabic newspaper pages, where it reached 91.90% for 13 labels, this result is very encouraging considering the previous work results where they reach 90.50% for 5 labels. With this, many problems are solved, such as merged deficient detection for a certain type of frame (a non-closed rectangle). We have a bigger range of labels,

and better enhancement of text/non-text segmentation.

3.2 Title Detection in Printed Arabic Newspaper

Now, we present a new text-line detection method for complex-structured documents, where the detected text is treated as a title or subtitle and each page contains many titles corresponding to the article number. Commencing with the preprocessing step using the Otsu binarization method [13], we then use the same formulas that were applied in a previous study (Printed Arabic newspaper), with constraints on the size of the CCs, the ratio of height and width, and the density of black pixels in the CC.
The detection of the titles is done by taking into account the fact that not just the height of the titles is important, but also the number of pixels in each component and its corresponding position. Horizontal RLSA [12] is then applied to the resulting image of the preceding step to remove spaces between words of the same line of text and Vertical RLSA is used to connect the diacritic marks to the corresponding words. Afterwards, the subtitle detection was done using two criteria; the size of the CC of the previous step and its position relative to the main titles, where the title/ subtitle conditions were found by using geometrical features of the CCs.

Table 4: Test results.

Font Type	Title extraction	Subtitle extraction
AL-Quds	98.18 %	97.96 %
AxTManal	98.15 %	98.23 %
Beirut	/	98.87 %
AL-Quds Bold	98.45 %	98.17 %
Kacstone	97.56 %	97.55 %
Alshrek Titles	97.78 %	/
Total	98.02 %	98.15 %

The algorithm was tested on three hundred scanned pages at 300 dpi got from the PATD (The databse from Section 2). The algorithm gives excellent scores, which may be as high as 98.02% for titles, and 98.15% for subtitles.
Table 4 lists the results obtained during the testing process with various font types, styles, and sizes. Here, we handled the problem of distinguishing text size. The results presented are superior to those of existing algorithms that perform the same task (see Table 5).

Thesis 3: We applied a new approach for title/subtitle detection in Arabic complex structure documents, without requiring an analysis of all the elements that exist on the page. Also, the proposed method can improve the article detection step, hence improve the document layout analysis process.

Table 5: A comparaison with other approach.

Algorithms	Tested on	Segmentation level	Recognition score
Proposed method	complexe structure	line segmentation (title/subtitle)	98.08%
Ibrahim	simple structure	line segmentation	97.8%
Soujanya et al.	simple structure	line segmentation	98%
Ayesh et al.	simple structure	line segmentation	99%

This approach reached 98.08% in complex structure, which is considered better to those of existing algorithms that perform the similar task.

3.3 Smartphone-captured Arabic Newspaper Analysis

In this section, a method for Arabic DLA using CNN is proposed. We commence with an Arabic document image as input for text/non-text classification and logical structure recognition [5]; The input is corrected and improved by sharpness/ smoothing filters, adaptive thresholding, morphological operations. The resulting image is transformed into CCs via a pixel connectivity technique, using Adaptive RLSA to reduce the numbers of CC, then we transform each CC into a patch (a small image covers the CC and its surrounding pixels) by a cropping technique, where each patch will be classified into six regions. These identified and localized regions are merged using a CNN score along with geometric features for label extraction. Finally, each article is segmented into blocks and lines by using a projection profile analysis.

Table 6: The performance of the proposed method on different newspapers.

LABEL/ Newspaper	**Text**	**Title**	**Legend**	**Author**	**Figure**	**Table**
Akhersaa	0.901	0.997	0.981	0.972	0.985	0.993
Alsharek	0.941	0.963	0.963	0.963	0.898	/
Alhadaf	0.759	0.780	0.905	0.916	0.732	/
Alnahar	0.915	0.921	0.969	0.945	0.903	/
Alshourouk	0.891	0.923	0.972	0.961	0.900	/
Alshorouk almisri	0.864	0.841	0.952	0.948	0.839	0.941
Alriad	0.853	0.889	0.971	0.973	0.845	0.959
Alsharek al-Awsat	0.851	0.894	0.958	0.952	0.832	/
Aswaq Qatar	0.798	0.761	0.915	0.909	0.747	0.967
Alayam	0.864	0.882	0.952	0.948	0.849	0.965
Total	**0.864**	**0.882**	**0.952**	**0.948**	**0.849**	**0.965**

To validate our method, all the algorithms utilized were implemented and the experiments were performed using JAVA and a computer system that had 14 GB RAM, INTEL (R) Xeon(R) CPU E550 @ 2.53 GHz with a Windows 7 operating system, with the Deeplearning4J (DL4J) framework and VGG-16 model. The system

was evaluated on 120 smartphone-captured document images taken from ten Arabic newspapers (The PATD databse from Section 2). The performance was measured based on a patch-based classification scheme (see Table 6). Here, our results cannot be directly compared to the previous studies because, to the best of our knowledge, none of the methods classifed these ten labels and because the test set used was not the same as that in the related papers (we used ten different newspapers). However, we should mention that some of our test images were extracted from the same newspapers (Alshourouk, Alsharek, and Alriad) that were used in the previous studies [6], so a comparison will provide an indication of how well our method actually works (see Table 7).

Table 7: A comparison of the performance of different approaches.

Algorithms	**Tested on**	**Used method**	**Recognition rate**
Previous method (Subsection 3.1)	55 images from one PDF newspaper	Mixed method	91.90%
Ibrahim et al	40 images collected from three PDF newspapers	CNN based on zone classification CNN based on patch classification	74.50% (2 classes) 91.00% (2 classes)
Proposed method	120 images collected from ten smartphone-captured newspapers	CNN based on patch classification/geometric features	92.06%

Here, we presented a simple and flexible machine learning-based method for an Arabic DLA in smartphone-captured conditions. Several tests were conducted to evaluate the performance of our system and the results we got are encouraging.

Thesis 4: We created an improved approach for Arabic layout analysis under varying conditions, where it reached 92.06% applied on 10 types of newspapers, this result is very encouraging considering the previous work results (they reached 91.00% applied on 3 types of newspapers).
We suggested new ideas for ARLSA implementation and showed that a pre-trained model can be effective for the extraction of new classes (document labels).

4 Handwritten Arabic Text Line Segmentation

Now, we describe the method for Arabic handwritten text line segmentation (word spotting), which is based on a self-organizing feature map (Kohonen map) [15], using CC labeling and the Run-length smoothing algorithm. The input of this method is a handwritten Arabic text image and the output is its segmentation result represented by extracted words.
The evaluation was carried out using the AHDB database [14], and we focused

on two types of errors, namely the spacing and overlapping error and the validity method (see Table 8).

Table 8: Test results for Arabic handwritten word detection.

Image No.	No.words in the line	No.Wrong Seg. words	Correct Seg. rate
1	14	4	71.42 %
2	15	1	93.34 %
3	14	0	100.00 %
4	19	5	73.48 %
5	13	0	100.00 %
6	20	4	80.00 %
7	18	2	88.89 %
.	.	.	.
.	.	.	.
300	17	1	94.11 %
Total	5400	672	87.54%

With this method, we partially solved the spacing and overlapping problems using huge amounts of information in the training data file which contains connected components features obtained from a wide variety of text word positions and writing shapes. However, it will not be effective when the two words have no distance between each other either in horizontal or in vertical segmentation and it will be treated as one word. In Table 9, results were presented along with those from other researchers working on handwritten Arabic texts, which ware got from previously published studies using the same dataset (AHDB), and other databases. We notice that our results are quite good compared to those using the other methods described in the literature, but due to the diversity of data (e.g. some of the databases have skewed data), we cannot guarantee that this method will be suitable for every single dataset.

Experiments were performed by us to demonstrate the effciency of our method. We chose a dataset that had several types of paragraphs, taken from several writers and scripts. We showed that our method is able to handle most situations such as irregular spaces and distinguish between the words.

Thesis 5: We presented an improved approach for handwritten Arabic word detection. We offered new ideas for text segmentation and showed that this approach is quite effective irrespective of the particular writing style.
The result of this approach reached 87.54 % which support our initial hypothesis and they are superior to those that tested Arabic text images on the same dataset.

Table 9: Comparison with other approaches.

Method	Measures	Tested on	Based on	Correct rate
Jawad et al. [8]	Within-word and between-word gaps	200 images extracted from the IFN/ENIT database	Bayesian criteria	85 %
Ayman et al. [1]	Within-word and between-word gaps	25 images extracted from the AHDB database	Kmeans	84.8 %
Al-Dmour et al. [2]	Within-word and between-word gaps + CC lengths	35 images extracted from the AHDB database	Kohonen Neural Network	86.3 %
Belabiod et al. [3]	batch features	200 line images extracted from the KHATT databse	CNN+ BLSTM+ CTC	80.1 %
Proposed method	CC features	300 line images extracted from the AHDB database	Kohonen Neural Network	87.54 %

Conclusions and Future Directions

This dissertation describes various studies on Arabic document and text analysis, and here we successfully enhanced and improved many methods for Arabic document analysis and Arabic handwritten word detection, and we created a new database for research purposes.
In spite of all this work, many experiments still remain for the future. In the following list we suggest some possible future research directions:

- We can create an automatic system that can extract not only the structures and labels, but also the content itself for indexing and searching purposes; then the exploration could be done by article-title, author or legend text for the extraction of the required information without needing to search for archive material by hand;
- We can improve the method of Arabic handwritten text word detection to cover touching words situations, and extract the words from images that have many strange shapes and lines along with actual words hastily written as scribble;
- We can generalize the word detection method to cover the datasets that have skewed images and suchlike.

Publications on Which the Thesis is Based

(1) Thesis 1: H.Bouressace and J.Csirik, Printed Arabic Text Database for Automatic Recognition Systems, 5th International Conference on Computer and Technology Applications, pp.107-111, 2019.
(2) Thesis 2: H.Bouressace and J.Csirik, Recognition of the logical structure of Arabic newspaper pages, 21st International Conference on Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence (Springer), Vol.11107, No.3, pp.251-258, 2018.
(3) Thesis 3: Hassina Bouressace, Title Segmentation in Arabic Document Pages, Journal of WSCG, pp.45-50, 2019.
(4) Thesis 4: H.Bouressace and J.Csirik, A Convolutional Neural Network for Arabic Document Analysis. IEEE 18th International Symposium on Signal Processing and Information Technology (ISSPIT), pp.1-6, 2019.
(5) Thesis 5: H.Bouressace and J.Csirik, A Self-Organizing Feature Map for Arabic Word Extraction. 22nd International Conference on Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence (Springer), Vol.11697, pp.127-136, 2019.

Co-author's declaration

I hereby certify that I am familiar with the thesis of the PhD applicant Ms Hassina Bouressace entitled ,,The application of new methods for offline recognition in printed Arabic documents.
Regarding our jointly published results that form part of this PhD dissertation, I declare the followings:

The applicant's contribution was dominant in obtaining the results in the following publications:
1. H.Bouressace and J.Csirik, Recognition of the logical structure of Arabic newspaper, 21st International Conference on Text, Speech and Dialogue, Lecture Notes in Articial Intelligence (Springer), Vol.11107, No.3, pp.251-258, 2018.
2. H.Bouressace and J.Csirik, Printed Arabic Text Database for Automatic Recognition Systems, 5th International Conference on Computer and Technology Applications, pp.107-111, 2019.
3. H.Bouressace and J.Csirik, A Self-Organizing Feature Map for Arabic Word Extraction. 22nd International Conference on Text, Speech and Dialogue, Lecture Notes in Articial Intelligence (Springer), Vol.11697, pp.127-136, 2019.
4. H.Bouressace and J.Csirik, A Convolutional Neural Network for Arabic Document Analysis. 18th Symposium on Signal Processing and Information Technology (ISSPIT), IEEE, pp.1-6, 2019.

and I did not and will not use these results in getting an academic research degree.

.... March 2020 Dr.János Csirik

Bibliography

[1] A. Al-Dmour and F. Fraij. Segmenting arabic handwritten documents into text lines and words. International Journal of Advancements in Computing Technology (IJACT), 6(3):109–119, 2014.

[2] A. Al-Dmour and R.A. Zitar. Word extraction from Arabic handwritten documents based on statistical measures. International Review On Computer and Software (IRECOS) 11, 2016.

[3] A. Belabiod and A. Belaïd. Line and Word Segmentation of Arabic handwritten documents using Neural Networks, University of Lorraine, 2018.

[4] A. Simon, J.C. Pret, and A.P. Johnson. A Fast Algorithm for Bottom-Up Document. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.19, No.3, pp. 273-277, 1997.

[5] B. Gatos, S. Mantzarisl, and A. Antonacopoulos, First International Newspaper Segmentation Contest. Proceedings of the 6th International Conference on Document Analysis and Recognition, pp.1190-1194, 2001.

[6] I.M. Amer, S. Hamdy, and M.G.M. Mostafa. Deep Arabic document layout analysis. In Proceedings of the IEEE Eighth International Conference on Intelligent Computing and Information Systems, pp.224-231, 2017.

[7] https://www.echoroukonline.com/newspaper/echorouk-yawmi/

[8] J. H AlKhateeb, J. Jiang, J. Ren, and S. Ipson. Interactive knowledge discovery for baseline estimation and word segmentation in handwritten Arabic text. Recent Advances in Technologies, Maurizio A Strangio (Ed.), 2009.

[9] J. Liang, J. Ha, R.M. Haralick, and I.T. Phillips. Document Layout Structure Extraction Using Bounding Boxes of Different Entities. in Proc. of 3rd IEEE Workshop on Applications of Computer Vision, pp. 278-283, 1996.

[10] K. Hadjar, R. Ingold. Arabic Newspaper Page Segmentation,7th International Conference on Document Analysis and Recognition, pp.895-899, 2003.

[11] K.Y. Wong, R.G. Casey and F.M. Wahl, Document analysis system. IBM Journal of Research and Development, Vol.26, pp.647-656, 1982.

[12] L. OGorman and R. Kasturi. Executive briefing: document image analysis, IEEE Computer Society Press, ISBN 0-8186-7802-X, 1997.

[13] N. Otsu. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. 9(1), pp.62-66, 1979.

[14] S. Al-Ma'adeed, D. Elliman, and C.A. Higgins. A database for Arabic handwritten text recognition research. In: Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition. pp.485-489, 2002.

[15] T. Kohonen. The self-organizing map. Proceedings of the IEEE, 78(9), pp.1464-1480, 1990.

[16] Y. Pan, Q. Zhao, and S. Kamata. Document Layout Analysis and Reading Order Determination for a Reading Robot. 10th IEEE Region Conference, pp. 1607-1612, 2010.