

**ARABIC OPTICAL CHARACTER RECOGNITION
USING ARTIFICIAL NEURAL NETWORK METHOD**

UNDERGRADUATE THESIS



ANA AINUL SYAMSI S.

1112001034

**INFORMATICS STUDY PROGRAM
FACULTY OF ENGINEERING & COMPUTER SCIENCE
UNIVERSITAS BAKRIE
JAKARTA
2016**

**ARABIC OPTICAL CHARACTER RECOGNITION
USING ARTIFICIAL NEURAL NETWORK METHOD**

UNDERGRADUATE THESIS

**Submitted as a partial fulfillment of the requirements for bachelor degree
of Computer in Informatics Study Program, Universitas Bakrie**



ANA AINUL SYAMSI S.

1112001034

**INFORMATICS STUDY PROGRAM
FACULTY OF ENGINEERING & COMPUTER SCIENCE
UNIVERSITAS BAKRIE
JAKARTA
2016**

**ARABIC OPTICAL CHARACTER RECOGNITION USING ARTIFICIAL NEURAL
NETWORK METHOD**

2 September 2016

ANTIPLAGIARISME STATEMENT PAGE

Writer, the undersigned below:

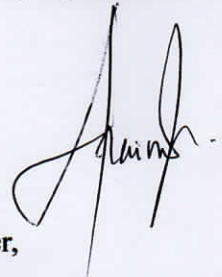
Name : Ana Ainul Syamsi Syamsuddin

NIM : 1112001034

Study Program : Informatics

Stated that this proposed research, is in my own work and prepared without using unjustified aids, just as usual in the preparation of proposal. All other written quotes and thoughts, both from published or unpublished sources (including books, journals, articles, lecturer notes, etc.), have been referenced properly according to academic standard rules. If there is mistake and error, I am ready to bear the guilt legal sanctions.

Jakarta, September 2nd 2016



Writer,

Ana Ainul Syamsi Syamsuddin

ENDORSEMENT PAGE

1. Title of research : Arabic Optical Character Recognition Using Artificial Neural Network Method
2. Field of study : Information Technology
3. Researcher : Ana Ainul Syamsi Syamsuddin
4. Student ID Number : 1112001034
5. Gender : Female
6. Address/zip code : Jl. Menteng Atas Selatan, RT.3, RW. 13, No.27, South Jakarta/ 12960
7. Phone : 085696403313
8. E-mail address : anaainul@gmail.com
9. LinkedIn : Ana Ainul Syamsi Syamsuddin
10. Proposed time : 2016

Jakarta, September 2nd 2016

Writer,

Ana Ainul Syamsi Syamsuddin

STATEMENT OF APPROVAL


This undergraduate thesis is prepared and submitted by:

Name : Ana Ainul Syamsi Syamsuddin
NIM : 1112001034
Study Program : Informatics
Faculty : Engineering and Computer Science
Title : Arabic Optical Character Recognition Using Artificial Neural
Network Method

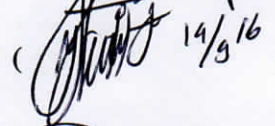
has been approved by the Board of Examiners and accepted as a partial fulfilment of the requirements to obtain Bachelor of Computer in Informatics Study Program, Faculty of Engineering and Computer Science, Universitas Bakrie.

Board of Examiners,

Thesis Supervisor : Irwan Prasetya Gunawan, Ph.D.

 14/09/16

Examiner I : Yusuf Lestanto, S.T, M.Sc.

 14/9/16

Examiner II : Berkah I. Santoso, S.T, M.T.I

 14/09/16

Authorized in : Jakarta

Date : September 2nd 2016

ACKNOWLEDGEMENT

All gratitude is in order to ALLAH SWT for his blessing so that this thesis entitled “Arabic Optical Character Recognition Using Artificial Neural Network Method” can be completed. This thesis submitted as a partial fulfilment of the requirements to obtain Bachelor of Computer in Informatics Study Program, Engineering and Computer Science Faculty, Universitas Bakrie.

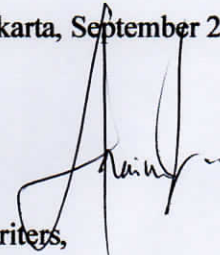
During this research there have been many people who have guided, helped, and inspired me. Therefore I also would like to express my sincere gratitude and appreciation to:

1. Irwan Prasetya Gunawan, Ph.D, as thesis supervisor who has really patiently gave the author so much precious guidance, knowledge and suggestions along the process of this research and writing report.
2. Prof. Hoga Saragih as Head of Informatics Study Program who always support and help the author during the process of this thesis.
3. Yusuf Lestanto, S.T., M.Sc. as preliminary thesis examiner who provide advice and correction for this research.
4. My greatest parents Drs. H. Syamsuddin, Dra. Hj. Nurhaeni M.Pd, and my lovely brother and sisters, Moh. Amar Ma’ruf S., Nurul Istiqamah S., Asmaul Husna S., for their unlimited love, support and never-ending-prays.
5. Special thanks to Siti Nur Fatimah, Khairunnisah, Riska Fitriawati, Muh. Rien Suryatama Idrus, Atikah Chairunnisa, Ayyu Andhysa, Mei Silviana and Rismunandar Winata for their unlimited support and motivation to finish this research as soon as possible.
6. All Maktabul Irsyad and Relawan Sahabat Tauhid members who always asked when the author finish this thesis because they always ready to attend my graduation day even though they are really busy person.
7. All Atlantis member for their unlimited support, prays, and unlimited friendship.
8. My lovely housemates-sister, Nur Fadhilah, Putri Dwi Ilhami, and Riska Amalia for their hospitality and kindness that always cheer me up through the tough time when working on this thesis.

9. All beloved friends, TIF 2011 members who always ready to accompanied and support the author working on this thesis.
10. PT. Bumi Resources Tbk, who has given me the scholarship for my study in Universitas Bakrie.
11. Contributors in Stackoverflow and Mathworks who have helped me to solve some technical problem and give comprehensive explanation in the process of coding and writing this report.

This thesis is still have a lot of imperfections. Therefore, constructive criticism and suggestion from readers are really needed for the sake of this thesis perfection. The author really hope that this research would bring benefits in academic society, OCR implementation and especially for all readers.

Jakarta, September 2nd 2016



Writers,

Ana Ainul Syamsi S.

DECLARATION OF PARTIAL COPYRIGHT LICENSE

As an academicians of Universitas Bakrie, I who assigned below:

Name : Ana Ainul Syamsi Syamsuddin
NIM : 1112001034
Study Program : Informatics
Faculty : Engineering and Computer Science

As author, whose copyright is declared on the title page of this document, for the development of science, agree and grant Universitas Bakrie a Non-Exclusive Royalty-free Right for this ungraduated thesis, entitled:

ARABIC OPTICAL CHARACTER RECOGNITION USING ARTIFICIAL NEURAL NETWORK METHOD

With the granted permission to use this material, Universitas Bakrie is allowed to keep or make digital copy, communicate, and publish this Undegraduated Thesis by providing full acknowledgement of the copyright and the source of the material.

Jakarta, September 2nd 2016

Approved by,

Ana Ainul Syamsi S.

Arabic Optical Character Recognition Using Artificial Neural Network Method

Ana Ainul Syamsi Syamsuddin¹

Abstract

Optical Character Recognition (OCR) is the process of converting scanned images of machine printed or handwritten text into a computer based format. It involves computer software that designed to translate images of text into machine printed editable text, or to translate pictures of characters into a standard encoding scheme representing them in ASCII or Unicode. This research will focus on OCR for Arabic script with all its unique characteristics. Feed Forward Neural Network Method with Back-propagation algorithm for training and testing stage were used. Using APTI data set, the research conducted with nearly 6000 images of both isolated and cursive characters. The research work in three main stages, pre-processing, feature extraction, and classification or recognition. Pre-processing stage consists of binarization, complement, normalization, and thinning. Segmentation stage also provided. Zoning, 2D DCT, and GLCM were implemented for Feature Extraction stage. Best algorithm that give best result respectively as follows: binarization, complement, segmentation, normalization, thinning, feature extraction, and classification. The proposed method yields the best accuracy rate up to 96.08% for 19 character classes experiment using Zoning method. While accuracy rate for 38 character classes experiment achieved up to 72.43% using 2D DCT method. K-fold cross validation also implemented and increased the accuracy rate for each method. So that, it proven effectively well support method for Artificial Neural Network.

Keywords : **OCR, Feed Forward Neural Network, Back-propagation algorithm**

¹ Undergraduated Student of Informatics Study Program, Bakrie University

TABLE of CONTENTS

Table of Content.....	iii
Pictures List	v
Table List	vii
Chart List.....	viii
Abreviation List.....	ix
Chapter I.....	1
Introduction	1
1.1 Background.....	1
1.2 Problem Statement	3
1.3 Purpose of Research	4
1.4 Scope of Research	4
1.5 Benefit of Research.....	4
1.6 Contribution of Research	5
Chapter II	6
Literature Review.....	6
2.1 Related Work.....	6
2.2 Off-line Optical Character Recognition	9
2.3 Characteristic of Arabic Script	11
2.4 Components of Recognition System.....	14
2.4.1 Preprocessing.....	14
2.4.2 Segmentation	24
2.4.3 Feature Extraction	26
2.4.4 Classification.....	28
2.4.5 Evaluation	33
Chapter III.....	35
Research Method.....	35
3.1 Research Framework	35
3.2 Research Tools	36
3.2.1 Dataset Access.....	36
Chapter IV	39
Research Implementation and Analysis.....	39

4.1 System Implementation	39
4.1.1 Data Preparation	41
4.1.2 Pre-processing	42
4.1.3 Feature Extraction	52
4.1.4 Classification and Recognition.....	56
4.2 Evaluation	59
4.2.1 19 Classes with diacritical Elimination.....	60
4.2.2 Classes based on APTI (with character labels).....	70
Chapter V.....	76
Conclusion and Future Work	76
5.1 Conclusion	76
5.2 Recommendation & Future Works.....	77
Bibliography	78
Appendix.....	84

PICTURES LIST

Figure 2.1	Types of character recognition technique [9].....	10
Figure 2.2	Arabic cursive script that written right to left	11
Figure 2.3	Shapes of Arabic characters based on position [17]	11
Figure 2.4	Arabic variation in size [12]	13
Figure 2.5	Secondary components of Arabic characters [33]	13
Figure 2.6	Words with different sub-words	13
Figure 2.7	Baseline, Ascender and Descender in a word	14
Figure 2.8	Character image after edge detection.....	17
Figure 2.9	Output of dilation process in image of character (original image)	19
Figure 2.10	Output of dilation from edge detection result	20
Figure 2.11	Output of erosion operation on image of character (original image)	21
Figure 2.12	Output of erosion operation from complement image result.....	21
Figure 2.13	Hole Filling Output	22
Figure 2.14	Skeletonization result for isolated character images	23
Figure 2.15	Output for thinning process on isolated character image.....	23
Figure 2.16	The differences output between Skeletonization and Thinning process..	24
Figure 2.17	Classification of Character Recognition System.....	25
Figure 2.18	Two Layer Feed Forward Network	29
Figure 2.19	Log Sigmoid Transfer Function.....	29
Figure 2.20	Confusion Matrix [28].....	33
Figure 3.1	Research Framework	37
Figure 3.2	10 Font types in APTI data set	39
Figure 3.3	Example of selected data set	39
Figure 4.1	Training Process.....	39
Figure 4.2	Testing Process.....	40
Figure 4.3	Original Image (Character Image)	43
Figure 4.4	Binary Image (Character Image).....	43
Figure 4.5	Complement output of binary image.....	44
Figure 4.6	Image Profile showed by histogram (show rising and the falling	

edge of image).....	45
Figure 4.7 Normalization result using first algorithm	46
Figure 4.8 Normalization result using second algorithm	47
Figure 4.9 Original Images (before thinning process).....	48
Figure 4.10 Thinning result	46
Figure 4.11 Histogram represent image profile	50
Figure 4.12 Segmentation Output (Isolated Character).....	51
Figure 4.13 Segmentation Output (Cursive Character)	52
Figure 4.14 Over-segmented character.....	52
Figure 4.15 DCT result	54
Figure 4.16 Two zigzag order in sorting 2D coefficients into one vector.....	55
Figure 4.17 Artificial Neural Network	57
Figure 4.18 Sample of plotperform of ANN.....	59

TABLES LIST

Table 2.1	Research Comparison	8
Table 2.2	Arabic variation shapes of characters based on position [11].....	12
Table 2.3	Vowel Sign of Arabic characters [26].....	20
Table 4.1	Data Set of the system after selection	40
Table 4.2	Comparison of Training Functions	58
Table 4.3	Arabic Character Classes after Diacritics Elimination	58
Table 4.4	Accuracy rate result using APTI 1 (with Zoning method).....	60
Table 4.5	Accuracy rate result using APTI 5 (with Zoning method).....	60
Table 4.6	Accuracy rate result using APTI 7 (with Zoning method).....	61
Table 4.7	Accuracy rate result using APTI 13 (with Zoning method).....	61
Table 4.8	Accuracy rate result using APTI 1 (with DCT method)	62
Table 4.9	Accuracy rate result using APTI 5 (with DCT method)	62
Table 4.10	Accuracy rate result using APTI 7 (with DCT method)	63
Table 4.11	Accuracy rate result using APTI 13 (with DCT method)	63
Table 4.12	Accuracy rate result using APTI 1 (with DCT method using PCA).....	63
Table 4.13	Accuracy rate result using APTI 1 (with GLCM method).....	66
Table 4.14	Accuracy rate result using APTI 5 (with GLCM method).....	66
Table 4.15	Accuracy rate result using APTI 7 (with GLCM method).....	67
Table 4.16	Accuracy rate result using APTI 13 (with GLCM method).....	67
Table 4.17	Accuracy rate using Zoning method	71
Table 4.18	Accuracy rate using 2D DCT method.....	71
Table 4.19	Accuracy rate result using APTI 13 (with GLCM method)	71
Table 4.20	Accuracy rate result using APTI 5 (with dct2 method)	72
Table 4.21	Accuracy rate result using APTI 7 (with dct transform matrix method)	73
Table 4.22	Training and test Performance using Zoning method.....	73
Table 4.23	Training and test Performance using 2D DCT method.....	74

CHART LIST

Chart 4.1 Accuracy rate based on Arabic Transparent Font	69
Chart 4.2 Accuracy rate based on 24 Font size	70

ABBREVIATIONS LIST

ANN	Artificial Neural Network
AOCR	Arabic Optical Character Recognition
APTI	Arabic Printed Text Image Database
ASCII	American Standard Code for Information Interchange
BPNN	Back-propagation Neural Network
CAPTCHA	Completely Automated Public Turing Test to Tell Computers and Humans Apart
GLCM	Gray Level Co-occurrence Matrix
MSE	Mean Squared Error
OCR	Optical Character Recognition