



***Developing an Algorithm for Arabic Document Image
Analysis***

Thesis submitted to the Department of Computer Science
Faculty of Computer and Information Sciences
Ain Shams University

*In partial fulfillment of the requirements for the degree
Of Master in Computer and Information Sciences*

By

Ibrahim Mohammed Ali Amer

B.Sc. in Computer and Information Sciences (2014)
Ain Shams University – Cairo

Under the supervision of

Prof. Dr. Mostafa Gadal-Haqq M. Mostafa

Professor of Computer Science
Faculty of Computer and Information Sciences
Ain Shams University

Dr. Salma Hamdy Mohamed

Lecturer, Computer Science Department
Faculty of Computer and Information Sciences
Ain Shams University
Cairo-2017

To my dear parents

Acknowledgements

I acknowledge my deep gratitude to ALLAH for providing me with the strength to complete this work on a level that I hope will please the reader.

I wish to express my sincere and heartfelt appreciation to my supervisor; Prof. Mostafa Gadal-Haqq for his supervision, resourceful ideas and his expert suggestions that helped in enhancing my work.

Special thanks are due to my supervisor Dr. Salma Hamdy for her continuous guidance and encouragement in achieving this work.

Without their assistance and dedicated involvement in every step throughout the process, this thesis would have never been accomplished. I would like to thank you very much for your support and understanding.

Finally, with all my appreciation and love that no words can describe, my deepest gratitude goes to my family for their love, prayers and endless support; my dear parents and my friends who supported me and endured a lot to lead me to success in my life and my career!

Abstract

Document layout analysis (DLA) is the process of identifying the regions of interest in document image which requires the separation of text regions from non-text ones. DLA is an essential step for Optical Character Recognition systems (OCR), document management systems, document-archiving systems and more. The text of the document fed to the OCR must be extracted first and isolated from images if exist. OCR systems recognize printed or handwritten text images, these images must contain text only and if the document contains text mixed with photos, graphs shapes or halftones; this will result in a negative effect on recognition accuracy. Thus, DLA is a crucial step before OCR. The DLA task is difficult as there is no fixed layout for all documents, but instead, there are several layouts based on the document type: a newspaper, a magazine, a book or a manuscript. There are various approaches for DLA for various different languages, but document layout analysis for Arabic scripts is

much more difficult compared to other languages because of the cursive nature of the Arabic language.

DLA is a method of defining and recognizing the structure of the document or it is the way of categorizing the important regions in the document. It includes separating its text components from non-text to feed them directly to the OCR, identifying the title of the document if any, etc.

Different font types and sizes are also factors that must be considered in document analysis, because if the font type/size provided to the OCR is different from what the OCR expects, the results might not be accurate. Font type and size recognition helps to classify the type of the font being processed so it can be segmented and classified using the proper segmentation and learning methods. Research on font recognition task has a lot of focus recently as they are very important for OCR systems.

Reading order determination is the process of identifying the order of the reading flow of the document (right to left or left to

right) which can be also used to enhance the DLA and produce reliable results.

This research proposes a method for document layout analysis for segmenting, localizing and separation of text regions from non-text regions. The method uses deep convolutional neural networks to classify regions as it achieves the best results for computer vision related tasks. The best results achieved was *Precision* = 0.96, *Recall* = 0.87 and *F-score* = 0.91 for text and non-text segmentation. The system evaluated on 40 document images (10 skewed and 30 non-skewed) collected from Arabic newspapers (Riyadh, Al-Sharq and Al-Youm).

In addition, methods for font type and font size recognition have been proposed. Both methods are based on classification using deep convolutional neural networks as well. For font type recognition, best results achieved was *accuracy* = 98.6%. For font size recognition, the highest accuracy achieved was 99.94% and the lowest accuracy was 95.39%.

List of Publications

1- Ibrahim M. Amer, Salma Hamdy, Mostafa, M. G. Mostafa, “Deep Arabic Document Layout Analysis”. *International conference on Intelligent Computing and Information Systems (ICICS)*. 2017.

2- Ibrahim M. Amer, Salma Hamdy, Mostafa, M. G. Mostafa, “Deep Arabic Font Type And Font Size Recognition”. *International Journal of Computer Applications (IJCA)*. 176(4):1-6, October 2017.

Table of Contents

- Chapter 1 Introduction.....3
 - 1.1. Overview 3
 - 1.2. Motivation..... 6
 - 1.3. Problem Statement..... 6
 - 1.4. Research Objectives..... 7
 - 1.5. Proposed Work 7
 - 1.6. Thesis Organization..... 8
- Chapter 2 Background..... 11
 - 2.1 Overview 11
 - 2.2 Image Thresholding..... 12
 - 2.3 Image Filtering 16
 - 2.4 Skew Estimation and Correction..... 20
 - 2.4.1 Projection profile based skew correction 21
 - 2.4.2 Hough lines based skew correction 24
 - 2.5 Deep Learning 25
 - 2.5.1 Deep learning in image recognition 27
 - 2.5.2 Deep Learning Algorithms..... 27
 - 2.6 Convolutional Neural Networks..... 28
 - 2.6.1 Convolutional layers..... 30
 - 2.6.2 Pooling layer..... 31
 - 2.6.3 Fully connected hidden layer (FC)..... 35

2.6.3 Toy example for CNN [52].....	36
Chapter 3 Related Work.....	51
3.1. Overview.....	51
3.2 Arabic Document Layout Analysis.....	51
3.3 Font Family and Font Size Recognition	54
Chapter 4 Deep Arabic Document Layout Analysis.....	59
4.1. Overview of the Proposed Work.....	59
4.2 Pre-processing.....	61
4.3 Classification	63
4.4 Text Lines and Words Segmentation	68
4.5 Results and Discussion	70
4.5.1 Datasets	70
4.5.2 Results	72
4.6 Summary and Conclusion	77
Chapter 5 Deep Arabic Font Family and Font Size Recognition.....	80
5.1. Overview of the proposed work	80
5.2 Font Family Recognition.....	81
5.2.1 Pre-processing.....	83
5.2.2 Architecture	83
5.3 Font Size Recognition.....	86
5.3.1 Pre-processing and feature extraction	87
5.3.2 Architecture	87
5.4 Results and Discussion	88
5.4.1 Dataset	88
5.4.2 Results	88
5.5 Summary and Conclusion	93
Chapter 6 Conclusion and Future Work.....	95
6.1 Summary of the work	95

6.2 Conclusion..... 96

6.3 Future Work..... 98

Chapter 7 References.....100

List of Figures

Figure 1-1 A document with multi-column format (the document contains text and non-text components).....	5
Figure 1-2 An example of what can be obtained from document layout analysis: structural and functional layouts	5
Figure 1-3 An example of what can be obtained from document layout analysis: structural and functional layouts	5
Figure 2-1 A demonstration of image filtering process with a 3X3 mask [7]	17
Figure 2-2 Pseudo code for image convolution process [7]	17
Figure 2-3 Edge detection: (a) original image, (b) canny edge detector, (c) Sobel filter, (d) Prewitt edge detector.	19
Figure 2-4 An Example for a document image skewed by 30 degrees	21
Figure 2-5 Histograms of (a) the deskewed document and the original skewed document (b)	23
Figure 2-6 A graph shows how deep learning behave better with increasing data compared with older learning algorithms [10].....	26
Figure 2-7 (a) Original image, (b) image after subsampling or pooling layer	32
Figure 2-8 Demonstration for max pooling 2x2 filter and stride 2 (source: CS231n Stanford's course).....	34
Figure 2-9 A demonstration for minimum pooling with 2x2 filter and stride 2 (source: CS231n Stanford's course).....	34
Figure 2-10 Demonstration of a CNN with multiple layers [51].....	35

Figure 2-11 Further demonstration of convolutional filters and its effect on the image [51]	35
Figure 2-12 A toy example of a CNN which classifies "X" and "O" characters	36
Figure 2-13 CNNs must be robust to different affine transformations.....	37
Figure 2-14 The decision between the original image and the image after scaling is hard.	37
Figure 2-15 CNNs can match pieces of the image	38
Figure 2-16 Filters that can be used to detect "X" character features.....	38
Figure 2-17 A demonstration for a filter detecting the upper left side edge of the character 'X'.	39
Figure 2-18 A demonstration for a filter detecting the upper right side edge of the character 'X'.	39
Figure 2-19 A demonstration for a filter detecting the middle edges of the character 'X' edge.....	40
Figure 2-20 A demonstration for a filter detecting the lower right side edge of the character 'X'.	40
Figure 2-21 demonstration of how can each filter detect a part of the image	41
Figure 2-22 A demonstration of the convolution process.....	42
Figure 2-23 The convolution process. Demonstrates a match found.....	42
Figure 2-24 The corresponding feature (activation) map after finding all matches using the convolution process.....	43
Figure 2-25 A demonstration of three feature (activation) maps for three different convolutional windows	43
Figure 2-26 A demonstration of max pooling process.....	44
Figure 2-27 The result of max pooling	45
Figure 2-28 A stack of feature (activation) maps became a stack of smaller maps	45

Figure 2-29 A demonstration of ReLU activation function and its effect on the activation map	46
Figure 2-30 A demonstration of ReLU activation function and its effect on the activation map	46
Figure 2-31 Deep stacking of layers (convolution layers, ReLU activation layers or pooling layers).....	47
Figure 2-32 Fully connected hidden layer and an output layer with a voting score for the network's decision.	47
Figure 2-33 Stacking of fully connected hidden layers	48
Figure 2-34 Stacking of different network layers (convolutional layers, ReLU activation layers, pooling layers and fully connected layers).	48
Figure 4-1 Proposed system architecture	60
Figure 4-2 Original document image (left) and the result after ARLSA (right)	62
Figure 4-3 Original document image (left) and the result after identifying connected components' regions (right).....	62
Figure 4-4 Vertical Projection Profile Histogram.....	69
Figure 4-5 Horizontal Projection Profile Histogram	70
Figure 4-6 APTI database fonts: (A) Andalus, (B) Arabic Trans- parent, (C) AdvertisingBold, (D) Diwani Letter, (E) DecoType Thuluth, (F) Simplified Arabic, (G) Tahoma, (H) Traditional Aatbic, (I) DecoType Naskh, (J) M Unicode Sara	71
Figure 4-7 From left to right: Al-Sharq, Al-Riyadh and Al-Youm newspapers used for training and testing.....	72
Figure 4-8 (a), (d): Original document images. (b), (e): zone based classification. (c), (f): patches based classification (green: text, red: non-text)	75
Figure 4-9 (a), (d): Original document images. (b), (e): zone based classification. (c), (f): patches based classification (green: text, red: non-text)	76

Figure 5-1 Proposed system architecture for font family and font size recognition	81
Figure 5-2 Confusion Matrix of the 4th network (CNN + SVM for all fonts' families and sizes). Labels shown on x and y axes from 0 to 9 represent font families Advertising Bold, Andalus, Arabic Transparent, DecoType Naskh, DecoType Thuluth, Diwani Letter, M Unicod.....	89
Figure 5-3 Font family recognition: a demonstration of the convolutional layers and the hidden fully connected layers of the fourth trained network.	92
Figure 5-4 Font family recognition: visualizing some convolutional filters of the first conv. layer for the fourth trained network.....	92