ARABIC DOCUMENT LAYOUT ANALYSIS USING MACHINE LEARNING AND CONNECTED COMPONENTS BASED FEATURES

By

Rana Sobhy Mostafa Saad

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Electronics and Communications Engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2018

ARABIC DOCUMENT LAYOUT ANALYSIS USING MACHINE LEARNING AND CONNECTED COMPONENTS BASED FEATURES

By

Rana Sobhy Mostafa Saad

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Electronics and Communications Engineering

Under the Supervision of

Prof. Dr. Neamt Sayed Abdel Kader

………………………………..

Professor of
Communications and Electrical Engineering
Faculty of Engineering, Cairo University

Prof. Dr. Samia Abdel-Razeq Mashaly

………………………………..

Professor of
Computers and Systems
Electronics Research Institute

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2018

ARABIC DOCUMENT LAYOUT ANALYSIS USING MACHINE LEARNING AND CONNECTED COMPONENTS BASED FEATURES

By

Rana Sobhy Mostafa Saad

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Electronics and Communications Engineering

Approved by the
Examining Committee

____
Prof. Dr. **Neamt Sayed AbdelKader**, Thesis Main Advisor

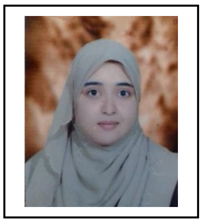
____
Prof. Dr. **Samia Abdel-Razeq Mashaly**, Advisor
Professor, Computers and Systems dept., Electronics Research Institute, Giza, Egypt

____
Prof. Dr. **Mohsen Abdel-Razeq Rashwan**, Internal Examiner

____
Prof. Dr. **Sherif Mahdi Abdo**, External Examiner
Professor and IT department head, Faculty of Computers and Information, Cairo University, Giza, Egypt

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2018

Engineer's Name: Rana Sobhy Mostafa Saad
Date of Birth: 18/1/1990
Nationality: Egyptian
E-mail: rana@eri.sci.eg
Phone: 01145449372
Address: 33 Dokki st.,Giza, Egypt



Registration Date: 1/10/2013
Awarding Date: …./…./2018.
Degree: Master of Science
Department: Electronics and Communications Engineering

Supervisors:
Prof. Neamt Sayed Abdel-Kader
Prof. Samia Abdel-Razeq Mashaly

Examiners:
Prof. Dr. Neamt Abdel-Kader (Thesis main advisor)
Prof.Dr. Samia Abdel-Razeq Mashaly (Advisor)
Professor, Computers and Systems dept., Electronics Research Institute, Giza, Egypt
Prof. Dr. Mohsen Abdel-Razeq Rashwan (Internal examiner)
Prof. Dr. Sherif Mahdi Abdo (External examiner)
Professor and IT department head, Faculty of Computers and Information, Cairo University, Giza, Egypt

Title of Thesis:

Arabic Document Layout Analysis Using Machine Learning And Connected Components Based Features

Key Words:
Page segmentation, Document Layout Analysis, Arabic dataset

Summary:
Document Layout Analysis (DLA) is a key preprocessing stage for optical character recognition (OCR). It locates and defines text and non-text regions of a document image. Arabic DLA is less addressed compared to other languages due to the lack of appropriate publicly available research datasets. A full pipeline of DLA procedure is composed of several stages: Input document Preprocessing, Document Physical layout Analysis (PLA), Document Logical Layout Analysis (LLA), and document analysis output representation.

In this thesis, CCs geometric features are used to represent the Arabic document images These CCs features are classified by means of Support Vector Machines (SVM) and Random Forests (RF) classifiers into text and non-text components to perform PLA for scanned Arabic book pages.

Experiments on BCE-v1, and other researcher's datasets showed remarkable performance of both the SVM and RF based solutions. Comparing to other classical and state-of-the-art systems showed much strength to the proposed system and promise further application to wider problem domains.

.

Acknowledgments

I would like to express special appreciation and thanks to my supervisors: Prof. Dr. Neamt Abdel Kader and Prof. Dr. Samia Mashaly for constant support, patience, and guidance.

I would especially thank Dr. Randa Elanwar for being my mentor in the Electronics Research Institute for her continuous support, conversations, and comments throughout this thesis.

Also, I must express my profound gratitude to my parents, my husband, my mother-in-law , my children Mohamed and Mariam for their unfailing support and continuous encouragement. This work would not have been possible without them.

Finally, I would also like to thank all of my colleagues at Electronics Research Institute who supported me at times of confusion, and urged me to strive towards my goal.

Table of Contents

Acknowledgments .... **i**

Table of Contents .... **ii**

List of Tables .... **iv**

List of Figures .... **vi**

List of Abbreviations .... **viii**

Abstract .... **x**

Chapter 1 : Introduction **1**

1.1. THESIS OBJECTIVE AND CONTRIBUTION 2
1.2. THESIS ORGANIZATION 2

Chapter 2 : Literature Review **3**

2.1. INTRODUCTION 3
2.2. ZONE BASED PLA SYSTEMS 4
2.2.1. Variable window size systems 4
2.2.1.1. Rule based segmentation approaches 4
2.2.1.2. Machine learning-based approaches 7
2.2.2. Fixed window size systems 9
2.2.2.1. Traditional machine learning-based approaches 9
2.2.2.2. Deep learning-based methods 10
2.2.3. Discussion 16
2.3. PIXEL BASED PLA SYSTEMS 20
2.3.1. Classical pixel-based segmentation algorithms: 20
2.3.2. Superpixel-based segmentation algorithms 20
2.3.3. Deep learning-based segmentation algorithms 20
2.3.4. Discussion 27
2.4. CONNECTED COMPONENTS BASED PLA SYSTEM 29
2.4.1. Rule based segmentation approaches 29
2.4.2. Machine learning-based approaches 32
2.4.3. Discussion 36
2.5. CONCLUSION 36

Chapter 3 Dataset Collection and Preparation **38**

3.1. SAMPLES SELECTION AND ACQUISITION 38
.3.2 GROUND TRUTH TOOLS AND ANNOTATION 39
3.2.1. MS-paint and Gimp: 41
3.2.2. Pixlabeler,2009 42
3.2.3. GEDI, 2010 43
3.2.4. Aletheia, 2011 44
3.2.5. DIVIDIA, 2015 46
3.2.6. TRUEVIZ, 2003 47
3.3. COMPARING ANNOTATION TOOLS 48
3.3.1. MS-Paint 49

3.3.2. Pixlabeler 50
3.3.3. Aletheia 51
3.3.4. GEDI 53
3.3.5. DIVADIA 54
3.3.6. TRUEVIZ 56
3.3.7. Conclusions 57

Chapter 4 Proposed System 60

4.1. INTRODUCTION 60
4.2. PREPROCESSING 61
4.3. FEATURE EXTRACTION 63
4.4. CLASSIFICATION 66
4.4.1. Support Vector Machines 66
4.4.2. Random Forests 68
4.5. POST-PROCESSING 69
4.6. OUTPUT REPRESENTATION 70

Chapter 5 Experiments and Results 71

5.1. SVM CLASSIFIER EXPERIMENTS (TRAINING AND EVALUATION) 71
5.1.1. Model training and parameter tuning using validation dataset 71
5.1.2. System evaluation using PLA system and SVM classifier for test dataset 73
5.1.3. Model training and parameter tuning using validation dataset (Preprocessing stage modification) 75
5.1.4. Effect of KNN on the SVM best model results 75
5.1.5. SVM-based system evaluation using test dataset using 16NN: 77
5.2. RF CLASSIFIER EXPERIMENTS (TRAINING AND EVALUATION) 79
5.2.1. Model training and parameter tuning using validation dataset 79
5.2.2. Model training and parameter tuning using validation dataset (Preprocessing stage modification) 81
5.2.3. Effect of KNN on the RF best model results 82
5.2.4. RF-based system evaluation using test dataset 82
5.3. SYSTEM EVALUATION USING OTHER DATASETS 84
5.3.1. Evaluation on Connected Components level 84
5.3.1.1. RDI Dataset 84
5.3.1.2. Hesham *et al.* private dataset: 89
5.3.1.3. PLA-SAB challenge (ASAR 2018) dataset: 93
5.3.2. Evaluation on block and pixel levels 101
5.3.3. Results conclusion of different datasets for block and pixel levels evaluation... 104
5.4. APPLYING OTHER SYSTEMS TO OUR BCE-V1-ARABIC DATASET: 106
5.4.1. RLSA Performance Evaluation 106
5.4.2. Zone and pixel evaluation For RDI Clever Page System 107
5.4.3. Zone and pixel evaluation for ECDP-system 107
5.5. SYSTEM OUTPUT FILES 109

Discussion and Conclusions 110

References 111

List of Tables

Chapter 2:
Table 2.1 Chen *et al.* [3] dataset division 23
Table 2.2 Wei *et al.* classification accuracy results[32] 13
Table 2.3 Chen *et al.* at [33] superpixel representation performance comparison 13
Table 2.4 Two CNN structures used by Pastor-Pellicer *et al.*[42] 15
Table 2.5 Zone based PLA systems 17
Table 2.6 Nandedkar *et al.* [52]Classification results 23
Table 2.7 Bukhari *et al.* [54]results on UW-II and ICDAR2009 25
Table 2.8 Pixel based PLA systems 28
Table 2.9 CCs-based PLA system algorithms 37
Chapter 3:
Table 3.1 Chen *et al.* evaluation for GT tools 47
Table 3.2 The average document image annotation time (in minutes) per annotation tool for every group of test set documents 57
Table 3.3 Zoning Properties offered by each tools 58
Table 3.4 Labeling Capabilities offered by each tools 59
Chapter 5:
Table 5.1 Coarse Tuning Results over Validation dataset 72
Table 5.2 Medium Tuning Results over Validation dataset 72
Table 5.3 SVM Fine Tuning Results over Validation dataset 73
Table 5.4 Finer SVM tuning over Validation dataset 73
Table 5.5 Medium Tuning Results over Validation dataset (with morphological preprocessing) 75
Table 5.6 Fine Tuning Results over Validation dataset (with morphological preprocessing) 75
Table 5.7 Per-document results for test dataset using best SVM model and 16NN post processing 78
Table 5.8 Validation data results of different features numbers combination with 100 trees 79
Table 5.9 Validation data results of different number of trees 80
Table 5.10Validation data results of different features numbers combination with 100 trees (with morphological preprocessing) 81
Table 5.11 Validation data results of different number of trees (with morphological preprocessing) 81
Table 5.12 Per-document results for test dataset using RF best model and morphological preprocessing 83
Table 5.13 SVM and Random Forests based solutions results for RDI Dataset. 85
Table 5.14 Per-document results (Classification Accuracy) for SVM and RF based solutions for RDI dataset 86
Table 5.15 SVM-based and RF-based solutions results for Hesham et al. dataset 90
Table 5.16 Per-document results (Classification accuracy) for SVM based and RF based solutions for Hesham et al. dataset 91
Table 5.17 SVM-based and RF-based solutions results for ASAR 2018, Set-A 96

Table 5.18 Per-document results (Classification Accuracy) for SVM based and RF based solutions for ASAR 2018, Set-A 96
Table 5.19 SVM-based and RF-based solutions results for ASAR 2018 Set-B 97
Table 5.20 Per-document results (Classification Accuracy) for SVM based and RF based solutions for ASAR 2018 Set-B 97
Table 5.21 SVM-based and RF-based solutions results for ASAR 2018 Set-C 108
Table 5.22 Per-document results (Classification Accuracy) for SVM based and RF based solutions for ASAR 2018, Set-C 98
Table 5.23 SVM- RF Segmentation evaluation For Different Datasets 103
Table 5.24SVM- RF Block and Pixel evaluation For Different Datasets 104
Table 5.25 SVM- Reclassification evaluation (Accuracy %) For Different Datasets . 106
Table 5.26 RLSA segmented blocks H and Rm parameters calculated over BCE training set 107
Table 5.27 Summary of segmentation evaluation For Different Systems 108
Table 5.28 Summary of Block and Pixel evaluation For Different Systems 108

List of Figures

Chapter 2:

Fig. 2.1 the left image is the original, and the right is the x-y cut resulting image [13].. 5
Fig. 2.2 white space analysis results [14] 5
Fig. 2.3 text line model with parallel base line and descender line. 6
Fig. 2.4 (Wang *et al.*) Nine classes zones [19] 8
Fig. 2.5 DSE pixel arrangements 9
Fig. 2.6 ECDP system Block diagram [29].. 11
Fig. 2.7 AE structure of Chen *et al*. PLA system [35].. 12
Fig. 2.8 the main body area MBA of a text line [47],.. 15
Fig. 2.9 The original image (after binarization and canny edge extraction) on the left and RLSA results on the right [50]... 21
Fig. 2.10 Nandedkar *et al.* proposed work results [57].. 22
Fig. 2.11 the comparison between Bloomberg's method results (upper row) and Bukhari *et al.* [59] improved results (lower row) [59] 24
Fig. 2.12 DeepLayout system stages.. 27
Fig. 2.13 (a) Part of original image (b) Nearest neighbor vectors overlaid on the image (c) Nearest neighbor vectors are shown [86]. 29
Fig. 2.14 Different threshold values effect on Voronoi regions [87]............................ 30
Fig. 2.15 Historical image system segmentation results of Bukhari *et al.* Work [5]..... 34
Fig. 2.16 Graphical components samples [100]... 34
Fig. 2.17 CCs merging process [2]. 35
Fig. 2.18 DSE corresponding values [2]... 35
Fig. 3.1 Samples from BCE-Arabic-v1 Dataset.. 40
Fig. 3.2 Pixel based representation [105].. 41
Fig. 3.3 Pixlabeler GUI 42
Fig. 3.4 GEDI interface tool [73]... 43
Fig. 3.5 GEDI: Zone attributes settings 43
Fig. 3.6 Aletheia image Enhancement tool.. 44
Fig. 3.7 Aletheia attributes settings.. 45
Fig. 3.8 Aletheia reading order determination 46
Fig. 3.9 Different labels offered by DIVIDIA [37] 46
Fig. 3.10 TRUEVIZ Interface.. 48
Fig. 3.11 MS-Paint color codes according to the zone type 49
Fig. 3.12 Different Zoning level using MS-Paint. a) block level, b) Text-Line level, c) Word level... 50
Fig. 3.13 Pixlabeler zoning for different layouts: text and image, text and charts, and multi column documents.. 50
Fig. 3.14 Pixlabeler XML output format 51
Fig. 3.15 Aletheia Zoning: (a) Automatic shrinking (smearing), irregular shape outline, (b) Rectangle Zoning, sharp edges.. 51
Fig. 3.16 Example of Aletheia XML output.. 52
Fig. 3.17 GEDI tool zone setting 53
Fig. 3.18 GEDI tool attributes Setting 53
Fig. 3.19 Example of GEDI XML output.. 54
Fig. 3.20 Examples of GEDI image output.. 54
Fig. 3.21 zoning using DIVADIA tool 55
Fig. 3.22 DIVADIA output XML 55

Fig. 3.23 TRUEVIZ tool interface 56
Fig. 3.24 TRUEVIZ XML output 56
Fig. 4.1 Different CCs composition within Arabic Words. 60
Fig. 4.2 different position of dots in Arabic script 60
Fig. 4.3 Diacritics position to the character stroke 60
Fig. 4.4 Proposed Arabic Physical Layout Analysis system block diagram 61
Fig. 4.5 Binarization effect on input image with line drawings. The non-text part has large amount of background pixels showing in between the drawing pen strokes. After binarization the pen strokes are broken and small CCs are generated. 62
Fig. 4.6 Morphological preprocessing to the binarized image 63
Fig. 4.7 RF Voting scheme to predict the relevant class 68
Fig. 5.1 SVM Sample results for test dataset (First test) 74
Fig. 5.2 Misclassified text CC from large size font 76
Fig. 5.3 Misclassified small size non-text CC 77
Fig. 5.4 RF errors for validation dataset, the red errors are corresponding to the wrongly classified textual components. 80
Fig. 5.5 Errors persisting despite morphological operations. a) Red colored components are due to misclassified text components, b) Purple colored components are due to misclassified non-text components. 82
Fig. 5.6 RDI dataset sample with text-only content 84
Fig. 5.7 RDI dataset sample with mixed text and images contents 85
Fig. 5.8 Font size variation affect the classification results for RDI dataset. Red colored errors are for misclassified text components, and purple colored errors are for misclassified non-text components 88
Fig. 5.9 "Image 78" a) errors in SVM (accuracy 51.6%), b) errors in RF (accuracy 51.9 %) Same errors appear with both solutions due to small size non-text CC. 88
Fig. 5.10 Samples from Hesham *et al*. self-collected dataset 89
Fig. 5.11 Example of random misclassified text components with (a) RF solution results, (b) SVM solution results on Hesham *et al.* dataset sample 'Book 6'. 90
Fig. 5.12 "Book-41" results as an example of errors due to binarization for both solutions. a) Original, b) RF results, and c) SVM results. 92
Fig. 5.13 Effect of binarization techniques on the document image quality, a) Original, b) Otsu, and c) Sauvola. 92
Fig. 5.14 Errors due to small non-text CCs (Purble colored), large text CCs (Red colored), and font size variation. 93
Fig. 5.15 Examples from ASAR 2018, Set-A [137] 94
Fig. 5.16 Examples from ASAR 2018, Set-B [137] 94
Fig. 5.17 Examples from ASAR 2018, Set-C [137] 95
Fig. 5.18 a) "KIC Documents 2016-02-19 13_Page_06" original image, b) errors due to small non-textual components in Set-A of ASAR 2018. 99
Fig. 5.19 “KIC Documents 2016-02-18 6_Page_03”a) original, b) errors due to non-text components spanning the page layout horizontally eliminated in preprocessing. 100
Fig. 5.20 Examples for large non-textual CCs error of ASAR 2018, Set-C 100
Fig. 5.21 Example of the proposed systems XML output file 101
Fig. 5.22 RLSA Segmented blocks generation by: horizontal, vertical, and smoothed thresholds 106
Fig. 5.23 RDI Clever Page system results on BCE-v1-Arabic test set 108
Fig. 5.24 Example of the proposed system recognized textual zones (a) the Original image, (b) the OCR output text. 109

List of Abbreviations

AGA	Adapted Genetic Algorithm
ASAR	Arabic And Script Driven Analysis And Recognition
ASFS	Adapted Sequential Forward Selection
BB	Bounding Box
BCE	Boston University-Cairo University-Electronics Research Institute
CAE	Convolutional Auto Encoder
CC	Connected Components
CNN	Convolutional Neural Networks
CRF	Conditional Random Field
CS	Correct-Segmentation
DIVA	Document, Image And Voice Analysis Group
DLA	Document Layout Analysis
DPI	Dot Per Inch
DSE	Document Structure Elements
DSECN	Document Structure Element Characteristic Number
ECDP	Ensemble Based Classification Of Document Patches
ER	Extraction Rate
FA	False Alarm
FV	Feature Vector
GEDI	Groundtruthing Environment For Document Images
GLCM	Grey Level Co-Occurrence Matrix
GUI	Graphical User Interface
HOG	Histogram Of Gradient
ICDAR	International Conference On Document Analysis And Recognition
IHP	Islamic Heritage Project
KNN	K-Nearest Neighbors
LLA	Logical Layout Analysis
MBA	Main Body Area
MDI	Multi Document Interface
ML	Machine Learning
MLP	Multi-Layer Perceptron
MR	Misclassification Rates
MSE	Missed-Segmentation Error
NN	Neural Networks
OSE	Over-Segmentation Error
PAGE	Page Analysis And Ground Truth Elements
PDF	Portable Document Format
PLA	Physical Layout Analysis
RBF	Radial Basis Function
RF	Random Forests
RLSA	Run Length Smearing Algorithm
ROI	Region Of Interest
SBS	Sequential Backward Selection
SFS	Sequential Forward Selection
SFTGS	Spectral Filtering Text-Graphics Separation Algorithm
SILC	Simple Linear Iterative Clustering
SOM	Self-Organizing Map

SVM	Support Vector Machines
SW	Stroke Width
USE	Under-Segmentation Error
UW	University Of Washington
WEKA	Waikato Environment For Knowledge Analysis
XML	Extensible Markup Language.

Abstract

Document Layout Analysis (DLA) is a key preprocessing stage for optical character recognition (OCR). It locates and defines text and non-text regions of a document image. Arabic DLA is less addressed compared to other languages due to the lack of appropriate publicly available research datasets.

A full pipeline of DLA procedure is composed of several stages: Input document Preprocessing, Document Physical layout Analysis (PLA), Document Logical Layout Analysis (LLA), and document analysis output representation. Preprocessing includes several image enhancement processes: binarization, noise removal, skew detection and correction, etc. PLA decomposes the document image into meaningful homogenous regions and then identify the type of their content as text or non-text. LLA identifies the functional role of textual regions within the document as header, footer, main body text, page number, and etc. The output representation is how the analysis resulting information is arranged and transcribed for text recognition systems to use.

In literature, PLA approaches are either: top-down segmentation, bottom-up segmentation, or hybrid segmentation. The top-down approaches perform recursive divisions of the top level (document page) until a desired region of interest (e.g. paragraph, textline) is reached. On the contrary, bottom-up approaches cluster the document's small primitives (pixels, Connected Components (CCs), or image patches) to form the Region of Interest (ROI). Both approaches could be implemented using: rule-based or learning-based algorithms.

Bottom-up approach achieves high performance systems regardless the high computational cost. Using CCs with bottom up approach is a better trade-off compared to pixels.

In this thesis, CCs geometric features are used to represent the Arabic document images. These CCs features are classified by means of Support Vector Machines (SVM) and Random Forests (RF) classifiers into text and non-text components to perform PLA for scanned Arabic book pages.

All classifier parameters tuning and testing experiments are performed on BCE-v1 Arabic dataset [1], the first publicly-available Arabic dataset of scanned book pages that have been collected to support DLA research. Experiments on BCE-v1, and other researcher's datasets showed remarkable performance of both the SVM and RF based solutions. Comparing the proposed system results to other classical and state-of-the-art systems showed much strength to the proposed system and promise further application to wider problem domains.

The results over BCE-Arabic on CCs level for SVM based system show 98.8% classification accuracy while other private datasets as RDI, Hesham et al., ASAR2018 Set-A, ASAR2018 Set-B, and ASAR2018 Set-C are 96.5%, 90.8%, 78.6%, 83.8%, and 97.19% respectively. However RF has 91%, 92.8%, 75.5%, 80.89%, 95.46 % classification accuracy for RDI, Hesham et al., ASAR2018 Set-A, ASAR2018 Set-B, and ASAR2018 Set-C respectively.

Chapter 1 : **Introduction**

Scanned documents are considered an important source of digital information, either these documents are resulting from daily data exchange production or resulting from projects of preserving ancient inheritance. These scanned documents are basically images of text rather than accessible text files; therefore there is a need to make their content accessible and editable. Accessibility requires extracting the elementary components of the document image and identifying their different types either: (1) graphics like diagrams, sketches, charts, maps, etc. , (2) images like photographs, halftones, paintings, etc. , (3) structured text like: tables or body text which could be recognized later by an optical character recognition (OCR) system to make it editable. This process is defined as document layout analysis (DLA).

Document layout analysis is a prerequisite stage to several information extraction procedures like OCR. Consequently, if a DLA process failed, OCR receives badly segmented text which leads to inaccurate recognition results in addition to meaningless symbols. Therefore, accurate DLA procedure is highly demanded

The generation of editable and searchable document content is necessary for several applications such as automatic indexing and retrieval, automatic summarization, automatic translation, Table of Content (TOC) generation, text to speech conversion, etc. and for visually impaired Assistive technology.

Since two decades, most DLA publications address English and Latin-script derived languages. Some attention is given to other languages like Chinese, Hindi, Urdu, Devanagari, Tamil, and Telugu. On the other hand, Arabic DLA is yet the least addressed. Regardless the fact that Arabic is spoken by more than 300 million people, and is ranked as the 5th top-spoken language worldwide[2]. Several reasons could be leading to this issue on the top of which is the absence of publicly-available annotated datasets to work on, and the complex nature of the Arabic script.

The DLA system includes several stages. Some stages are optional according to the desired output representation. These stages include: preprocessing, physical layout analysis, logical layout analysis, and output (document) representation.

The preprocessing stage details depend on the input document quality. Degraded historical documents, old newspapers could suffer low resolution, skews, tears, ink bleeds, shadowing, see-through and many defects. Therefore, most preprocessing basically includes:

1. Binarization
2. Noise detection and removal.
3. Image quality enhancement
4. Skew detection and correction.

Physical layout analysis (PLA) aims to decompose the document image into homogenous regions and identify the type of content in these regions as text, and non-text. On the other hand, logical layout analysis (LLA) defines each textual component's functional role within the document for example as being a header, footer, figure caption, text body, etc.