AIN SHAMS UNIVERSITY
Faculty of Computer and
Information Sciences
Computer Science Department

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

List of Figures