



**Arabic Offline Handwritten Isolated Character
Recognition System Using Bayesian Network and
Neural Network**

التَّعرف إلى الحروف العربيَّة المعزولة والمكتوبة بخط اليد باستخدام
الشبكة البيزية والشبكة العصبية

By

Ahmed Subhi Abdalkafor

(401320041)

Supervisor

Dr. Sadeq ALHamouz

Master Thesis

Submitted In Partial Fulfilment of the Requirements for the

Master Degree in Computer Science

Faculty of Information Technology

Middle East University

Amman – Jordan

January 2016

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

إِنَّا أَنْزَلْنَاهُ قُرْآنًا عَرَبِيًّا لَعَلَّكُمْ تَعْقِلُونَ

صدق الله العظيم

سورة يوسف اية (٢)

Authorization statement

I, Ahmed Subhi Abdalkafor, authorize the Middle East University
To supply a copy my Thesis to libraries, establishment or individuals.

Name: Ahmed subhi Abdalkafor

Date: 10 / 01/ 2016

Signature:

A handwritten signature in blue ink, appearing to be 'A. S. Abdalkafor', written over a horizontal line.

أقرار تفويض

انا احمد صبحي عبد الغفور أفوض جامعه الشرق الاوسط بتزويد نسخ من رسالتي

للمكتبات او المؤسسات او الهيئات او الافراد عنده طلبها.

الاسم: احمد صبحي عبد الغفور

التوقيع : 

التاريخ: 2016 /01/ 10

Middle East University
Examination Committee Decision

This is to certify that the thesis entitled "**Arabic Offline Handwritten Isolated Character Recognition System Using Bayesian Network and Neural Network**" was successfully defended and approved in 2016.

Examination Committee Members

Signature

Dr. Sadeq AlHamouz (Supervisor & Member)



Associate Professor, Department of Computer Information Systems
Middle East University (MEU)

Dr. Mudhafar Al-Jarrah (Chairman)



Associate Professor, Department of Computer Information
Systems
Middle East University (MEU)

Prof. Ahmad Sharieh (External Member)



Computer Science Department
King Abdullah II School for Information Technology
The University Of Jordan

Acknowledgements

First and before everything, I give thanks, praise to **Allah** for his mercy, and reconcile and for granting me knowledge, confidence, patience to pass this master thesis successfully.

And I give thanks for the first teacher to the nation, who saved it from the darkness of ignorance to the light of learning, illiterate Prophet **Muhammad (PBUH)**.

Also, I would like to express my thanks and gratitude to my supervisor, **Dr. Sadeq AlHamouz** for his guidance, encouragement, as well for the support on the way. I thank him again for his insightful conversations and I would like to express my thanks to **Dr. Ahmed Sahlol** for answer all my questions on this subject and **Dr. Nicola Nobile** for facilitating access the database.

Dedication

To My

Father, Mother, Wife and my family for their full support, for their great patience, for their great attention and pray for me.

I dedicate my effort

Table of Contents

Cover Page	I
الآية الكريمة	II
Authorization statement	III
اقرار تفويض	IV
Examination Committee Decision	V
Acknowledgements	VI
Dedication	VII
Table of Contents	VIII
List of Tables	XII
List of Figures	XIII
List of Abbreviations	XVII
ABSTRACT	XIX
الملخص	XX
Chapter 1: Background of the study and importance	1
1.1. Introduction	1
1.2. Problem Statement	2
1.3. Questions of the Study	3
1.4. Motivation of the Study	3
1.5. Contribution	4
1.6. Methodology of the Study	4
1.7. Limitations of the Study	5
1.8. Thesis Outlines	5
Chapter 2: Theoretical Background and Literature Review	6
2.1. Introduction	6
2.2. Arabic Language	7
2.3. Databases in Arabic optical Character Recognition	11
2.4. Optical Character Recognition	18
2.5. Bayesian Network	21
2.5.1. Introduction	21
2.5.2. Naïve Bayesian as Likelihood Estimator	23

2.5.3. Optical Character Recognition Systems based on Naïve Bayesian Network: Literature Review.....	23
2.6. Artificial Neural Networks	25
2.6.1. Introduction	25
2.6.2. Biological and Artificial neurons	26
2.6.3. Artificial Neural Network Mechanisms	27
2.6.4. Multi-Layer Perceptron.....	28
2.6.5. Learning	29
2.7. Optical Character Recognition Systems Based on ANN	30
2.7.1. Optical Character Recognition Systems Based on ANN: Literature Review	31
Chapter 3: Proposed BBAOCR System: Implementation and Methodology	39
3.1. Introduction	39
3.2. Phase (I): Dataset Collection	41
3.2.1 Character Labels Codification	42
3.3. Phase (II): Character Recognition	44
3.3.1 Stage (I): Image Pre-processing	44
3.3.1.1 Background Noise removal	45
3.3.1.2 Binarization.....	45
3.3.1.3 Skeltonoization	46
3.3.1.4 Universe of Discourse of Character Image	47
3.3.2 Stage (II): Features Extraction	48
3.3.2.1 Directional Features	48
3.3.2.1.1 Zoning	49
3.3.2.1.2 Starters, Minor Starters and Intersections	52
3.3.2.1.2.1 Starters.....	53
3.3.2.1.2.2 Intersections	55
3.3.2.1.2.3 Minor Starters	58
3.3.2.1.3 Character Skeleton Traversal	60
3.3.2.1.4 Distinguishing Individual Line Segments	60
3.3.2.1.4.1 Line Segment Information Labelling	61
3.3.2.1.5 Estimation of Feature vectors Through Zoning	64
3.3.2.2 Regional Features	67

3.3.2.2.1 Euler Number	68
3.3.2.2.2 The Eccentricity	71
3.3.2.2.3 Minimal Bounding Box.....	72
3.3.2.2.4 The Extent	73
3.3.2.3 Features Normalization	74
3.3.3 Stage (III): Features Validation Algorithm in This System	76
3.3.3.1 The Naïve Bayesian Probabilistic Model.....	77
3.3.3.2 Model Parameter Estimation	79
3.3.4 Stage (IV): Character Classification	81
3.3.4.1 Back Propagation	81
3.3.4.1.1 Transfer Function	82
3.3.4.2 Back Propagation Training Algorithm	83
3.3.4.3 Proposed BBAOCR System Training Phase.....	86
3.3.4.4 Proposed BBAOCR Testing Phase	88
Chapter 4 : Experimental Results	89
4.1 Experimental Setup	89
4.2 Training and Testing Data Sets Statistics	89
4.3 Experimental Results of Feature Validation Phase	91
4.4 Experimental Results of Back Propagation ANN based OCR system	92
4.4.1 Experimental Results of the BPANN based OCR system: Training Phase	92
4.4.2 Experimental Results of the BPANN based OCR system: Testing Phase	96
4.5 Comparison between Training and Testing Phases of our OCR System	98
4.6 Comparisons of Our Proposed OCR system to other OCR systems.....	106
4.6.1 Comparison of proposed BBAOCR System and other OCRs using CENPRIM Dataset.....	106
4.6.2 Comparison of proposed BBAOCR System and other OCRs Datasets	111

Chapter 5: Conclusions and Future work	114
5.1 Conclusions	114
5.2 Recommendations and Future Work.....	116
References	118
Appendix	130

List of Tables

Table (2.1): Different Shapes of Arabic Language Character	8
Table (2.2): Isolated Handwritten Arabic Characters Databases	12
Table (3.1): Arabic Letters: Decimal and Binary Representations	42
Table (4.1): The Parameters and Topology of Back Propagation ANN: Training Phase.....	92
Table (4.2): Achieved Recognition Rates for Arabic Characters in Training Phase	93
Table (4.3): Achieved Recognition Rates for Arabic Characters in Testing Phase	96
Table (4.4): Ten worst Characters recognized by (Shalol, et.al, 2014A)	100
Table (4.5): Ten worst Characters recognized by our proposed	101

List of Figures

Figure (2.1): Arabic OCR system testing tools	6
Figure (2.2): Character Ayn (ع) in Four Shapes.	9
Figure (2.3): Main body and Secondary Component of Character Tha (ظ)	9
Figure (2.4): Different Secondary component (dots) of Character Thaa (ث)	10
Figure (2.5): Connected Secondary Component with the main body	11
Figure (2.6): CENPARMI_2 filled form	16
Figure (2.7): CENPARMI_3 filled form	17
Figure (2.8): Logical Components of OCR System.....	20
Figure (2.9): Character Recognition Systems Classification.....	20
Figure (2.10): Analogy between biological and artificial neurons	26
Figure (2.11): (a) linearly Classes (b): Non- linearly Classes	28
Figure (3.1): BBOCR System Block Diagram	40
Figure (3.2): The general structure of CENPARMI dataset	41
Figure (3.3): The Stages of Character Recognition Phase.....	44
Figure (3.4): Character Kaaf (ك) (a) Before Filtering. (b) After Filtering	45
Figure (3.5): Character Kaaf (ك) (a) Before Binarization. (b) After Binarization	46
Figure (3.6): The Skeleton of character Waaw (و).....	47
Figure (3.7): (a) Original Ha (هـ) Character image. (b) Universe of Discourse	47
Figure (3.8): (a) Original Saad (ص) Character image. (b) Universe of Discourse	48
Figure (3.9): (a) Original Ha (هـ) Character image. (b) Skeletonized image (c) Column zone_1 (d) Column zone_2 (e) Column zone_3	50
Figure (3.10): (a) Original Ha (هـ) Character image. (b) Skeletonized image (c) row zone_1 (d) row zone_2 (e) row zone_3.....	51
Figure (3.11): (a) Skeletonized Ha(هـ) image (b) zoned Ha (9 zones)	51
Figure (3.12): Zone that contain the innermost circle of Ha(هـ) character	52
Figure (3.13): Highlighted Line Segments of a Particular Zone.....	53

Figure (3.14): Direct and Diagonal Neighborhoods of a Pixel	54
Figure (3.15): The skeleton Pixels (ones) and Starters of a Particular Zone	54
Figure (3.16): starters of Tha (ط) and Meem (م) characters	55
Figure (3.17): Intersection Points in Ta(ط) Character	55
Figure (3.18): Neighbors pixels: Diagonal and Direct in the window around pixel under consideration	57
Figure (3.19): Five Pixels- Neighborhoods of a Pixel	58
Figure (3.20): Minor Starters of a Particular Pixel	59
Figure (3.21): pixel direction	62
Figure (3.22): one zone of nine zones of character image	62
Figure (3.23): Identified Segments of the given zone	63
Figure (3.24): Pixels' directions of segment1	63
Figure (3.25): Pixels' directions of segment2	64
Figure (3.26): Feature Vector of a character image	66
Figure (3.27): Euler Number Concept	69
Figure (3.28): Number of regions and Holes in Faa (ف) Character	70
Figure (3.29): Number of regions and Holes in Saad(ص) Character	70
Figure (3.30): High and Low Eccentricity	71
Figure (3.31): Eccentricity of Alif(ا) and Hamza (ء) Characters	72
Figure (3.32): (a) Bounding Box Concept. (b) Bounding Box of Ghayn (غ) character	73
Figure (3.33): The Normalized dataset	75
Figure (3.34): Normal Distribution Function	79
Figure (3.35): Kernel Distribution function	81
Figure (3.36): Information processing by the i^{th} neuron of the j^{th} layer	82
Figure (3.37): Sigmoid function	83
Figure (3.38): The Back Propagation Training Algorithm	85
Figure (4.1): The Statistics of Training, Testing and Validation	90
Figure (4.2): Offline Handwritten Arabic Isolated Letters	90
Figure (4.3): Epochs Tuning	94

Figure (4.4): Back Propagation ANN Structure (Number of neurons /Layer) Tuning	95
Figure (4.5): Recognition Accuracy Comparison: Training and Testing Phases of our Proposed OCR system.....	97
Figure (4.6): Accuracy Comparison of our proposed OCR system to proposed by (Sahlol, 2014A)	
Where shows our fully recognized characters to that recognized by (Sahlol, 2014A)	99
Figure (4.7): Accuracy Comparison of our proposed OCR system to proposed by (Sahlol, 2014A)	
Where shows our achieved accuracy rates of characters that fully recognized by (Sahlol, 2014A).....	99
Figure (4.8): (a) Tha (ظ) Character (b) غ (Ghayn) Character	101
Figure (4.9): 3x3 zones of (a) Tha (ظ) Character (b) Ghayn(غ) Character	102
Figure (4.10): Zaay (ز) Characters written as Thaal (ذ) Characters	102
Figure (4.11): The skeleton zones of Daad (ض) and Sheen (ش) Characters where Sheen (ش) written in Al-Ruqa'a Style	103
Figure (4.12): The skeleton zones of Daad (ض) and three Sheen (ش) Characters	104
Figure (4.13) Recognition Accuracy Comparison: Our Proposed BBAOCR system to other AOCR systems	107
Figure (4.14): (a) the word Khelal (خلال) composed of two classes: non-end word Khel (خلا) and end-word Laam (ل).	
(b) The word Hunak (هناك) composed of two classes: non-end word Huna (هنا) and end-word Kaaf (ك).	
(c) The word Alyaom (اليوم) composed of two classes: non-end word Alyao (اليو) and end-word Maam (م).	
(d) The word Allthi (الذي) composed of two classes: non-end word Allth (الذ) and end-word Yaa (ي).....	109
Figure (4.15): Graphical Comparison between (Jamal, 2015) and (Sahlolo, 2014) AOCR systems and our proposed BBAOCR system	110
Figure (4.16): Graphical Comparison between other up-to-date AOCR systems and our Proposed BBAOCR	111

Figure (4.17): Graphical Comparison between other AOCR systems	112
Figure (4.18): Graphical Comparison of BBAOCR system to High-performance AOCR systems	113

LIST OF ABBREVIATIONS

Abbreviation	Meaning
ANN	Neural Network
BBAOCR	Bayesian Backpropagation Arabic Optical Character Recognition System
BF	Bayesian Filter
BP	Back Propagation
BPANN	Back Propagation Artificial Neural Network
CEMPAMI	Center for Pattern Recognition and Machine Intelligence
DCT	discrete cosine transform
ESL	End-Shape Letter
GT	Ground Truth
HCR	Handwritten Character Recognition
HMM	Hidden Markove Models
HOG	Histogram of Oriented Gradients
IDE	Integrated Development Environment
IFN/ENIT	1Institute for Communications Technology/ 2Ecole Nationale d'Ingénieur de Tunis
KDE	Kernel Density Estimation
KNN	K nearest neighbour
LR	logistic recognition
MLP	Multi-layer Perceptron
NB	Naïve Bayes
NFPD	Neighborhood Foreground Pixels Density
NFPD	Neighborhood Foreground Pixels Density
NOUN	Novel Object and Unusual Name
NOUN	Novel Object and Unusual Name
OCR	Optical Character Recognition
PCA	Principal component analysis
POW	Part Of the Word
SIFT	Scale Invariant Feature Transform
SOM	Self-Organized Map

SURF	Speeded-Up Robust Features
SUST ARG	University for Science and Technology
SVM	Support Vector Machine
UPOW	Unknown Part Of Word

Abstract

Optical Character Recognition (OCR) has been an active and dynamic research field that has a wide applicability in a variety of areas since the early days of computer science. Although this research area is considered mature, Arabic language has been one of the major languages that have little attention in this field of research by Arab researchers in particular and foreign researchers in general.

Due to the highly cursive nature of handwritten Arabic language, Arabic character recognition is considered one of the most challenging problems in contrast to working with Latin, Japanese or Chinese character recognition.

In this thesis, we proposed Arabic off-line handwritten recognition system based on novel feature extraction techniques, Bayesian network as features validation and Backpropagation Artificial neural network as classification engine, we called it **Bayesian Back propagation Optical Arabic Character Recognition (BBAOCR)**.

The presented work is implemented and tested via CENPARMI database. Competitive recognition accuracy reached up to (94.75%) has been achieved. This result motivate us and other researches in this field to employ the features extraction techniques that we have used in this research with other Arabic character shapes or integrate our proposed system into handwritten Arabic text recognition systems.

Keywords: Arabic Optical Characters Recognition, Directional Features, Regional Features, Zoning, Validation, Classification, Neural Network, Bayesian Network.

الملخص

ان التعرف الضوئي على الأحرف (OCR) تعد واحدة من مجالات البحوث الفعّالة والنشطة والتي لها استخدامات واسعة في كثير من المجالات منذ الأيام الأولى لعلوم الحاسوب. على الرغم من أن هذا الحقل يعتبر من الحقول الناضجة، إلا أن اللغة العربية تعتبر واحدة من اللغات الأساسية التي حازت على القليل من الاهتمام في هذا المجال البحثي من قبل الباحثين العرب على وجه الخصوص والباحثين الأجانب بشكل عام.

ونظرا للطبيعة المتصلة للأحرف العربية المكتوبة بخط اليد، فإن التعرف إلى الأحرف العربية يعد واحدة من المشاكل الأكثر تحدياً مقارنةً مع التعرف إلى أحرف اللغات اللاتينية واليابانية والصينية.

في هذه الأطروحة، تم اقتراح نظام للتعرف إلى الأحرف العربية المعزولة والمكتوبة بخط اليد استناداً إلى تقنيات جديدة في استخراج الميزات وتم استخدام الشبكة البيزية كأداة لقياس فاعلية هذه الميزات واستخدام الشبكة العصبية ذات الانتشار العكسي كأداة للتصنيف وقد أطلقنا على هذا النظام (BBAOCR).

تم تنفيذ واختبار هذا النظام باستخدام قاعدة البيانات (CENPARMI) حيث تم الحصول على دقة تصنيف منافسة وصلت إلى (٩٤,٧٥%) وتشكل هذه النتيجة حافزاً لنا وللأبحاث الأخرى لتوظيف هذه التقنيات المستخدمة في استخراج الميزات للتعرف على الأشكال الأخرى للحروف العربية او دمج هذا النظام المقترح كجزء في أنظمة التعرف على النصوص العربية المكتوبة بخط اليد.

الكلمات المفتاحية: التعرف الضوئي للحروف العربية، الخصائص الاتجاهية، الخصائص

الإقليمية، التقسيم، التحقق، التصنيف، الشبكة العصبية، الشبكة البيزية.

Chapter One

Background of the Study and Importance

1.1 Introduction

Optical character recognition (OCR) can be defined as the mechanism in which the images of written text, word or characters can be converted into machine editable text.

Since the early days of computer science, the optical character recognition (OCR) has been an active research field, and still one of the most challenging, and dynamic areas of research in computer science. However, the researches in this discipline have a tendency toward being mature associated with a large body of published works.

Nevertheless, Arabic character recognition has been in the last place of researcher attention, despite that the Arabic language is considered one of the major languages that spoken by more than 280 million people (Abuzaraida, Zeki & Zeki, 2013). Where the issue of Arabic character recognition is important not for native Arabic speakers but for non-Arabic countries too, where the Arabic language has been adopted for alphabet writing of the non-Semitic languages , such as Kurd, Malay, Urdu and some West African countries come as second language (Alijla & Kwaik, 2012 ; AlKhateeb, 2015).

Moreover , the researches in the field of optical Arabic character recognition have limited success due to the complex natural of Arabic language in general and Arabic scripts in particular.

The principle goal of OCR is to recognize the classes (labels) of unknown characters utilizing an already database established of character classes. Any OCR system can became in one of two types: printed and handwritten; where in the former, the characters to be identified are printed using machine, whereas the latter concern in

recognition (identification) of handwritten characters that written on paper and then scanned by machines.(Asebriy, Bencharef, Raghay,& Chihab,2014).

Handwritten recognition system, on the other hand, comes in two major types: On-line and Off-line. In case of On-line handwritten recognition system, the pressure is used on the digital display of an instrument to create a series of points that traced by pen. However, offline is built on optical character recognition system and is applied on optical scanned texts. As a thumb of rule, and as a natural result, Off-line recognition is much harder than online handwritten recognition (Rashad & Semary, 2014; Lawgali, 2015).

Our research concerns in the identification of Off-line Arabic handwritten isolated characters, where the huge variations in the writing styles from one person to another, and even one for person itself poses a strong motivation for many researches and for this research to work in the direction of enhancing the recognition accuracy and precision for this type of optical character recognition systems.

1.2 Problem Statement

Recently, the field of optical character recognition (OCR) has become a dynamic area of research, due to its broad band of applications including postal sorting, signature recognition, bank cheque processing and automatic data entry and other applications.

Robust and stable handwritten Arabic characters' recognition system present a challenging task, where a number of various techniques and algorithms have been proposed. This thesis proposes a stable and robust handwritten Arabic characters recognition system, with high recognition rate and high accuracy built using back propagation artificial neural network.

1.3 Questions of the Study

The major questions in this thesis are identified as follows:

- How to build an optical handwritten Arabic characters recognition system that have high recognition rate for each isolated Arabic character?
- Is it possible to achieve high recognition rate in the case of isolated Arabic letter with hardest letterforms?
- Is it possible to build an optical handwritten Arabic characters recognition system that can be implemented in the real time and involved in real application?

1.4 Motivation of the Study

Due to its broad applicability in a variety of disciplines including signature recognition, automatic data entry, bank cheque processing and other important applications; optical character recognition designer has been an active define field of research and development either text ,word or isolated character recognition .

A variety of Arabic handwritten isolated techniques have been proposed, developed and implemented but they do not reach to a level they are fit to achieve a high recognition accuracy. Therefore, in this thesis a robust, and reliable optical recognition system for the Off-line Arabic handwritten isolated characters, based on back propagation artificial neural network is proposed, which takes advantage of two powerful feature extraction techniques, that leads to high overall rate, and stable optical recognition system.

1.5 Contribution

Low recognition accuracy represents the main challenge that the majority of currently existing optical handwritten Arabic recognition systems face. This metric can be considered as strong indicator of falsely classification of isolated characters. This issue is due to the complicated nature and the broad variation in the Arabic character shapes.

The proposed optical handwritten Arabic character recognition system based on a combination of novel feature extraction techniques and naïve Bayesian network as a likelihood estimator and back propagation artificial neural network as core classifier achieved high overall accuracy reached up to **(94.75%)** and high recognition rates reached up to **(100%)** for many letters.

Moreover, since the recognition reliability of isolated characters and dealing efficiently with large amount of character streams are principal challenges that face real recognition system. Our proposed recognition system proved its time efficiency and its capability to be improved in real-time applications with high recognition stability.

1.6 Methodology of the Study

In this thesis, we will consider the following methodology:

1. Studying of literatures in Optical Character recognition, Bayesian Networks, Artificial neural networks, and data mining and machine learning algorithms.
2. Developing Off-line Arabic Handwritten Isolated Characters that based on a combination of Naïve Bayesian and Back Propagation Neural Network algorithms.
3. Implement the proposed BBAOCR system using Matlab 2015a as an Integrated Development Environment (IDE).
4. Experimenting the system operation (Training, Testing and Validation) using a benchmark data set of isolated Arabic Characters.

5. Evaluating the performance of the system and comparing its performance with other existing Arabic Handwritten Isolated Character recognition techniques.
6. Discuss the results, give conclusions and suggest recommendations for future work.

1.7 Limitations of the Study

The recognition accuracy of this research is highly dependable on many factors related to the used database, namely, CENPARMI in our case. These factors include the way in which the letters are written by the writers who involved in the process of creating this database such as if the letters are poorly written or drawn in unusual style or using different writing style such as AL-Rukaa style, and since CENBARMi contains most of those writing defects, we expect a percent of overall performance degradation.

1.8 Thesis Outlines

The rest of this thesis is organized as follows. In Chapter II, we summarize the significant previous systems and techniques that had been used to build isolated Arabic character recognition systems. Implementation strategies and the proposed methodology of our proposed Off-line Arabic handwritten isolated characters recognition system are explained in Chapter III. The experimental results of the proposed recognition system associated with comparisons to other existed character recognition schemes are presented in Chapter IV, the conclusions and recommendations for future work are presented in Chapter V.

Chapter Two

Theoretical Background and Literature Review

2.1 Introduction

All OCR systems are built on common test tools in particular databases validation, dictionaries vocabularies and statistics about language under consideration. Figure (2.1) shows the tools that are considered essential in the process of design and implementation of any OCR system. Moreover, it is considered the most effective way to validate the experimental results of the implemented recognition systems. (Amara, e t .al, 2005).

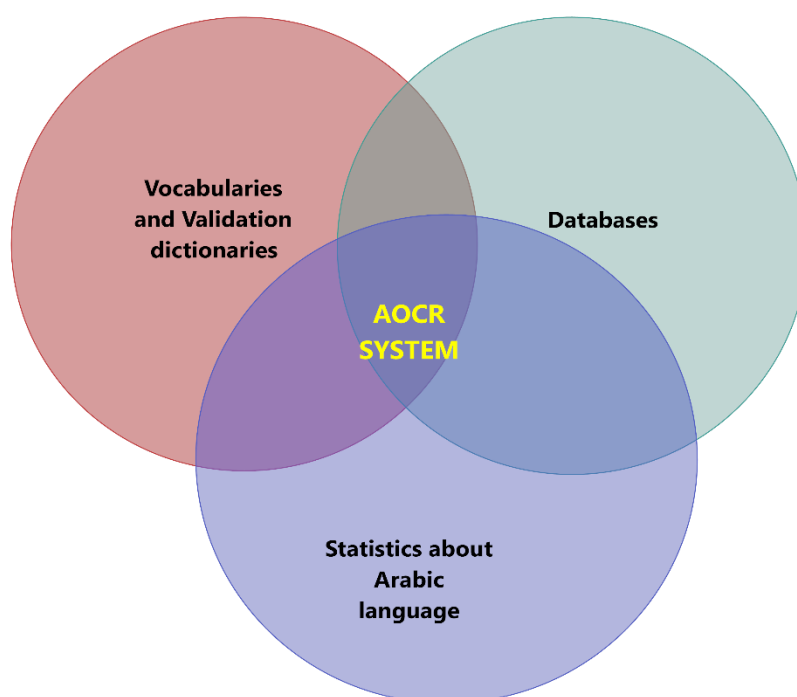


Figure (2.1): Arabic OCR system testing tools

2.2 Arabic Language

Arabic Language is considered one of most ancient languages that used by broad slice of people all over the world beside Arabic world (especially in Africa and Asia) such as Kurdish, Urdu, Malay and Farsi. More than three millions of historical manuscripts were written in Arabic language (Khorsheed, 2000; Sahlol & Suen, 2014)..

Arabic writing system is another story, Arabic language is different from Latin language such as English; Arabic is written from right to left whereas English (and other Latin language) in the reverse direction. English language can be written in non- cursive style whereas the Arabic language is always cursive (El qacimy, Hammouch, kerroum, 2015).

Basically, Arabic language (28) characters. Depending on the different shapes of characters, Arabic alphabet can be expanded to (92) according to the position of the letter ; namely if the letter comes at beginning, middle, end or in isolated form as well as according to the writing style such as Nasekh, Farsi, Roqa'a, Andulisy and others. Table (2.1) shows the different shapes of Arabic language character (Aburas& Rehiel, 2008).

Table (2.1) Different shapes of Arabic language characters

Name	Isolated	Initial	Medial	Final
Alif	أ	أ	ا	ا
Baa	ب	ب	ب	ب
Taa	ت	ت	ت	ت
Thaa	ث	ث	ث	ث
Jeem	ج	ج	ج	ج
Haa	ح	ح	ح	ح
Kha	خ	خ	خ	خ
Daal	د	د	ـ	د
Thaal	ذ	ذ	ـ	ذ
Raa	ر	ر	ـ	ر
Zaay	ز	ز	ـ	ز
Seen	س	س	س	س
Sheen	ش	ش	ش	ش
Saad	ص	ص	ص	ص
Daad	ض	ض	ض	ض
Ta	ط	ط	ط	ط
Tha	ظ	ظ	ظ	ظ
Ayn	ع	ع	ع	ع
Ghayn	غ	غ	غ	غ
Faa	ف	ف	ف	ف
Gaaf	ق	ق	ق	ق
Kaaf	ك	ك	ك	ك
Laam	ل	ل	ل	ل
Meem	م	م	م	م
Noon	ن	ن	ن	ن
Ha	ه	ه	ه	ه
Waaw	و	و	ـ	و
Yaa	ي	ي	ي	ي

Arabic character is considered *Isolated* when it is written alone. However, the Arabic character takes three other forms when it is written in connected manner to other characters in the word. To illustrate this idea, Figure (2.2) elaborates the different forms of Ayn(ع) character.

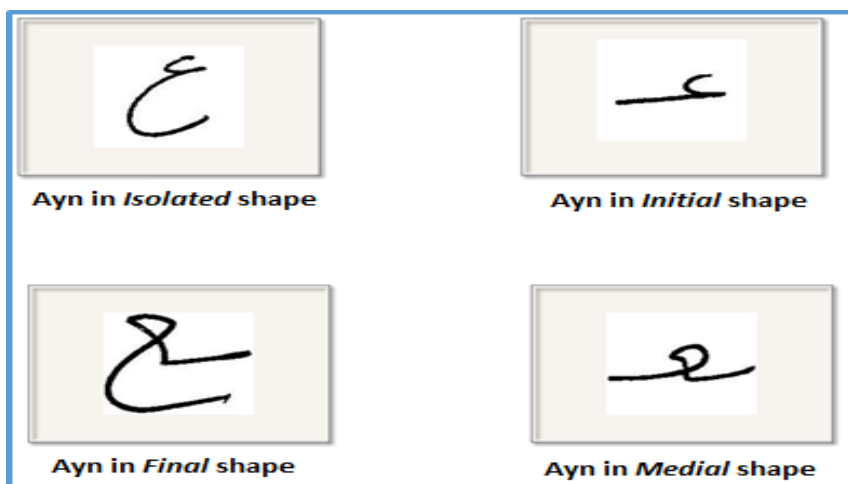


Figure (2.2): Character Ayn (ع) in Four Shapes.

Among of 28 Arabic characters, there are sixteen Arabic letters that have secondary components (dots), where secondary components are that components that disconnected from the main body of character. Figure (2.3) shows **Tha** (ظ) character and its main body and secondary component of it.

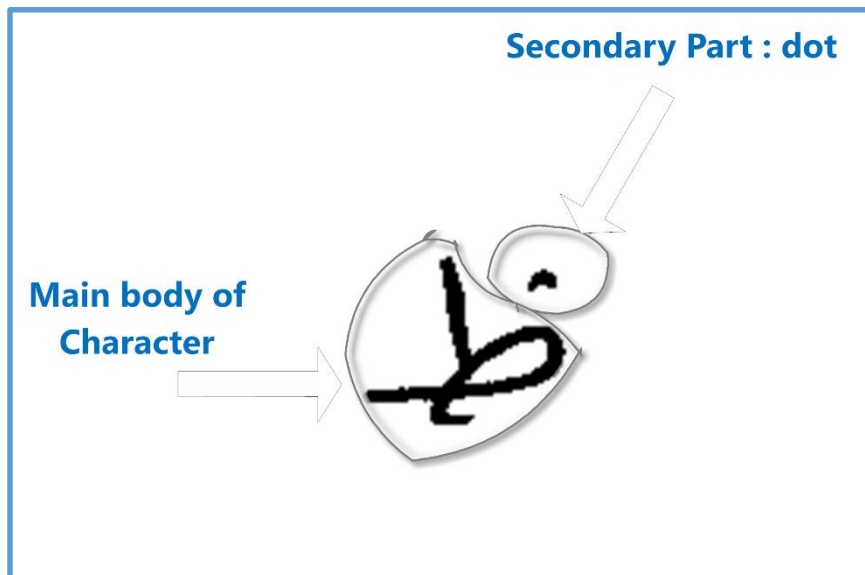


Figure (2.3): Main body and Secondary Component of Character Tha (ظ)

The position and type of secondary components have principal effect on the features (attributes) of the Arabic character. For instance, **Seen** (س) and **Sheen** (ش) differ

by secondary components (three dots). However they have the same character body, whereas **Raa** (ر) and **Zaay** (ز) by one secondary component (one dot).

Among many variations that can be made in drawing the secondary components is in drawing two or three dots. Figure (2.4) shows these variations in drawing secondary component of **Thaa** (ث) character.

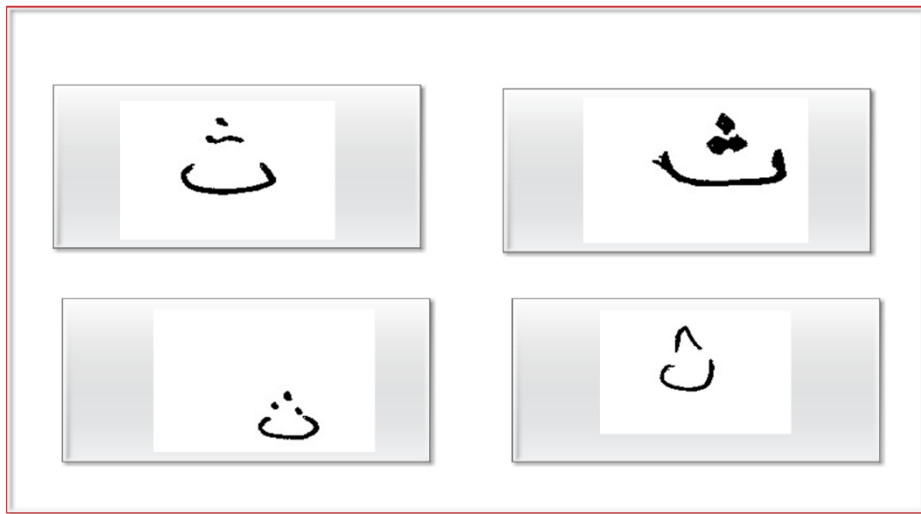


Figure (2.4): Different Secondary component (dots) of Character Thaa (ث)

One of the problems that associated the secondary components is the quick writing style of it. As Arabic writers draw them connected to the main body as shown Figure (2.5) that illustration this idea via **Noon** (ن), **Gaaf** (ق) and **Faa** (ف), characters.



Figure (2.5): Connected Secondary Component with the main body

2.3 Datasets Used in Arabic optical Character Recognition

Databases that contain a comprehensive representation of handwritten and printed Arabic characters. They are important tools variant of OCR system and it is considered as essential tool in the advanced research in the area of optical recognition.

Strictly speaking, these databases enable researchers to evaluate, upon common ground (standard reference), the performance of a variety of proposed techniques and algorithms. The process of error analysis allows researchers to move the wheel of development forward in this field of research.

In the ideal situation, these databases are issued from real documents written (filled) by real Arabic writers. They are as important as their significant size. The number of samples of written characters that used in the phase of training has a noticeable impact on the overall performance of an AOCR system.

Once the number of samples of database is limited to small set, the capacity of recognition undergoes a degradation in the recognition performance of the AOCR.

The confidentiality of handwritten Arabic database is considered one of the most important issues that arises to guarantee that samples of established database written by different writers, hence the impotence of Ground Truth (GT) that associated with the database is of high importance.

A variety of databases for different OCR system exist and implement for foreign languages such as Japanese, and Latin which contribute significantly in the course of advance and development of their OCR research such as: NIST (Geist,et .al ,1992), VNIPEN (Guyon, Haralick, Hull, & Phillips,1997), CEDAR (Kavallieratou, et .al,2001) and IAM(Marti and Bunke,2002) .

In case of Arabic language, many databases were established and many of it have recently had birth and they are offered to the use of researches either free available or commercially use. Table (2.2) lists some of databases that used in particular to build ACOR systems that designed to recognize isolated Arabic handwritten characters.

Table (2.2): Isolated Handwritten Arabic Characters Databases

Researchers	Date	Dataset Used
Abed,et .al	2015	scanned 252 image of offline isolated handwritten Arabic characters
Salouan , et .al	2015	Handwritten Arabic isolated characters
Hammad , et .al	2015	Handwritten Arabic isolated characters (SUST ARG)
Abed	2014	Total 168 characters with five shapes written by different persons
Bahashwan et. Al	2014	5,600 images and Each character has 200 images divided into four groups (beginning, medium, end and isolated) by 50 writers
Zawaideh	2012	Total 100 characters written by 10 different persons

Elglaly, et al	2011	Total 280 characters written by 5 different persons twice for each latter
Amrouch, et al	2008	Total 6188 characters written by 26 different persons 14 times for 17 characters
Aburas, et al	2007	Arabic isolated Handwritten characters collected from 48 persons

The IFN/ENIT proposed and implemented by (Pechwitz,et al, 2002) is the most broadly used database in the area of Arabic research which is provided by the Tunisian signal and processing systems laboratory in collaboration with the German Institute of telecommunication and it built on the names of Tunisian of town and villages.

(26400) names were written by (411) writers, which contains more than (210,000) characters database is available for researchers.

Most of people that contributed to this database are chosen for narrower range of the Ecole Nationale d'Ingénieurs de Tunis (ENIT), where each writer filled a form with handwritten pre-specified names of Tunisian town and villages as mentioned before.

IFN/ENIT is composed of large database of Arabic handwritten words where (Pechwitz, et.al, 2002) claim that the difficulty, time consumption and the error that associated with the process of generating ground truth for Arabic language on the level of character forced them to establish IFN/ENIT on the level of words rather than on the level of characters. Therefore, for this reason we do not use this database as training and testing database.

Another database that composed of word, signature, digits and sentences has been developed by computer science and electrical engineering department in the British Columbia University, Canada (Kharma , Ahmed & Ward,1999) where 500 students have

participated in the process of data collection that took place in Al-Isra University in Amman, Jordan.

AL- Isra database contains (73,000) Arabic words, (10,000) Arabic and Indian digits and more than 2,500 signatures. Al-Isra database does not contain Arabic characters in its isolated forms; therefore, it is not suitable as training and testing database for our research.

A rich database contains isolated handwritten Arabic characters have been established by more than 400 witters, most of them are Tunisian and it is called ARABASE (Amara, et .al, 2005).it contains a broad data issued from machine handwritten and printed documents.

Although this database seems applicable to our research goals, it have many drawbacks that might affect the overall performance of our proposed AOCR system .This is due to approach that used in the collection phase of this database, where the process of character writing is performed in distance academic environment and only by Tunisian people. Therefore, ACOR system will bias toward this band and might show degradation in performance if it were used in real-time application.

(AlKhateeb, 2015) construct a database that is eligible for Arabic handwritten recognition researches. AlKhateeb-database is composed of (28,000) digital images of the Arabic alphabets that were written by (100) native Arabic writers and it is freely available.

AlKhateeb database is freely available and it is suitable for our research since it contains isolated Arabic characters. However, the number of available samples for each character is small in comparison with CENPARMI_2, which will limit the capacity of our proposed AOCR.

In 2013, (Lawgali,Angelova and Bouridane) have created a database for Arabic character recognition purposes that cover all shapes of Arabic characters including overlapping ones. The characters of this database were written by (50) witters and contains (6,600) shapes. Ignoring the fact that the number of samples that available in this database is very limited, it contains isolated forms of Arabic characters. However, the isolated forms of Arabic characters are written without dots, which limit this database for a limited group of Arabic characters.

The Centre for Pattern Recognition and Machine Intelligence (CENPARMI) has developed three Arabic databases: CENPARMI_1 (Al-Ohali, Cheriet & Suen, 2003), CENPARMI_2 (Alamri,Sadri, Suen & Nobile, 2008), and CENPARMI_3 (Jamal,2015). All of these databases are available for public but in commercial use.

CENPARMI_1 developed in collaboration with Al Rajhi Banking and Investment Corporation where about (7,000) real word grey-level cheque {'.TIFF'} images collected and divided into numbers, Indian digit and sub-word. This version of CENPARMI does not contain Arabic character in its isolated shape.

CENPARMI_2 created by (328) writers from two countries: Montreal Canada and Saudi Arabia where they randomly selected from various ages, genders, education levels and backgrounds and nationalities.

The participants were asked to fill a form contains all classes of Arabic language: word, isolated letters, and Numerical, String Dataset and special symbols. The participates were asked to fill a form shown in Figure (2.6).

CENPARMI_2 contains three databases in one database; therefore, these databases have been separated into three different series: series-1, series-2 and series-3.

ARA0142

Arabic Handwritten Collection Form
Concordia University (Montreal, Canada)
Email: hs_alamri@concordia.ca

Address:
4420 de Maisonneuve St. - 8712-8975
Montreal QC H3J 1M8, Canada
<http://www.concordia.ca>

الاسم: _____
العنوان: _____
الهاتف: _____

الرقم الوطني: _____
الرقم القومي: _____
الرقم المدني: _____
الرقم العسكري: _____
الرقم المدني: _____

الطول: _____
الوزن: _____
الحجم: _____

الطول: متر
الوزن: كيلوجرام
الحجم: ليتر

Figure (2.6): CENPARMI_2 filled form (Alamri, et al., 2008)

Where series-1 composed of the character samples of the first (100) witters. Series-2 contains the character samples of the last 228 witters. The combination of series-1 and series-2 compose series-3. Hence, we have different series, then we can conduct a variety of experiments. Totally, (656) pages were filled and then processed.

Due to the highly heterogeneous properties of CENPARMI_2 that owned by different writers who constructed it and due to relatively large size of available character samples that associated with a ground truth, all of these reasons force as to adopt this database in the phases of training and testing.

(Jamal, 2015) created CENPARMI_3 database, which is called CENPARMI ESL database that composed of (63), words and (10) digits written by (650) participants from: Maka and Jeddah cities in Saudi Arabia. Figure (2.7) illustrate the form that used in the phase of dataset collection.



Figure (2.7): CENPARMI_3 filled form (Jamal, 2015)

2.4 Optical Character Recognition

The subject of optical character recognition (OCR) has been an area of research and development. Since it belongs to the field of pattern recognition, it has three major steps: observation, pattern segmental and pattern classification.

Optical character recognition can be defined as the process by which a large amount of documents either handwritten or printed alphabets are transformed into machine encoded text without any noise, resolution differences or other effects (Bhatia, 2014).

Our research handling handwritten recognition, which in turn can be mainly classified into two types: Off-line and On-line character recognition. The automatic process of text conversion into an image then into letter codes, which are usable within computer and text-processing applications, is called the Off-line character recognition.

Whereas, in case of On-line character recognition, the recognition system deals with a data stream comes out of a transducer while the user is writing. Smart phones, tablets are examples of typical hardware that used to collect data in case of On-line character recognition.

These types of hardware are electromagnetic or pressure sensitive in its nature, so, when the user write on it the successive pen movement are transformed into a series of electronic signal which then stored in the memory and then undergoes an analysis process by computer (Bansal, Garg, & Kumar,2014).

In contrary to On-line handwritten recognition, the Off-line recognition is considered more difficult since it depends on the user itself, which is a generator of different handwriting styles.

Pattern recognition, machine vision, artificial intelligence and signal processing are typical examples of fields of researches that involve optical character recognition.

Broadly speaking, optical character recognition can be used for a vast range of applications that includes anything that based on the idea of transformation of anything humanly readable into machine manipulatable representation (Shah, et al, 2009).

In summary optical character recognition can be defined as the process that translate any optical scanned Bitmaps of printed or written text characters into character codes such as ASCII (Alginahi, 2013).

In general, the main goal of OCR system is to enable computer to recognize optical system symbols written human mediation (Wu, Manmatha& Riseman, 1999; Li, Doermann & Kia, 2000).This process is accomplished by looking for a match between the extracted feature of symbol image under consideration and a library image models.

Hardware and computer systems that are provided with OCR systems shows a notable improvement in the speed of input operation, fast retrieval, compact storage capability, reduction in possibility of human error , and file manipulation enhancement; all of these benefits save both time and money (Mani & Srinivasan,1997).

Figure (2.8) shows the logical components of any OCR system: camera, software that implement OCR system and an output interface between machine language and user. The interface is the gate by which the final output represented friendly to the user in form of a series of characters.

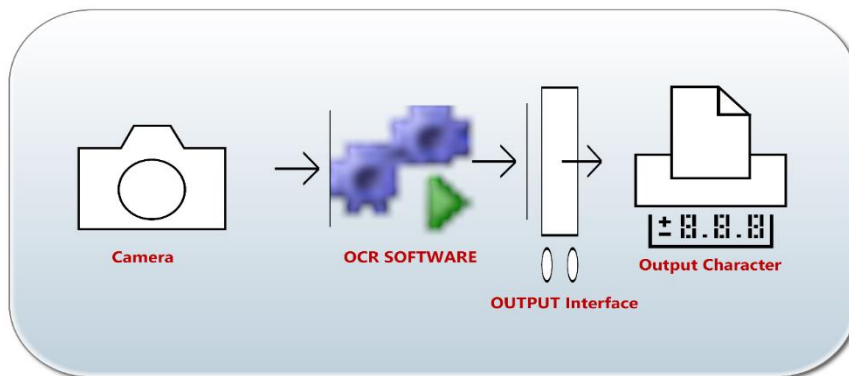


Figure (2.8): Logical Components of OCR System

As shown in Figure (2.8) camera will capture the text images optically then these text images undergo a processing with OCR software algorithm, which involves the following phases: Filtration, Segmentation of characters, Thinning and Character recognition using learning systems or artificial intelligence.

The output interface is an interface hardware the user and the OCR software, which is in charge of communicating OCR system to the outside real world (Shah, et al, 2009).

Generally, Character recognition systems can be classified in to different categories as illustrated in Figure (2.9) (Lawgali, 2015).

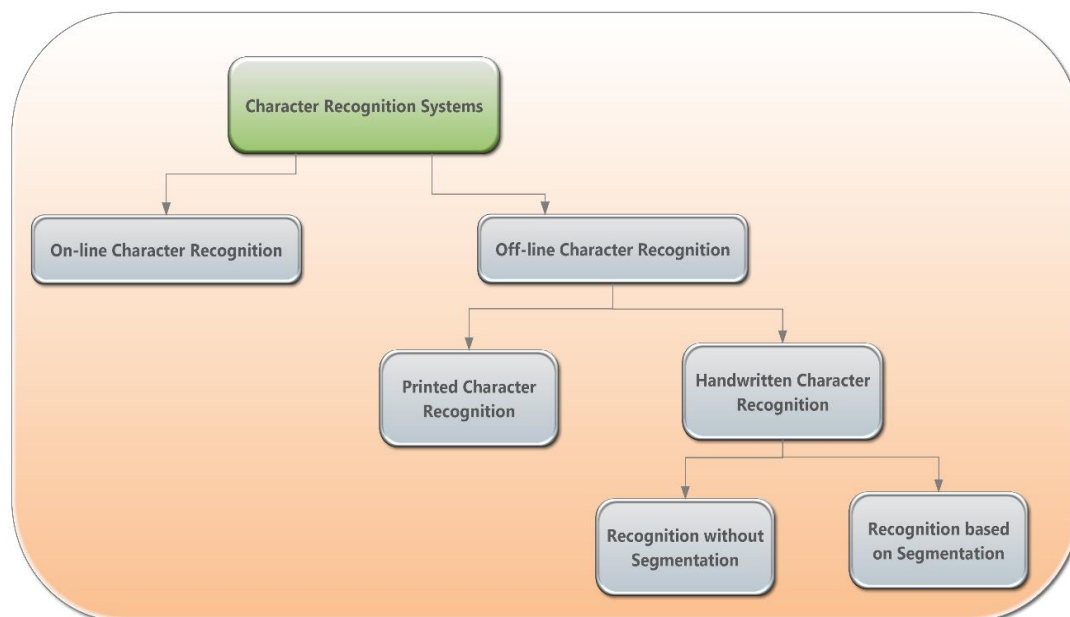


Figure (2.9): Character Recognition Systems Classification (Lawgali, 2015)

As illustrated in Figure (2.9), the character recognition system can mainly lie in one of two major categories: On-line or Off-line character recognition

Off-line recognition system of document can be further splitted into two major categories: handwritten text or printed one. Printed characters have one size and one style for any given font. On the other hand, the handwritten characters have different sizes and styles with different writers or with the writer himself.

Now, each handwritten word can be recognize as (without segmentation) or recognized based on segmentation. (Lawgali, Angelova & Bouridane, 2014).

2.5 Bayesian Network

2.5.1 Introduction

Bayesian network is considered one of the well-known supervised and statistical learning method that used for classification purposes (Hemalatha, et al., 2002) Bayesian classifier used when the model under consideration is probabilistic in nature. Hence, the Bayesian can capture the uncertainty a about the model and process it in a systematic way that determines the probabilities of the model outcomes. Therefore, Bayesian classifier is very suitable to solve predictive problems.

This classification technique built on the Bayes theorem that proposed by Thomas Bayes (1702-1761). This theorem used to calculate probabilities a particular problem hypothesis explicitly and at the same time it does not affect by the noise that associated with input data. Bayesian classifier has high capability to minimize the misclassification probability. (Vijaykumar, Vikramkumar & Trilochan, 2014).

The core idea of Bayesian classifier can be simplified as follows: Given a set of different objects (the character images in our case), and let us assume that each of these

objects belongs to known class (character labels) and each object (character image) has a known vector of variables (features in our case).

Now, our goal is to establish a rule that will allow us to assign the class (the character label) to a future object (future character or the Character under testing) provided that we have only the vectors of variables (features) that describe the future objects.

Naïve Bayes Technique is considered very important due to its simplicity mathematically and when it has been implemented. Therefore, it has many titles that refer to this property such as idiot's Bayes, independence Bayes and simple Bayes (Wu, et al, 2008).

This simplicity contribute to enhanced ease in the mathematical construction phase, namely, it does not need complex iterations to estimate the model parameters. This can make a conclusion, that naïve Bayesian could be readily utilize to huge amount of data.

Another important property is the ease of interpretation, where even unskilled users in classifier techniques and algorithms can understand the way that followed by Bayesian in the classification process. (Wu, et al, 2008).

On the other hand, although Bayesian is very simple in comparable to other classification techniques, it often does (classify) surprisingly well. Generally speaking, Naïve Bayesian can do quite well in some applications. However, it is not the best possible classifier in any specific application. (Song, et al, 2014).

2.5.2 Naïve Bayesian as Likelihood Estimator

In real-world applications, features in a variety of learning tasks are correlated to each other. Therefore, Naïve Bayesian conditional independence assumption may impair its classification performance according to the likelihood percentage between objects that undergo the classification (Wu, et al., 2015)

In our research, we exploit this property not only for classification purposes but also in order to estimate the correlation between feature vectors of characters.

If the correlation is high, then our feature extraction techniques must be modified or changed correspondingly, otherwise, the proposed feature extraction techniques are acceptable and can be used to distinguish characters via the Back-propagation ANN with expected high performance depending on the percentage of correlation that yield by Naïve Bayesian classifier.

2.5.3 Optical Character Recognition Systems based on Naïve Bayesian Network: Literature Review

This section presents and discusses past works that used the Bayesian network classifier in the processing phase of Arabic and non-Arabic optical character recognition systems.

(Bansal, Garg, & Kumar, 2014) presented an off-line handwritten Gurmukhi character recognition using three algorithms: SVM, MLP, and Naïve Bayes in separated manner.

The proposed recognition technique proposed a novel feature extraction-technique called Neighbourhood Foreground Pixels Density (NFPD); the feature dimensionality has been reduced using Principal component analysis (PCA).

The proposed technique achieved high recognition accuracy up to (91.95%) using Support Vector Machine (SVM) with radial basis function as kernel utilizing ten-fold cross validation test method and (77.6%) in case of Naïve Bayesian classification.

(Albashiti & Tamimi, 2012) proposed a lexicon Based offline handwritten recognition that was built using the Naïve Bayesian classifier. The suggested AOCR system has five major phases: Pre-processing, feature extraction, classification, probability computation and word recognition.

After the sample word image is pre-processed, the features were extracted out of each Part Of the Word (POW), then the Naïve Bayesian network classifier is used to assign the class of each Unknown Part Of Word (UPOW) depending on the Gaussian distribution. Once the probability of the combination between POWs and lexicon was calculated, the lexicon is used to validate the given combination namely, the word.

The proposed technique can guess the word and give the user the most similar words to the word under testing provide the data from lexicon is given and the success of this technique is tested and reported where the success rate reached up to 92.87 % for (26) recognize words.

An Arabic on-line characters recognition system based on dynamic Bayesian network was proposed by (Tlemasni & Benyettou, 2012). The utilization of dynamic Bayesian networks enabled an effective treatment of the isolated set of Arabic characters by keeping their special properties and the writing order for each character. The proposed technique uses the Novel Object and Unusual Name (NOUN), which is a database that established to initiate the research and development in the field of on-line Arabic recognition. NOUN composed of (2800) isolated Arabic characters.

The proposed technique was implemented on all Arabic letters and it achieved an overall recognition rate of (66.78%). Although we have used Naïve Bayesian network in save feature likelihood estimator, we have achieved higher classification (character recognition) rate reached up to **(70%)** using kernel distribution.

(Araki, Okuzaki, Konishi, & Ishigaki, 2008) proposed a novel handwritten Japanese character recognition techniques for two types of Japanese characters build up on statistical algorithms that in turn based on Bayes theorem.

The proposed technique was called Bayesian Filter (BF). Although the proposed technique was relatively simple, it archived a recognition rate above (90%) with only a few learning data.

2.6 Artificial Neural Networks

2.6.1 Introduction

Artificial Neural Network (ANN) can be defined as computational modelling tools that inherently proposed to model the complex (highly non-linear) real-world problems (Rouhani & Ravasan, 2013).

The structure of ANN composed of densely interconnected adaptive simple processing unit called artificial neurons. These Artificial neurons have the capability to perform enormous parallel computational data processes in parallel way (Hecht-Nielsen, 1990; Schalkoff, 1997; Basheer & Hajmeer, 2000).

This property of ANNs are attractive for the fields that need high non-linearity processing, robustness, and high parallelism and in applications that handle information of fuzzy and imprecise nature.

ANNs can be considered a robust abstraction of biological counterparts. However, the core idea of ANN is not to replicate the duties (or working algorithm) of biological neurons, but to make use of well-known neurons functionality in sake of solving complex problems. Therefore, the major objective of ANN is to develop mathematical algorithms that make ANN capable to learn by emulating the algorithm that used by brain to process information or knowledge acquisition.

2.6.2 Biological and Artificial neurons

A rough analogy between artificial neuron and biological neuron is that the connection between the nodes in artificial neuron represent the axons and dendrites in the biological one whereas the weights upon each connection represent the synapses in biological neurons (Jain, sperauct al., 1996).

The activity in the soma of biological neurons can be approximated as threshold in the artificial neurons (Jain, sperauct al., 1996).

Figure (2.10) elaborates the Analogy between biological neuron and artificial one.

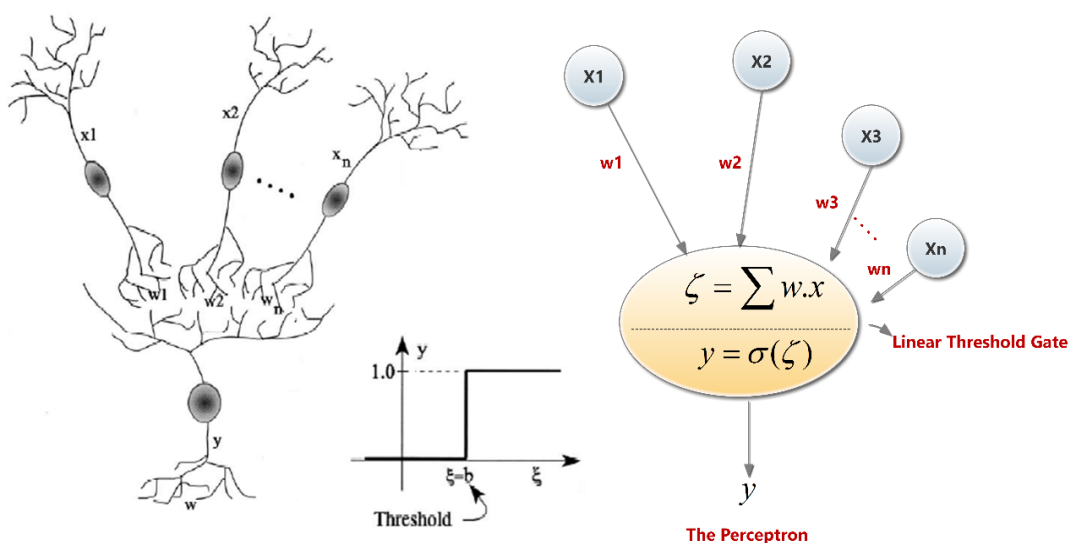


Figure (2.10): Analogy between biological and artificial neurons (Basheer & Hajmeer, 2000)

As shown in Figure (2.10), (n) biological neurons associated with different signals of intensity (\mathbf{x}) and synaptic strength (\mathbf{w}), that fed into a neuron with a threshold (b) and the analogy equivalent artificial one.

Both of neural networks, artificial and biological one learn by adjusting the weights magnitudes (synapses strengths in biological systems) in incremented way (Zupan and Gasteiger, 1993).

2.6.3 Artificial Neural Network Mechanisms

The mechanics of single artificial neuron was introduced by Rosenblatt in 1958, where he proposed the concept of "*perceptron*" to solve many problems arose in the field of character recognition (Hecht-Nielsen, 1990; Basheer& Hajmeer, 2000).

The principal findings that extracted out of the biological neuron operation enabled early researchers such as (McCulloch and Pitts, 1943) to model the operation of simple artificial neurons.

As Figure (2.10) shows, the artificial processing neuron receives the inputs as stimuli from the environment. Then, the stimuli are collected in particular manner to form the resultant (net) input (ξ) that will pass a linear threshold gate (the transfer function here is linear; However it can be a non-linear function).

Then the output of threshold gate (y) is forwarded to neighbor neuron or the neighbor environment. The neuron is activated only and only if (ξ) exceeded the neuron threshold.

The net (resultant) input is evaluated as the dot product of the input signals vector (\mathbf{x}) that impinging on the neuron and their associated weights (\mathbf{w}). (Basheer& Hajmeer, 2000).

2.6.4 Multi-Layer Perceptron

Two-object classification problems can be categorized into two major categories: linearly and non-linearly separable.

In the former, a linear hyper-plane can separate one class of object on one side of the plane and the other class of object on the other side of plane. Whereas, in case of non-linear problems one cannot establish such discriminator line due to the semi-fuzzy distribution of different classes of object. Figure (2.11) elaborate this idea.

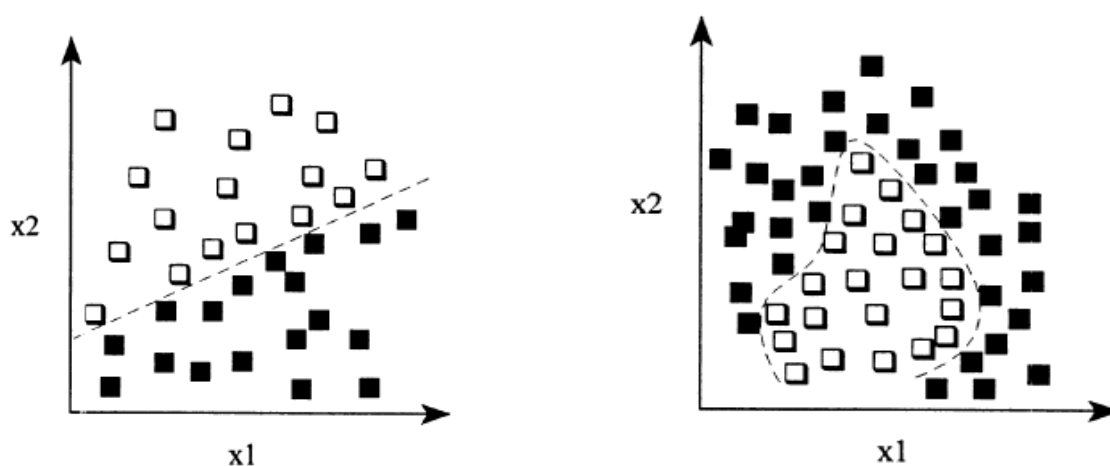


Figure (2.11): (a) linearly Classes (b): Non-linearly Classes (Basheer & Hajmeer, 2000)

In order to cope up with the nonlinearity nature of non-linear separable problems, additional layers of neurons were added between the input layer that contains the input nodes and the output neuron which compose what so called multilayer perceptron (MLP) architecture (Hecht-Nielsen, 1990; Basheer & Hajmeer, 2000).

These layers are called the hidden layers since it do not interact with the external environment, or it is isolated from the external environment through the input and output layers. The hidden layers extension will extend the capabilities of perceptron in dealing with non-linear classification problems.

2.6.5 Learning

The capability of learning is a distinguishing feature appertain to intelligent systems either biological systems or otherwise. In the artificial systems, the learning process can be viewed as the process of updating the internal representation of the system as a result of responding to external stimuli which enable such systems to perform a particular task.

This process of updating includes network architecture modification which involves weights adjustment, creating or pruning some connection links, and /or changing the activation rules of the individual nodes (Schalkoff, 1997; Basheer & Hajmeer, 2000).

The learning process of artificial neural networks is implemented through iterative manner, where the training samples are presented and passed through it in order to modify the weights exactly in the same manner that we follow when we learn from our experience.

Upon different learning rules, different ANNs can be established and defined. One of the most widely used ANN is the Back propagation ANN that is considered the workhorse of ANNs (Rumelhart et al., 1986).

This type of ANN composed of at least three layers: input, one hidden and output layer and using the supervised learning can learned through error back propagation process where the error that result from the difference between the input and output layers

is back propagated fed from output layer to first hidden layer associated with weights modification.

In our research, we have used this powerful type of ANN in the process of character recognition and it will be explained in more details in next chapter.

2.7 Optical Character Recognition Systems Based on ANN

The utilization of Artificial Neural Network (ANN) in the processes of optical character recognition system applications can considerably improve the performance quality of the recognition with an extremely simplification in the code. In addition, the utilization of artificial neural network can increase the extensibility of the OCR system, namely, it extends the capabilities of OCR system to recognize more characters sets that initially defined to work with it.

For example, in our thesis, the proposed OCR system can be extended to recognize numerical Arabic and Indian digits, or handwritten Arabic characters in beginning, medial or final shapes, which is can be considered promising in terms of building an integrated OCR system that have the ability to recognize handwritten Arabic characters in all shapes. This will be considered as a future work for us in final chapter of this thesis. The following give a literature review of OCR systems that built using artificial neural networks.

2.7.1 Optical Character Recognition Systems Based on ANN:

Literature Review

An optical character recognition system that build based on a new divide and conquer algorithm was proposed by (Hammad & Elhafiz, 2015).The proposed system utilized the curve tracing and number of dots of each character as discriminating pieces of information.

A robust connected components labelling algorithm was used in order to divide the characters into four major sets. Then, a neural network was used to classify the characters in each set. The proposed system was realized based on an off-line isolated character database that was collected and processed by the recognition group of Sudan University for Science and Technology (SUST ARG).

Despite of the complexity of the recognition system, the elementary experimental result of this system seem promising and the performance of four ANNs outperform the performance of one neural network where it archived (88.11%).

(Abed & Abed, 2015) proposed a recognition system that utilized the powerful properties of the back propagation artificial neural network and features extraction techniques. The technique that extracted the features perform when each binary image of each character was fixed, and stored in matrix form of size (18x18). Then, this matrix is divided into nine square sub –matrixes of size 6x6 where each sub-matrix represent a separate area and its elements are either one or zero depending on the existence of black pixels(that represent character or not). Although the proposed system was tested using just (12) off-line isolated Arabic handwritten characters (**Alif**(ا),**Taa** (ت),**Geem** (ج),**Thaal**(ذ), **Sheen**, (ش),**Ta**(ط), **Gyn**(غ), **Kaaf**(ك), **Meem**(م), **Waaw**(و), **Yaa**(ي)), it archived high recognition accuracy reaches up to (93.6%).

(Alkhateeb, 2015A) suggested an off-line Arabic isolated character recognition system that employed hybrid types of features extraction techniques and the back-propagation ANN a classification technique.

The proposed system was built in four major phases, First, character images acquisition, and then the characters are extracted, which form training and testing dataset of this system. Then in the second stage, a pre-processing operations are performed on character images to be ready for feature extraction phase,

In this phase, a hybrid features that composed of structural and statistical features combination are extracted where a Discrete Cosine Transform (DCT) coefficients of it are extracted then the structure features such as: position of dots, number dots, style of dots and number of holes are extracted and combine with the DCT coefficients to form the hybrid features vectors database.

The final phase represents the classification, where the Back-propagation ANN was used to classify the testing features vectors dataset and a very good accuracy of (87.75%) was achieved.

An off-line Arabic handwritten character recognition system that built utilizing HMM had been proposed by (AlKhateeb, 2015 B). The first phase of the proposed system is to detect all the variations in the character images and remove it. A sliding window technique with HMM is used to extract features. Then, the HMM is used in the process of classification and recognition.

The proposed system attained a recognition rate of (88.25%) which proved the high efficiency of HMM in both features extraction and classification phases.

A new segmentation methodology of an Arabic handwritten text line into words was proposed by (Jamal, Nobile, and Suen, 2015). This new technique utilized the Arabic writing characteristics in the process of Arabic text segmentation.

This shape analysis -based technique has two main stages: (1) metric- based segmentation and (2) recognition- based segmentation or End-Shape Letter (ESL) -based segmentation. A binarized text line will be the input for this system, where first the connected components of the text line were extracted, and secondary components of words (as dots) were removed in order to prepare the pre-processed text to be an input to the main phases that composed proposed system.

In the metric- based segmentation phase, the distance between adjacent components was evaluated utilizing a gap metric that will calculate the mean gap between text words based on estimated threshold. This stage segmented the text line into its words.

The ESL –based segmentation will specify the word segment. ESL can come into two forms: Isolated or Part of Word (PAW), this stage begins by end-shape recognition where the isolated letter or the last letter of a PAW will be identified and then the last part of the word will be extracted based on width, highs and position of baseline. The system was trained and tested using IFN/EINT database and achieved high performance reached up to (93.88%) recognition accuracy and ESL (Isolated) reached up to (90.88%) using CENPARMI database.

(Salouan, Safi& Bouikhalene,2015) suggested an isolated handwritten Arabic character recognition system that utilized novel six hybrid feature extraction techniques that built based on the idea of combination between profile technique and particular effective moments.

The feature techniques are profiled separately with Gegenbauer moment, Racah, moment, Tchybechev moment, orthogonal Fourier-Mellin moment, Hahn moment or profiled with a combination of these moments simultaneously.

Median filter, Thresholding, Centring, edge detection were used as characters images pre-processing techniques whereas the recognition process was done employing Support Vectors Machine (SVM). The simulation results of the proposed system demonstrated the effectiveness of the profile that based on a combination of moments where it had achieved a high performance reached up to (89.01%), however, it need much time more than it needed if it is profiled in separate manner.

An efficient technique for Arabic off-line handwritten characters recognition was proposed by (Sahlol , el.al,2014A) .This novel approach was built based on two principle building blocks : Novel prepressing operations which include a variety of noise removal and dilation techniques, and structural, statistical and topological features that extracted out of the main and secondary components of each Arabic character .

Structural features include the upper and lower profiles that used to capture the outlining shape of the connected portion of the character, which is composed of the horizontal and vertical projection profiles too. Where the vertical profile represents the sum of white pixels that come in normal position to the Y-axis whereas the horizontal projection profile represented by the sum of black that come in normal to the X-axis.

The statistical features are those features that resulted from discrimination between the individual foreground pixels and the set of all foreground pixels, which yields the connected components of the character. Finally, the topological features that composed of the end- points, pixel ratio and height to width ratio features of the character.

The classification (Recognition) phase is performed via back propagation ANN utilizing log sigmoid function. The proposed recognition system was trained and tested with CENPARMI -database.

The experimental results of the proposed system outperform the existing off-line handwritten character recognition schemes where it achieved an overall average recognition accuracy reached up to 88% and 100% for some letters.

A new approach for Arabic character recognition was described by(Sahlol , et.al ,2014B) this new approach is built based on new pre-processing methodology ,structural, statistical and topological features that extracted from both (1) character body (2) secondary components .

The pre-processing phase includes binarization, noise removal, median filtering and dilation. The proposed system used a new morphological noise removal approach consists of four stages: firstly for the binary from of character image the isolated interior pixels were filled if the number of non-zero neighbours are seven pixels. Then, if all neighbours are zeros, the pixel under consideration will be cleaned out. Finally, the adjacent and diagonal neighbours were specified and if three of them are zeros then it set to zero.

Feature extraction techniques include vertical and horizontal projections, where the vertical profile of a character image represent the sum of white pixels perpendicular on the y-axis.

Whereas, the horizontal projection represented by the sum of black pixels that are perpendicular to the x-axis, then each character is divided into four triangular parts and then each part is cropped by specifying the boundaries of the last non-zero pixel.

Number of secondaries, number of holes, and secondaries components positions features are extracted too.

Once the features vectors dataset is created, it was normalized and used as inputs for SVM classifier. The experimental results of the proposed recognition system shown a high achieved performance of the proposed feature extraction techniques where it attain (89.2%) recognition accuracy which is considered high performance with respect to existing AOCR systems.

In (Zawaideh, 2012), an Arabic handwritten character recognition system that based on using four cascaded networks was proposed. In this system, the features data of the unknown character are inserted to the first neural network, while the resultant output is then divided into three types that are then inserted to the three other neural networks. In the meeting of neural network, Artificial Neural Network (ANN) is considered as an information processing system. It consists of very simple and particularly interconnected cells which are linked with each other by weighted connections. Thus, the ANN input is distributed all through the network. The proposed system achieved (67%) recognition accuracy.

(Elglaly, and Quek, 2011) presented a character recognition system that built separately using two powerful techniques: K nearest neighbour (KNN) and Back propagation ANN. In the first stage of this system, the input character images are pre-processed and the main features of images that include: (1) High and width of image. (2) The number black pixels and the number of white pixels in the image. (3)The number of horizontal transitions and the number of vertical transitions are extracted and normalized in vectors to be classified in classification stage.

The classification results of this system belong to KNN and BBANN where BBANN achieved (60%) recognition accuracy whereas KNN achieved a high recognition accuracy reached up to (90%).

(Amroch, Elyassa, Rachidi,& Mammass,2008) proposed An off-line Arabic handwritten characters recognition system built based on a Hidden Markov Models. The Hough accumulator of the character image was divided into equal horizontal bands that in turn will be utilized in the process of directional information extraction. This information was translated into sequences of observations that are employed to train the model for each character during the learning step.

The Hidden Markov Models (HMM) of various Arabic characters are trained by the conventional procedure that proposed by Baum-Welch in order to adjust their parameters. Encouraging results with high detection rate of (85.71%) achieved by this system, where they proposed using a hybrid technique that combining HMM and ANN to achieve more recognition accuracy.

The use of the wavelet decomposition method (Aburas & Rehiel, 2007) in recognizing the isolated handwritten Arabic characters was proposed. This method facilitates the analysis of nonlinear signals which have many features. It is considered as the projection of the required signal on several wavelet vectors that include the most useful features. This method represents a signal with enhanced resolution in both frequency and time based on using wavelets. The wavelet transforms are divided into two types; continuous and discrete wavelet transforms. The wavelet decomposition based recognition system utilizes a discrete wavelet transform because the image characters are discrete. By using this method, the results achieved 80% recognition accuracy

A handwritten and typed Arabic character recognition system was proposed by (Al-Shridah & Sharieh, 2000). It is built based on multi-layer feed forward back propagation ANN as classification engine. The proposed system presented enhanced and modified algorithms in skeleton representation, segmentation and classification phases in the field of Arabic letters recognition within a text.

After the skeleton of each character is determined, a special feature for every two connected vertices of the skeleton were extracted in association with other generic features of the skeleton vertex, where the vertices of skeleton can be classified into four types depending on the number of vertices that are connected to it.

The classification phase of the proposed system consists of eight ANNs where a special NN is built for every pair of skeleton vertices. Although the high complexity of the proposed OCR system, it achieved high accuracy rate reached up to (97%) for (28) isolated characters in case of typed letters and (90%) in case of handwritten letters, then take the percentages of isolated handwritten letters and compare it with our proposed in experimental result.

Chapter Three

Proposed BBAOCR System: Implementation and Methodology

3.1 Introduction

The principal goal of optical character recognition system is to recognize the classes of unknown handwritten characters using a previously stored dataset of characters' classes

In this research, an efficient optical character recognition system is presented for Off-line Arabic handwritten isolated character recognition that depends on three essential mechanisms as basic building blocks: Feature extraction mechanism, features likelihood estimation and recognition mechanism.

The recognition phase of our proposed system has two major phases: Training and Testing.

Our proposed system, which we called it: Bayesian Backpropagation Arabic Optical Character Recognition System (BBAOCR) consists of two main phases: Dataset Collection and Character Recognition where the Character Recognition Phase can be further subdivided into four major stages: Image Pre-processing, Features Extraction, Features Validation and Classification. Figure (3.1) shows the block diagram of our proposed system.

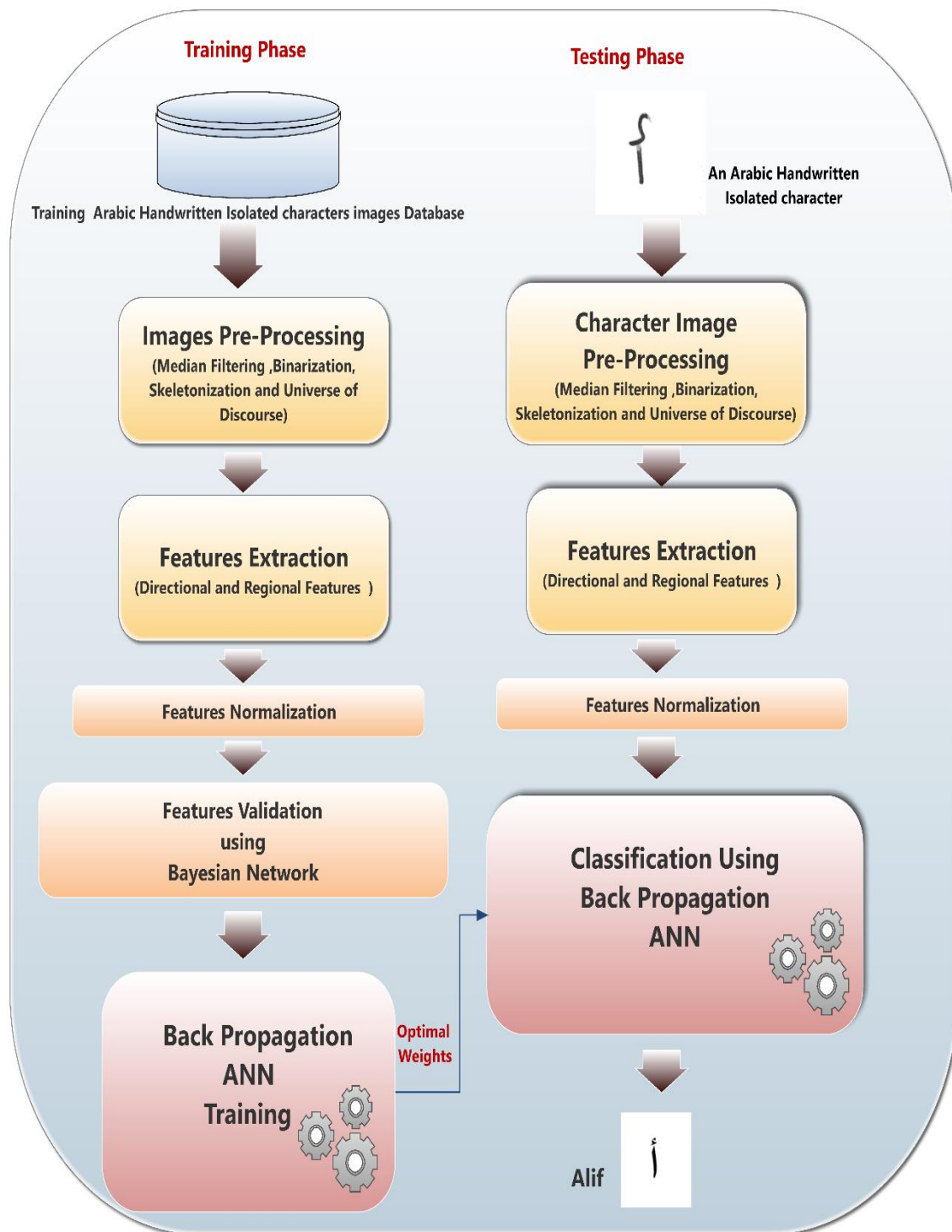


Figure (3.1): BBOCR System Block Diagram.

3.2 Phase (I): Dataset Collection

As illustrated in chapter 2, some of the popularly used datasets in the field of Arabic optical handwritten character recognition is CEMPAMI. In our research we have used CEMPAMI dataset since it is a novel dataset that composed of Off-line Arabic handwritten isolated characters and it has been used by pioneer works in the field of Arabic handwritten character recognition (Shalol, et.al, 2014A), (Shalol, et.al, 2014B) and (Jamal, 2015).

Figure (3.2) shows the general structure of CEMPAMI Arabic dataset. In our research, we use the isolated grayscale letters.

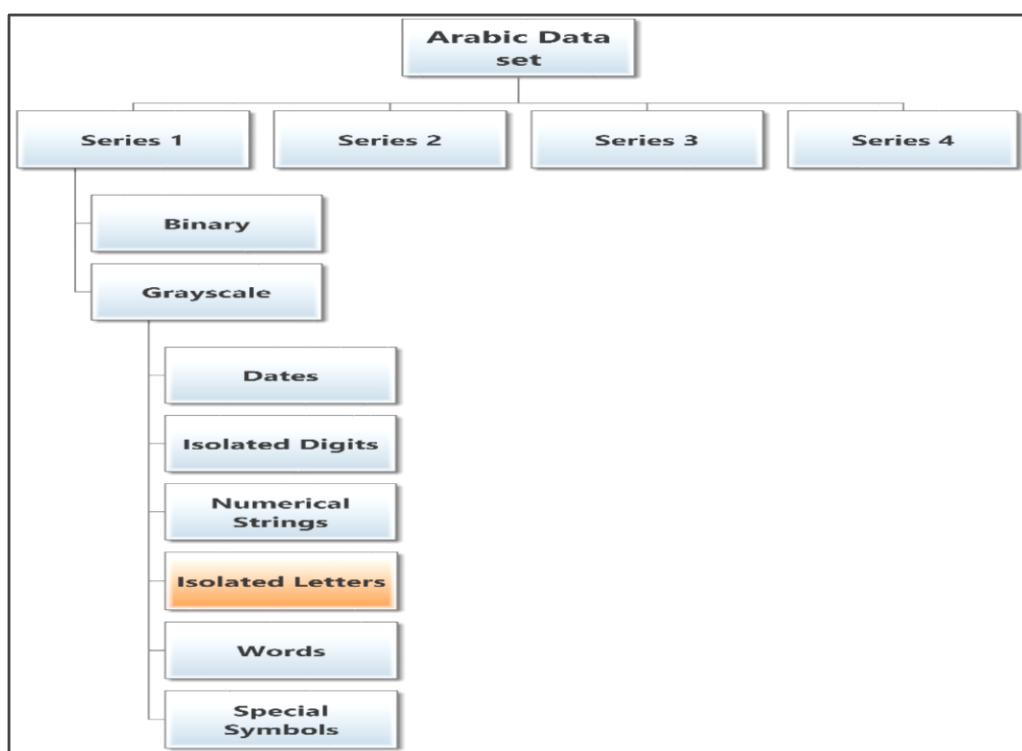


Figure (3.2): The general structure of CENPARMI dataset form (Alamri, et al.,2008)

3.2.1 Character Labels Codification

Coding can be define as the process of examine the raw qualitative data by assign labels or codes to any piece of data whether phrases, words sentences or paragraphs

In our case, dataset itself is just a set of extracted scanned character images, however, the codification here involves the string that represent the label of the letter that used to label the extracted feature of each character.

Therefore, in order to be used in the process of neural network training and testing, these characters' labels (final column of dataset of extracted features) must be encoded in such a way, each character has a unique decimal number that represent it.

Where these labels will represents the targets of input vectors that will be used to train and evaluate the back propagation ANN performance.

Then, these decimal numbers must be encoded in terms of binary (1and 0) in sake of neural network training and recognition purposes. Table (3.1) shows the decimal representation and its corresponding binary one.

Table (3.1): Arabic Letters: Decimal and Binary Representations

Arabic Character	Decimal Representation	Binary Representation
Alif	0	00000
Baa	1	00001
Taa	2	00010
Thaa	3	00011
Jeem	4	00100
Haa	5	00101
Kah	6	00110
Daal	7	00111
Thaal	8	01000

Raa	9	01001
Zaay	10	01010
Seen	11	01011
Sheen	12	01100
Saad	13	01101
Daad	14	01110
Ta	15	01111
Tha	16	10000
Ayn	17	10001
Ghayn	18	10010
Faa	19	10011
Gaaf	20	10100
Kaaf	21	10101
Laam	22	10110
Meem	23	10111
Noon	24	11000
Ha	25	11001
Ha2	26	11010
Waaw	27	11011
Waaw2	28	11100
Hamza	29	11101
Yaa	30	11110
Yaa2	31	11111

3.3 Phase (II): Character Recognition

This phase consists of four major cascaded stages, where first the character image undergoes a pre-processing step that enable the character image ready for the feature extraction stage. Then, the dataset of features vectors must be checked if it suitable to represent the character image and can this representation distinguish it from another character, which ensure that the classification via back propagation ANN will yield high performance. Figure (3.3) shows these stages of this major phase.

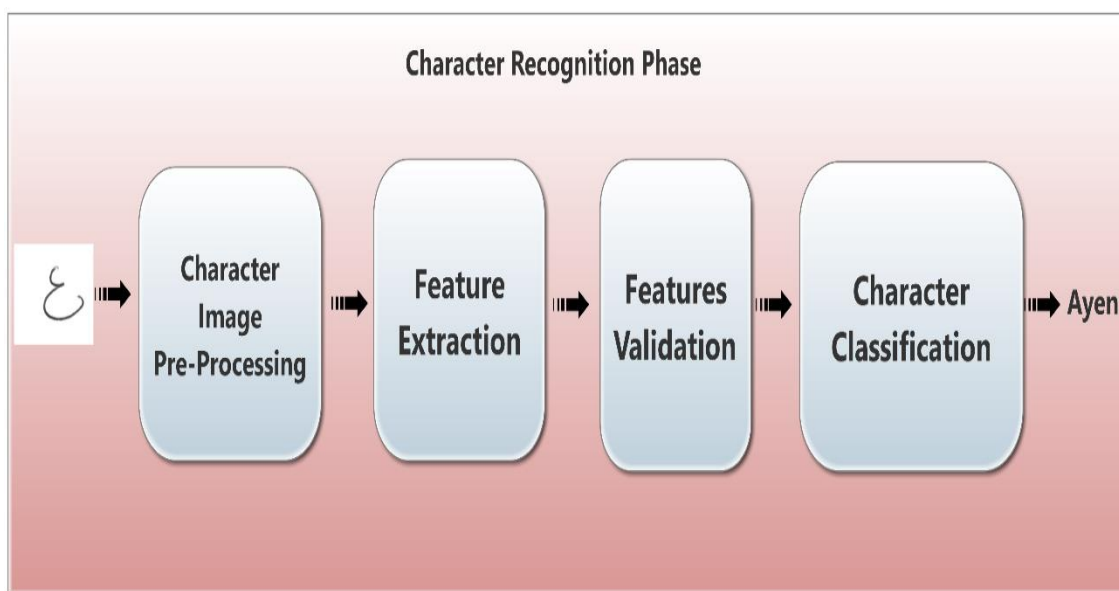


Figure (3.3): The Stages of Character Recognition Phase

3.3.1 Stage (I): Image Pre-processing

Image pre-processing involves three major steps:

- Background Noise removal.
- Binarization.
- Skeletonization.
- Universe of Discourse.

3.3.1.1 Background Noise removal

Although all character images of CENPARMI database are filtered and de-noised, we have used the Median filtering technique (Lim, 1990) in order to over enhanced grey images. This type of filter reduce the random noise that may exist in the character grey images.

Therefore, we have utilized a 3*3 median filter since it yielded the best de-noised results. Figure (3.4) shows the image of Kaaf (ك) character before and after median filter pre-processing.

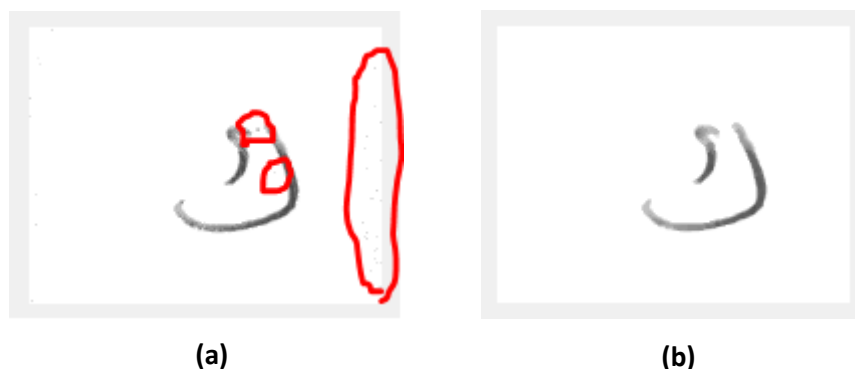


Figure (3.4): Character Kaaf(ك) (a) Before Filtering. (b) After Filtering

3.3.1.2 Binarization

Our database, which consists of grey images, so it needs to be converted into binary image, which is normalized intensity that lies between (0) and (1).

This process is called the binarization .We have used Otsu's method (Otsu, 1975) to convert the grey images into binary ones.

We can use the default value of binarization of (**0.5**). However, we will lose the useful information of the character.

Since MATLAB enable you to select the intensity level and calculated it by Otsu's method; we have replaced all pixels in the input image with luminance that greater than the Otsu's level with the value 1 (white) and all other pixels with the value 0 or (black). The process of binarization can be done easily in MATLAB in the following code:

```
% Call image and convert it to bitwise image
image=imread('kaaf.tif');
image=im2bw(obw,graythresh(image));
```

Figure (3.5) illustrates the idea of binarization.



Figure (3.5): Character Kaaf (ك) (a) Before Binarization. (b) After Binarization

3.3.1.3 Skeletonization

Where the skeletonization is considered one of the morphological operations where the pixels on the boundaries of object (character) are removed but at the same time it does not allow the object (character) to break apart. The remaining pixels make up the image skeleton. In addition to binarization, the skeletonization can easily done in MATLAB too as the following code:

```
% Create the skeleton of an image
image = bwmorph(image,'skel',inf);
```

Figure (3.6) shows the character **Waaw** (و) undergoes a skeletonization process.

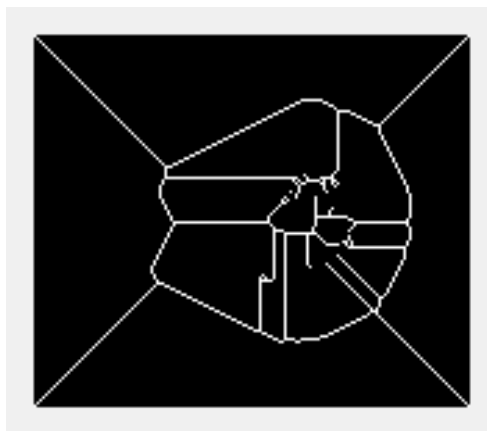


Figure (3.6): The Skeleton of character Waaw (و)

3.3.1.4 Universe of Discourse of Character Image

The universe of discourse of a character image can be defined as the shortest universe that encloses the character itself. In our research, the image is represented as a matrix of ones and zeros that represent the character itself and the white space around it respectively. Therefore, the universe of discourse will be the shortest matrix that fits the whole character skeleton. Figure (3.7) and Figure (3.8) show the universe of discourse of letters: **Ha** (هـ) and **Saad** (ص) respectively.



Figure (3.7): (a) Original Ha Character image. (b) Universe of Discourse



Figure (3.8): (a) Original Saad(ص) Character image. (b) Universe of Discourse

The universe of discourse is important to be performed before the zoning take place because the features extracted out of the character image depends on the positions of different line segments in the character image so it is of essential importance that the character image should be independent of its size.

3.3.2 Stage (II): Features Extraction

Two major types of feature extraction techniques were investigated in the implementation of our proposed BBAOCR system: (1) Directional Features and (2) Regional Features. The following subsections will discuss these types of features extraction more profoundly.

3.3.2.1 Directional Features

We highly depend on the novel directional feature extraction technique that proposed by (Blumenstien, Verma& Basli, 2003) which were used with English handwritten letters. This novel technique sought to simplify (divide and conquer) the skeleton of a character through identification of individual stroke on line segments in the character image.

Dileep (2012) implement this feature extraction technique through zoning the character image into 3x3 (nine) zones and extract the features of each zone individually.

In our research, we have used three zoning techniques. Two of which are proposed and implemented for English letters by (Dileep, 2012). Where we have divided the character image into three horizontal zones and 3x3 zones.

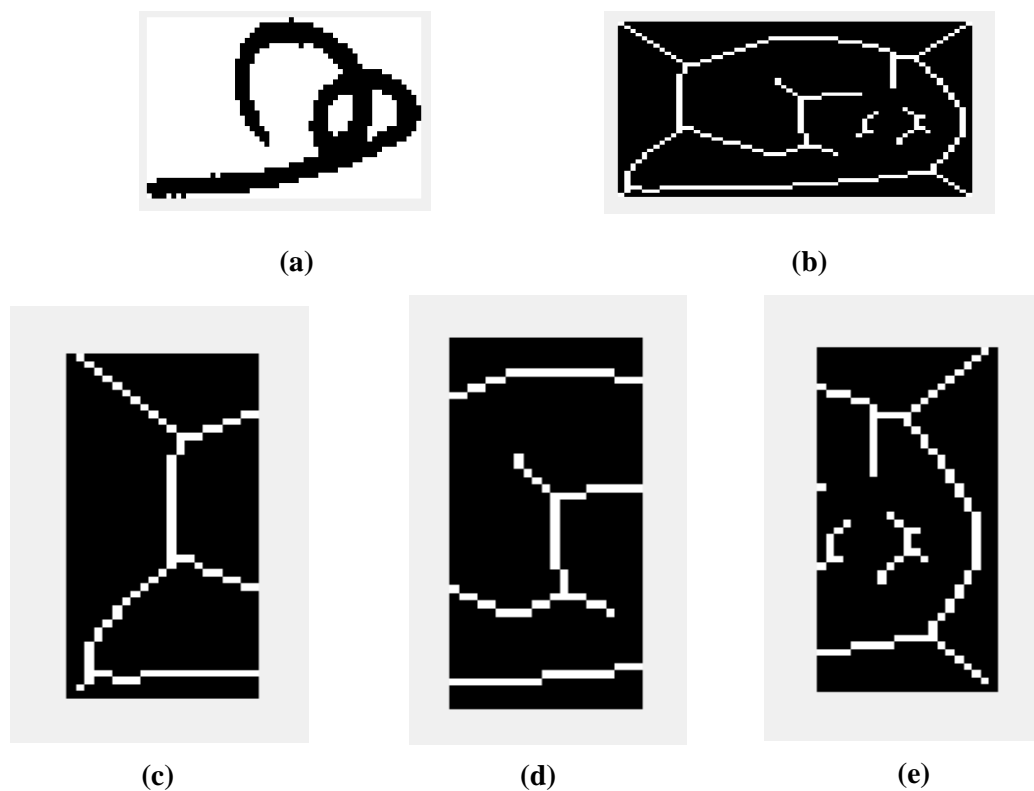
For further enhancement, in our research, we have divided the image into three vertical zones then we have extracted the features of each zone individually and concatenated the features of each type of zoning in one feature vector for each character image.

By this approach, we guarantee that all fine details of Arabic character are taken under consideration which help the back propagation ANN in the classification stage dramatically.

3.3.2.1.1 Zoning

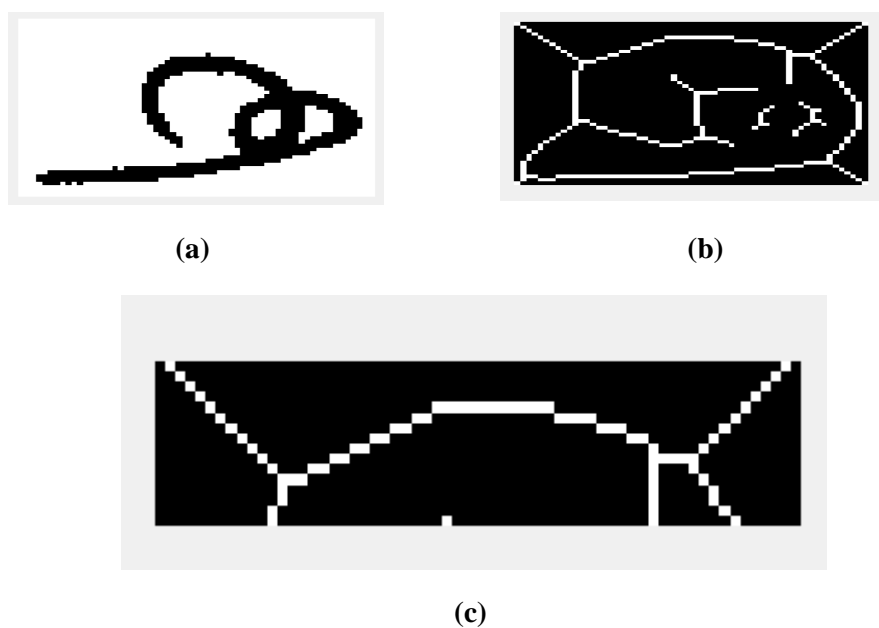
After the universe of discourse is determined, the character image is divided into two types of zoning: (1) zoning along one dimension (along the row or column dimension of image) or (2) two-dimension zoning.

In the case of one-dimension zoning, the character image has been divided into three zones along one of image dimension, namely, one time along the rows and the other zoning along with column dimension. Figure (3.9) shows zoning along the column dimension of the character image respectively.



**Figure (3.9): (a) Original Ha (ハ) Character image. (b) Skeletonized image
(c) Column zone_1 (d) Column zone_2 (e) Column zone_3.**

Figure (3.10) represents the row zones that yielded when the skeletonized character image divided into three horizontal zones.





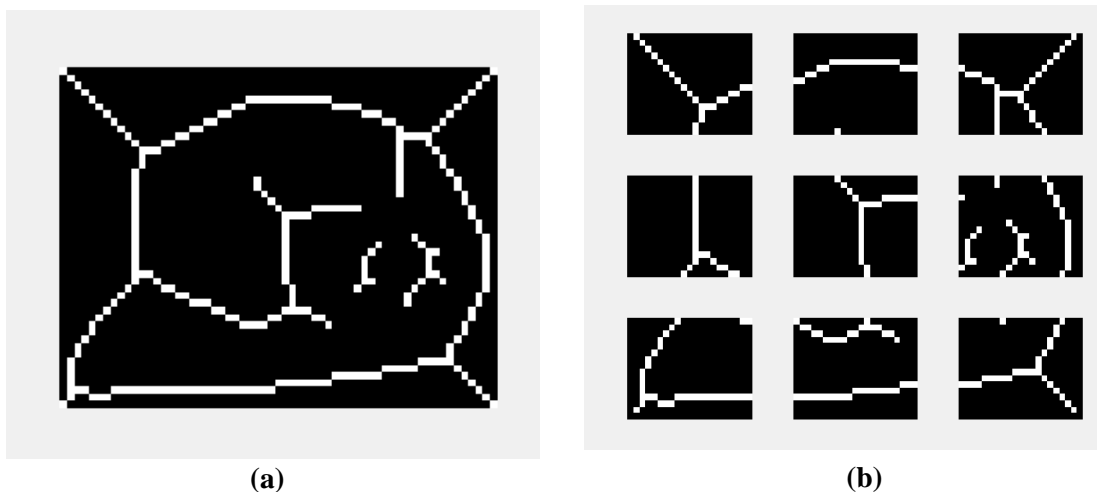
(d)



(e)

**Figure (3.10): (a) Original Ha (↵) Character image. (b) Skeletonized image
(c) row zone_1 (d) row zone_2 (e) row zone_3**

In case of two-dimensional zoning, the character image zoned into nine equal sized windows as shown below in Figure (3.11)



(a)

(b)

Figure (3.11): (a) Skeletonized Ha (↵) image (b) zoned Ha (9 zones)

Feature extraction will be applied on these individual zones rather than the entire character image. This yields more information about very fine details of the character skeleton.

Moreover, the zoning technique enable us to consider the positions of different line segments in a character skeleton as features and it is a natural result because a specific line segment of a character may occur in particular zone and do not occur in other in almost cases of characters.

As an example, the “innermost circle in character **Ha** (↵) occurs in one zone of the entire character zone as illustrated in Figure (3.12)



Figure (3.12): Zone that contain the innermost circle of Ha (↵) character.

3.3.2.1.2 Starters, Minor Starters and Intersections

As elaborated in the zoning section, the feature extraction technique highly depends on the different typed of segments that can be defined in each zone as shown in Figure (3.13), where if we consider the zones of the image it has the following matrix:

$$\begin{bmatrix} \text{zone11} & \text{zone12} & \text{zone13} \\ \text{zone21} & \text{zone22} & \text{zone23} \\ \text{zone31} & \text{zone32} & \text{zone33} \end{bmatrix}$$

In Figure (3.13), zone of 33 positions in the matrix above is extracted and the line segments of this particular zone are highlighted.

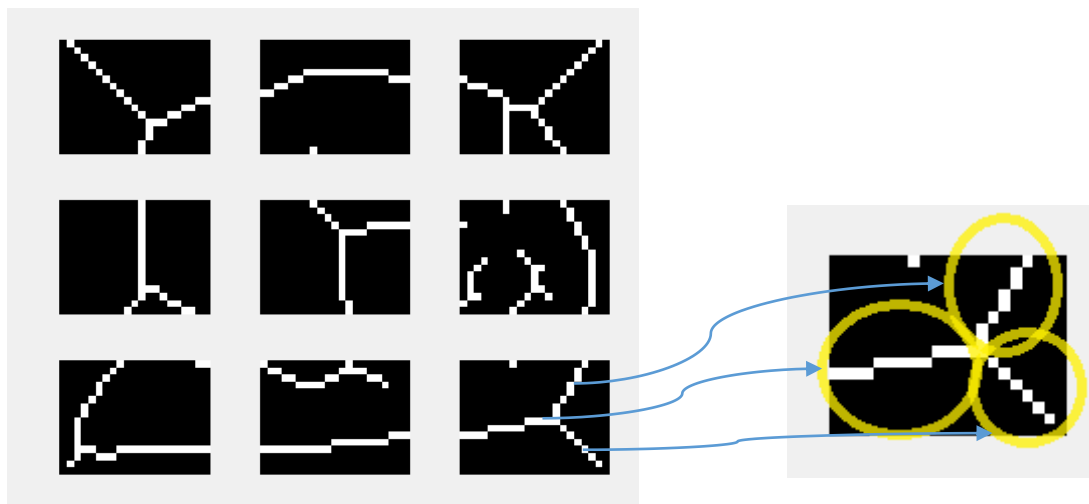


Figure (3.13): Highlighted Line Segments of a Particular Zone

To identify the line segments that will establish our features database, certain pixels in the character skeleton were first must be defined as: starters, minor starters, intersections, and the entire skeleton of that zone will be traversed entirely pixel by pixel.

3.3.2.1.2.1 Starters

Starters are defined as those pixels with one neighbour where the neighbourhood of a pixel is defined as all pixels that immediately surrounds the pixel under consideration. Figure (3.14) elaborates both of these concepts, where the pixel under consideration is coloured in darker colour rather than that of neighbours around it.

Neighbourhood can come in direct or diagonal directions, which means that each pixel has eight neighbour pixels: four of them in diagonal, two vertical and the other horizontal.

<-----Zone----->

1	0	0	0	1	0	0	1	1
0	1	0	1	0	0	0	0	0
0	1	1	0	0	0	1	0	0
0	0	1	0	1	0	1	0	0
0	0	1	0	1	0	0	0	0
1	1	0	0	0	1	0	0	0
0	1	0	0	0	1	0	0	0

Figure (3.14): Direct and Diagonal Neighbourhood of a Pixel

Before the traversal of the character skeleton begin, all the starters in the particular zone of the character image are identified and then populated in an array. In order to illustrate the concept of starter pixel let Figure (3.15) represents the skeleton of a particular zone of a character.

<-----Zone----->

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	0	0	0
0	1	1	0	1	0
1	0	0	0	0	1

^
-----Zone-----
v

Figure (3.15): The skeleton Pixels (ones) and Starters of a Particular Zone.

Pixel in the left top corner is numbered as (1,1) and the pixel in the right lower corner is labelled as (5,6) where we have five rows and six columns in this particular zone and let the standard convention for numbering matrices is assumed so forth.

Starters list in this image would be [(1,1), (1,6), (5,1), (5,6)] and can be noted that these pixels are really represent the starting points in this particular zone. Figure (3.16) shows the starters of **Tha** (ظ) and **Meem** (م) characters respectively:

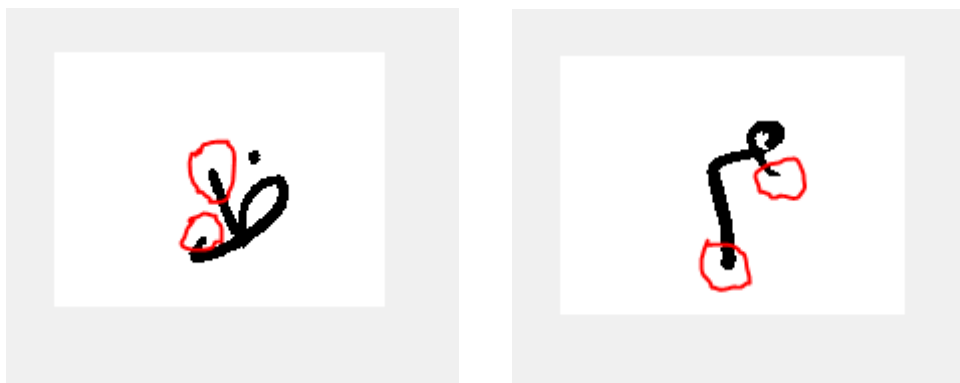


Figure (3.16): starters of Tha (ظ) and Meem (م) characters

3.3.2.1.2.2 Intersections

Intersections can be defined as those pixels that have more than two neighbours. Although this criterion is not sufficient but it is necessary to define a pixel as an intersection point between multi line segments. Therefore, the process of identification of intersection points is somewhat more complicated than that in case of starters. Figure (3.17) shows some of intersection points in **Ta** (ط) character.



Figure (3.17): Intersection Points in Ta (ط) Character

To overcome this issue, a new property called true neighbours is defined for each traversed pixel based on the number of true neighbours for a particular pixel it is classified as an intersection point(pixel) or not.

For this reason, neighbouring pixels have been classified into two main categories: Direct pixels and diagonal pixel as was illustrated above.

Now, in sake of finding the number of the neighbours for specific pixel (pixel under consideration). Firstly, it has to be classified further based on the number of neighbours it has in the character skeleton.

We have three cases of neighbourhood that will determine if the current pixel is a true intersection or not (Dileep, 2012):

➤ **Case (1): Three neighbors**

if

Any one of the direct pixels is adjacent to anyone of the diagonal pixels

Then

The pixel under consideration cannot be an Intersection

else if

none of neighbouring pixels are adjacent to each other

then

The pixel under consideration is an intersection

To illustrate this case, lets us return back to our matrix of ones and zeros that represents the binary pixels of a particular zone or as shown in Figure (3.18):

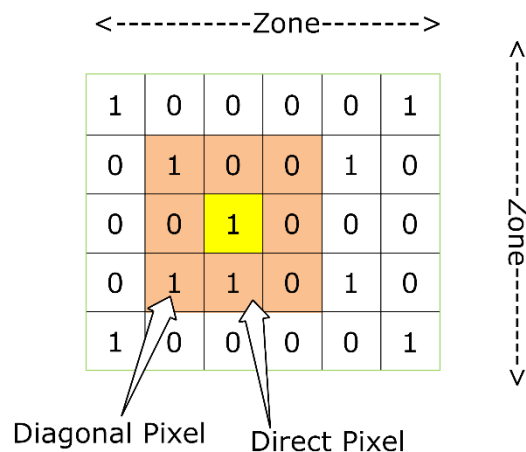


Figure (3.18): Neighbour pixels: Diagonal and Direct in the window around pixel under consideration

Let ones represent the skeleton of character in this particular zone. The pixel under consideration is highlighted in yellow colour. As noted in Figure (3.18), one of the direct pixel is adjacent to diagonal pixel; therefore, the pixel under consideration cannot be an intersection.

➤ **Case (2): Four neighbors**

If

Each and every direct pixel have an adjacent diagonal pixel or vice versa

Then

The pixel under consideration cannot be considered as an Intersection

➤ **Case (3): Five neighbors or More**

If

pixel under consideration have five or more neighbors

Then

it always considered an Intersection

Figure (3.19) shows this case when the pixel under consideration has five neighbours:

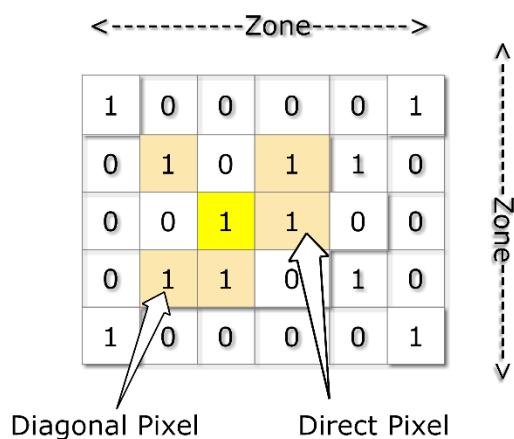


Figure (3.19): Five Pixels- Neighbourhood of a Pixel.

Once all the intersections are identified in each zone of character image, then they have been populated.

3.3.2.1.2.3 Minor Starters

Minor starters are created when the pixel under consideration has more than two neighbours. Minor starters pixels can be found when the skeleton of character is traversed. There are two major conditions when they met, minor starters can occur.

➤ **Condition (1):** Intersections

When the pixel under consideration is an intersection, then the line segment under consideration will end there and consequently all the unvisited neighbours are populated in the list of minor starters (considered as minor starters). Figure (3.20) illustrates this concept.

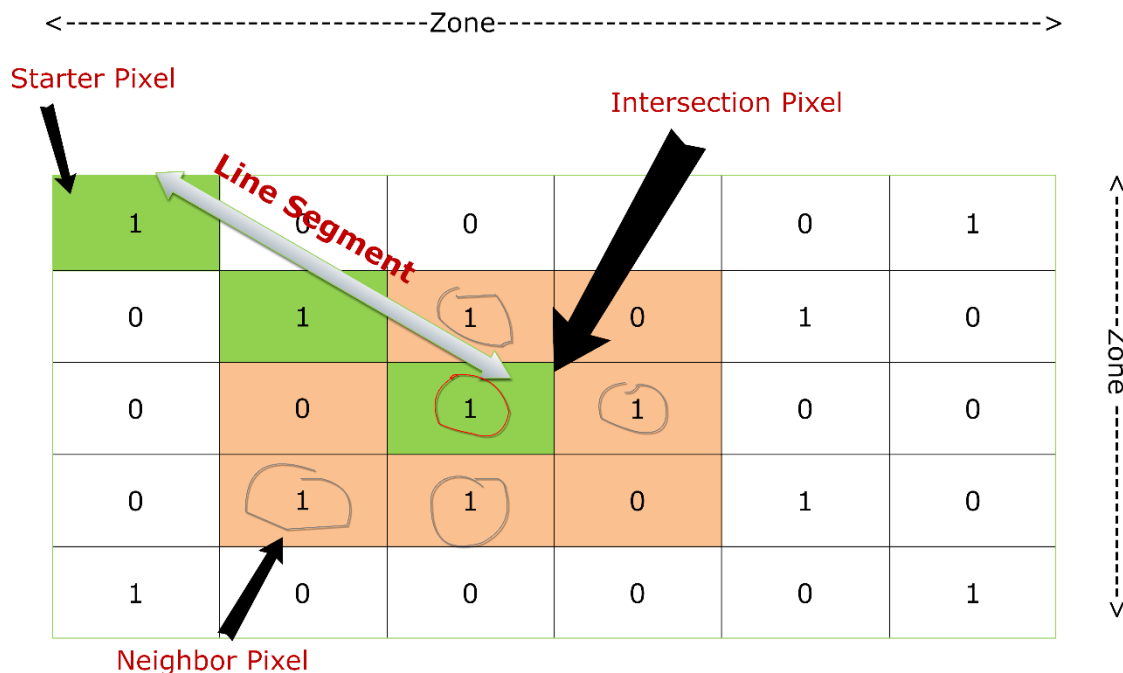


Figure (3.20): Minor Starters of a Particular Pixel.

➤ **Condition (2): Non-Intersections**

This condition occurs when the pixel under consideration has more than two neighbours but still it is not an intersection. In such condition, we lost the direction since intersections do not appear yet. Therefore, the current traversal is found by using the position of the previous pixel.

Now, if any of the unvisited pixels in the neighbourhood is in this traversal direction, then it is considered as the next pixel in the direction and all other pixels are considered minor starters to be listed in the minor starters list.

On the other hand, if none of the pixels is not in the current direction, then the current segment line is ended there and all the rest of pixels that exist in the neighbourhood are populated in the minor starters list.

3.3.2.1.3 Character Skeleton Traversal

Once the zoning phase has been finished, the skeleton of character image undergoes traversal process where each zone of is separately subjected to the process of line segments extraction.

In aim of this phase, the starters and intersections are firstly found and then they populated in arrays. At the same time of skeleton traversal, minor starters are populated simultaneously.

The novel algorithm that proposed by (Dileep, 2012) which we used it in to extract features of our BBAOCR system, begin first by finding the starters list. Once all the starters are processed in order to obtain the line segments, and the minor starters at the same time processed., then the algorithm starts with minor starters and all the line segments that obtained are populated and stored to be labelled and processed later on. Once all the pixels of the character skeleton have been visited, the algorithm stops.

3.3.2.1.4 Distinguishing Individual Line Segments

Now, once line segments have been extracted from the character image, they have to be classified into one of the following line segment types:

- Right Diagonal Line.
- Left Diagonal Line
- Vertical Line
- Horizontal Line.

Each character pattern compromised of these four types of line segments that mentioned above. After starters and intersections have been determined, the neighbouring

pixels along the thinned character skeleton were followed from the starting points until we reach an intersection point.

Once we arrive at intersection, the clockwise searching begins in order to determine the beginning and the end of the individual line segments.

The incipience of a new line segment is located **IF**:

1. The previous direction was up-right or down-left **AND** the next direction is down-right or up-left **OR**
2. The previous direction is down-right or up-left **AND** the next direction is up-right or down-left **OR**
3. The direction of a line segment has been changed in more than three types of direction **OR**
4. The length of the previous type is greater than three pixels.

These rules that illustrated above reviewed for each set of pixels (that compose the segments) in each character Skeleton (pattern).

3.3.2.1.4.1 Line Segment Information Labelling

In the previous subsection, we discussed the issue of individual line segments location. Now, the white pixels that composed each segment will be encoded with a direction number in the following manner.

After the starters and intersections of each segment is identified, the segments of the zone should be labeled, that means the directions pixels that compose each segment are identified in one of eight directions: Up, right, left, down, up-right, up-left, down-left and down-right, where the directions take the following values that illustrated in Figure (3.21):

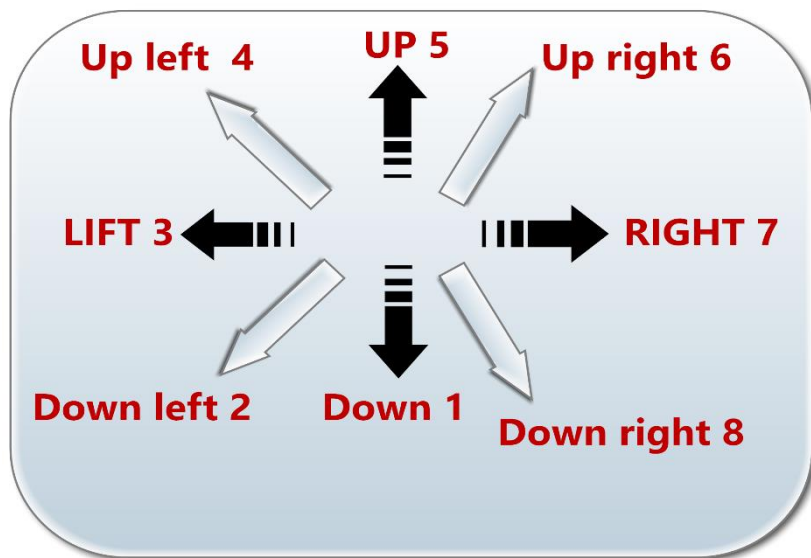


Figure (3.21): Pixel direction

Then the segment takes the label of the most frequent direction. As an example, suppose that we have the following zone shown in Figure (3.22)

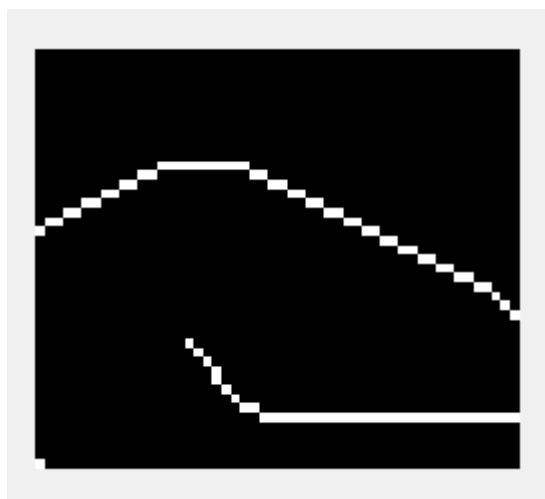


Figure (3.22): One zone of nine zones of character image.

We can identify two segments in this zone, as illustrated in the Figure (3.23):

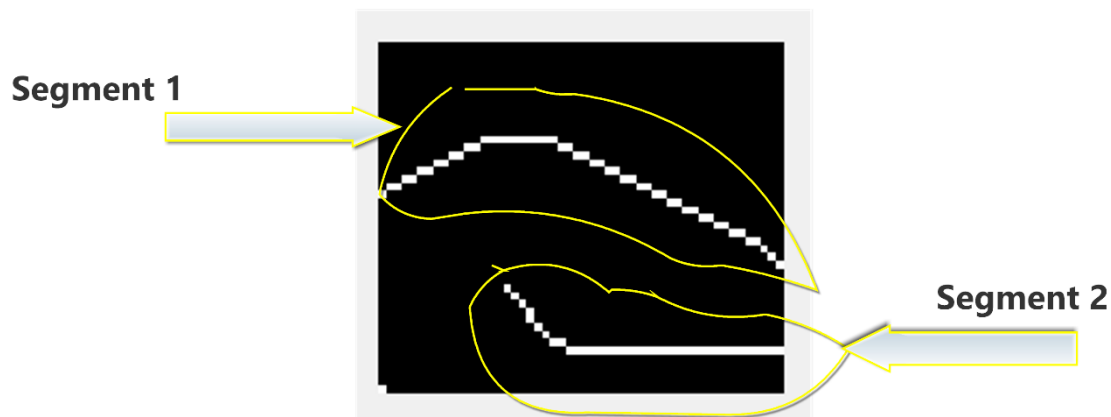


Figure (3.23): Identified Segments of the given zone

However, these segments contain spurious pixels that composed another direction, thus, we identify the direction of each pixel and the most frequent direction is given as a label for that segments as illustrated in the Figures (3.24) and (3.25):

Variables - Truelines{1, 1}			
Truelines{1, 1}			
	1	2	3
41	8		
42	7		
43	8		
44	7		
45	8		
46	7		
47	8		
48	7		
49	8		
50	8		
51	8		
52			
53			

Figure (3.24): Pixels' directions of segment1

As we note, the direction (8) which represents down right is the most frequent direction in segment 1, whereas the pixel direction of (7) (which represent right direction) is the most frequent direction in segment 2 as illustrated in Figure (3.25):

	1	2	3	4	5
7	8				
8	7				
9	8				
10	7				
11	7				
12	7				
13	7				
14	7				
15	7				
16	7				
17	7				
18	7				
19	7				

Figure (3.25): Pixels' directions of segment2

Therefore, we can conclude that segment (1) has the label of (8) and segment (2) take the label of (7), which are numerical values that will be processed in order to be suitable as inputs for the Naïve Bayesian classifier on one hand, and suitable inputs for the back propagation artificial neural network on the other hand.

3.3.2.1.5 Estimation of Feature vectors Through Zoning

(Blumenstein, et.al, 2003) who suggested the feature extraction techniques mention above, have also developed a methodology for creating appropriate feature vectors in such a way, it is suitable and uniform in its size to be used as inputs for both: Naïve Bayesian classifier and Back propagation artificial neural network.

The first step of this methodology is to zone the character pattern that marked with direction information into windows of equal size. Now, if the image matrix was not divisible in equal manner, then it was padded with additional background pixels along the length of both: its rows, and its columns.

The next step is to extract the direction information out of each individual window where the direction information includes: line segment direction , the intersection points, length, starter points and it expressed in floating point values between (-1 and 1).

The extraction and storing algorithm of line segment information proceeds in the following couple of steps:

Frist: locate the starting point and the intersection points in the window under consideration.

Second: Extract the number and the length of line segments.

These steps yield an input vector composed of nine floating-point values. Now, each value in this vector is defined as follows:

1. The number of right diagonal lines.
2. The total length of right diagonal lines.
3. The number of horizontal lines.
4. The total length of horizontal lines.
5. The number of vertical diagonal lines.
6. The total of vertical diagonal lines.
7. The number of left diagonal lines
8. The total length of left diagonal lines.
9. The number of intersection points.

Figure (3.26) illustrates a portion of feature vector (blue highlighted) that formed by the algorithm explained above.

	7	8	9	10	11	12	13	14	15	16
1	0	0.9778	0.0169	0.8000	1	0.6000	1	0.3488	0	0.6
2	0	0.9778	0.0169	0.8000	1	1	1	0.9808	0	
3	0	0.9778	0.0169	0.6000	1	1	1	0.9773	0	
4	0	0.9778	0.0169	1	1	0.8000	1	0	0	0.9
5	0	0.7857	0.0211	0.6000	1	1	1	0.9744	0	
6	0	0.6027	0.0275	0.4000	0.8000	0.8000	0.6000	0.7064	0.0734	0.0
7	0	0.9778	0.0169	1	1	0.8000	1	0	0	0.9
8	0	0.9778	0.0169	1	1	0.8000	1	0	0	0.9
9	0	0.9778	0.0169	0.6000	1	1	0.8000	0.9394	0	
10	0	0.9778	0.0169	1	1	0.8000	1	0	0	0.9
11	0	0.9778	0.0169	0.8000	1	1	1	0.9545	0	
12	0	0.9778	0.0169	0.2000	0.6000	0.8000	1	0.7481	0.1778	0.0

Figure (3.26): Feature Vector of a character image

Example:

If we assume that, the value of (1) represents 'no line' state in the window under consideration, and if the window contains a vertical line, then the value decreases by (0.2) to give a value (0.8). (Blumenstein, et.al, 2003) suggested a decrement by (0.2) because in their preliminary experiments, they found that the average number of lines that find a specific direction in a specific window is (5); so ($\frac{1}{5} = 0.2$).

The equation that controls this process, namely, the equation that evaluates the values of a specific line type in a particular window can be expressed mathematically as in Equation (3.1):

$$Value = 1 - \left(\left(\frac{\text{number of lines}}{10} \right) * 2 \right) \quad (3.1)$$

Hence, for each value that recorded the number of lines that present in a specific window, there is a corresponding input value that recorded the total length of lines was also stored.

As an illustration, let the number starts at (0) to represent "no vertical lines " in the window under consideration. Then if the window contains a vertical line, then the input will increase by the length of the line divided by maximum window length multiples by (2) which is mathematically expressed as in Equation (3.2):

$$length = \frac{\text{number of pixels in a particular direction}}{(\text{window length or width}) * 2} \quad (3.2)$$

Example:

If the vertical line mentioned above has length of (8) pixels and the window height (or maximum length) is (12) pixels by (15) pixels, then the line length will be:

$$8 / (15 * 2) = 0.2666$$

The encoding operations that have been discussed for the vertical line information should be drawn on the rest of directions.

Now, the last input value of the feature vector is the number of intersections points in the character image, which is calculated in the same approach, followed for the number of lines present in Equation (3.1).

3.3.2.2 Regional Features

In addition to the directional features that extracted above, we have extracted the regional features of an image, which include *Euler number*, image *Eccentricity*, image *minimal bounding box* , *Orientation* and *Extent*.

3.3.2.2.1 Euler Number

Euler Number of an image is considered one of the most important regional features that majorly used to describe the topological structure of an image.

This type of regional feature is not affected by a variety of image transformations such as rotation, scale-changed, projects, affinities, translation and some of non-linear transformation such as the process of deformation of the shapes that exist in the image (Sossa-Azuela, et al, 2010).

Euler number of a binary image is expressed mathematically as explained in (Rosenfeld & Kak, 1982; Gonzalez, Woods, 1993; Zenzo, Cinque & Levivald, 1996) or as Equation (3.3):

$$E = N - H \quad (3.3)$$

Where: N: is the number of regions of the image or in other words, it represents the number of connected components of object.

H: represents the number of holes in the image or the number of isolated regions of the image's background. Figure (3.27) elaborates this important concept.

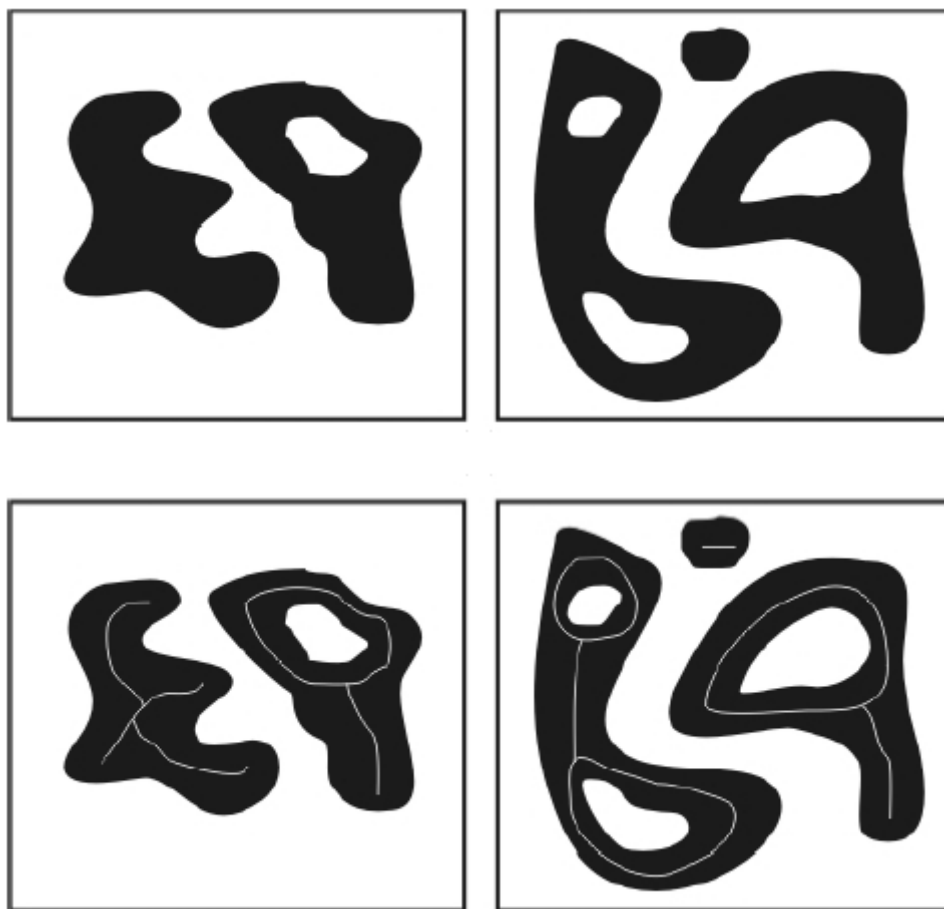


Figure (3.27): Euler Number Concept (Sossa-Azuela, et al, 2010)

In order to illustrate the Euler number implemented in our proposed BBAOCR, we apply Equation (3.3) to two Arabic characters: **Saad** (ص), **Faa** (ف). Figure (3.28) illustrates the application of Euler number.

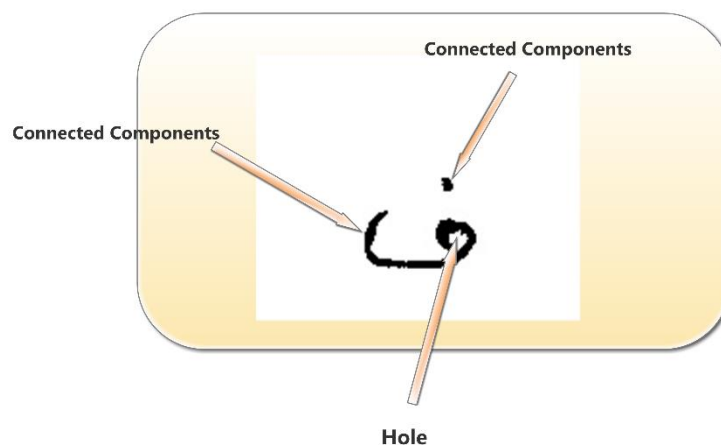


Figure (3.28): Number of regions and Holes in Faa(ف) Character.

As shown in Figure (3.28) **Faa(ف)** character has two regions and one hole. Therefore, Euler number equals: $2-1=1$ whereas **Saad(ص)** character have one hole and one connected region, thus, Euler number equal zero .as can be shown in Figure (3.29).

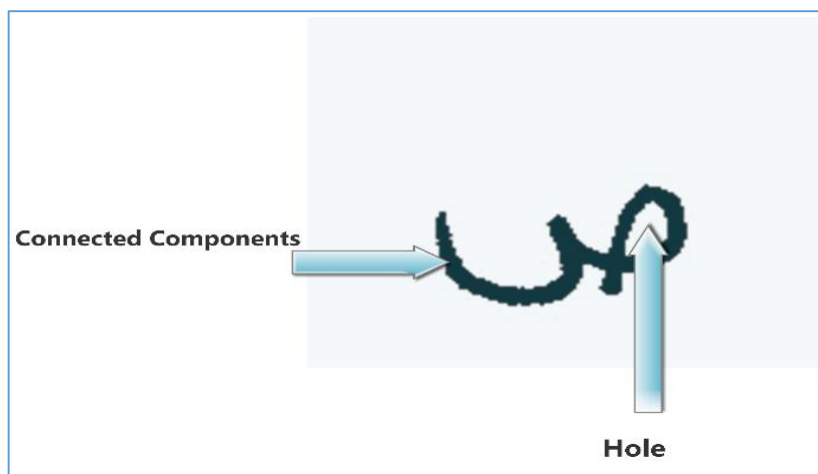


Figure (3.29): Number of regions and Holes in Saad(ص) Character

3.3.2.2.2 The Eccentricity

The **Eccentricity** of an object is called elongation too. It is defined as the ratio of the l_c to l_p or as illustrated in the Equation (3.4):

$$\text{Eccentricity} = \frac{l_c}{l_p} \quad (3.4)$$

Where l_c and l_p represent the lengths of the longest and the longest perpendicular line that connected between two pixels on the object. Figure (3.30) shows a couple of examples of objects with low and high eccentricity.

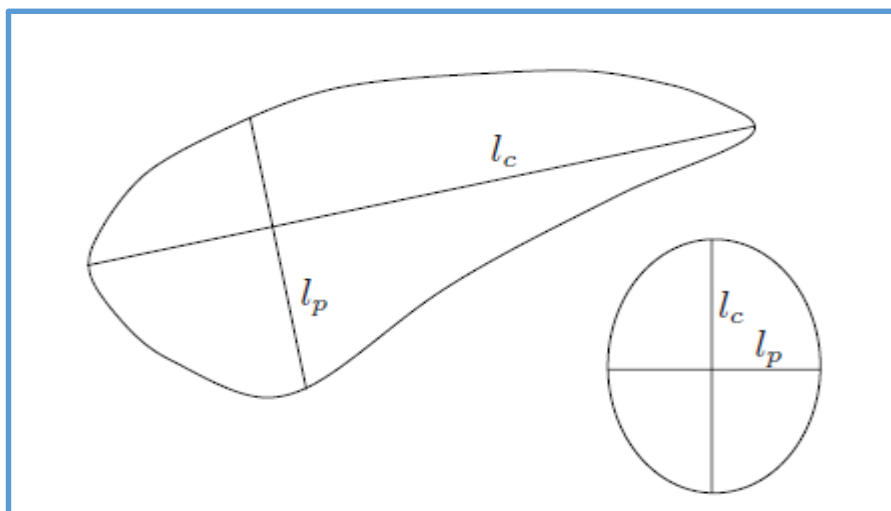


Figure (3.30): High and Low Eccentricity (Maintz, 2005)

In order to illustrate this concept on our case, we apply extract this feature out of two Arabic characters: **Alif (ا)** and **Hamza(ء)**.

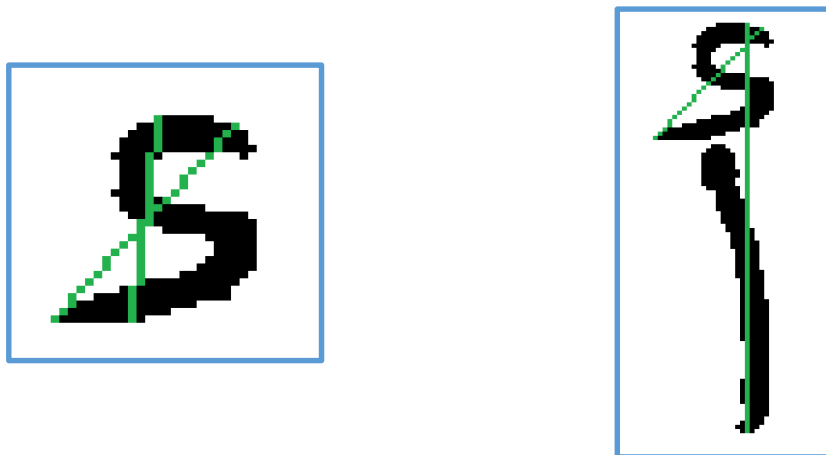


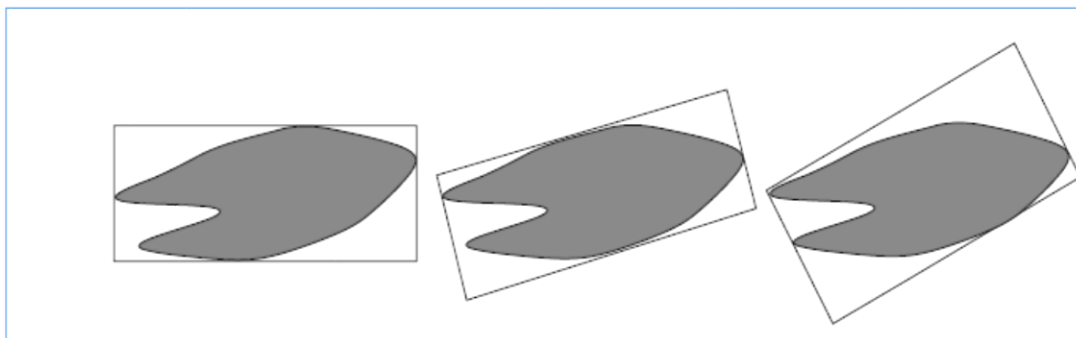
Figure (3.31): Eccentricity of Alif (ا) and Hamza (ء)

As shown in Figure (3.31), the eccentricity of **Hamza (ء)** is totally different than that for **Alif (ا)** which proves the high efficiency of this regional feature in the classification phase.

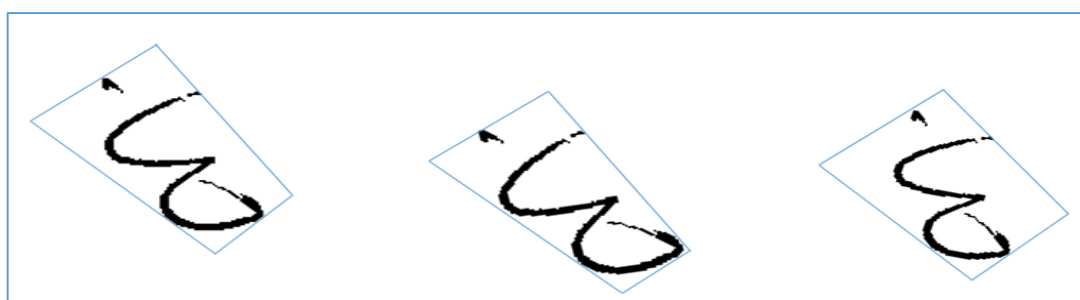
3.3.2.2.3 Minimal Bounding Box

Minimal Bounding Box area is defined as the bounding box that can be drawn around the object (character in our case) in an image in a rectangular form containing the object and where the sides of this rectangle touch the object. However, the question is what is the relation between the boundary box of an image and the extended orientation of an image?

To answer this question, let us assume that the character in an image is represented as an object, where the bounding box of an object works exactly in same manner as the universe of discourse work, namely, it contains the object only. This is illustrated through a graphical comparison between arbitrary object in different orientations and **Ghayn (غ)** takes different bounding boxes associated with each orientation shown in Figure (3.32)



(a) **Bounding Box Concept** (Maintz, 2005)



(b) **Bounding Box of Ghayn (Ġ) character**

Figure (3.32): (a) Bounding Box Concept (b) Bounding Box of Ghayn (Ġ)

As shown in Figure (3.32), by examining the Bounding boxes of **Ghayn(Ġ)** character, where the image varies its orientation, we can note how the area of its bounding boxes change according to changes in character orientation (Maintz, 2005).

3.3.2.2.4 The Extent

The Extent of an image is defined as the ratio A/A_m where A is the *Area* of object and A_m represents the *Area* of minimal bounding box that can be drawn around the object. Therefore, this ratio equal one if the object is a rectangle and less than one otherwise.

3.3.2.3 Features Normalization

After features of each character images are extracted, the features dataset need to be normalized before it used as inputs to the Naïve Bayesian and back propagation ANN. The normalization phase is of essential importance for the classification where features vectors must be unique; otherwise, it will affect the similarity between each couple of feature vectors, which leads to suppression the influence of the small-valued input variables by the higher -valued input variables.

There are two basic normalization techniques to normalize the i^{th} features vector of a particular character: (1) min-max (minimum-maximum) normalization and (2) z-score normalization.(Jain, Nandakumar, & Ross, 2005)

In our research, we will use the z-score normalization which it is going in the following steps (Hu, et al., 2006):

Step (1): Features' means calculation:

In this step, the mean for each feature (column in our dataset) is calculated using the following mathematical expression in Equation (3.5):

$$\mu = \sum_{i=1}^n y_i \quad (3.5)$$

Where, y_i is the feature value for the i^{th} features vector, $i = (1, 2, \dots, n)$

Step (2): Standard deviation calculation:

In this step, the standard deviation of the respective attributes is calculated using the following mathematical formula in (3.6):

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \mu)^2} \quad (3.6)$$

Step (3): Attribute normalization in (3.7):

$$\bar{y}_i = \frac{y_i - \mu}{\sigma} \quad (3.7)$$

Where: \bar{y}_i is the normalized feature value for the i^{th} feature vector, $i = (1, 2, \dots, n)$

Figure (3.33) shows a snippet of normalized data set that obtained before the features validation phase in Bayesian classifier begins.

	23	24	25	26	27	28	29	30	31	32
1	0.4855	0.0942	0.2174	0.0942	0.0520	0.8000	0.6000	1	0.8000	0.1228
2	0	0	0.5000	0.4706	0.0256	1	0.6000	0.8000	0.8000	0
3	0.1039	0.2338	0.4805	0.1299	0.0290	1	0.4000	0.8000	0.8000	0
4	0.5556	0	0.4167	0	0.0271	1	1	0.8000	0.8000	0
5	0.1728	0	0.4815	0.2840	0.0305	1	0.8000	0.8000	0.8000	0
6	0	0.2581	0.7097	0	0.0234	0.2000	0.6000	0.8000	0.8000	0.4659
7	0.6000	0	0.3733	0	0.0282	1	0.6000	0.8000	1	0
8	0.5753	0	0.3973	0	0.0275	1	0.6000	0.8000	0.8000	0
9	0	0	0.4571	0.5143	0.0264	1	0.8000	1	0.8000	0
10	0	0	0.4571	0.5143	0.0264	1	1	0.8000	0.8000	0
11	0.5352	0	0.4366	0	0.0267	1	1	0.8000	0.8000	0
12	0.2981	0.0192	0.3558	0.2308	0.0392	1	1	0.8000	0.8000	0

Figure (3.33): The Normalized dataset

As shown in Figure (3.33) the normalized dataset is represented as a matrix, where the columns represent features (attributes) and the rows represent the character images (samples).

3.3.3 Stage (III): Features Validation Algorithm in This System

A Naïve Bayesian classifier can be defined as the simple probabilistic classifier that can be built based on the idea of Bayes' Theorem which has its roots in the Bayesian statistics with robust (naïve) independence assumption.

Therefore, Naïve Bayesian classifier can be described as independent feature model due to the underlying probability theory that form the basis for this type of classifier (Dong and Shang, 2011).

Simply, a naïve Bayesian classifier depends heavily on an essential assumption that the presence (or absence) of a specific feature of a particular class is not in relation with the presence or absence of any other features (Dong and Shang, 2011).

Naïve Bayesian classifier has inherently a probabilistic nature, hence it can be trained in a supervised learning and in efficient manner (Jiang,et.al., 2009). Thus, we first introduce the major basis of the Bayesian classifier:

1. Prior Probability

Prior probability is the probability that in built based on a subjective judgement, or historical data to assign the probability for each event

2. Posteriori Probability

This type of probability refers to the probability that can be obtained via Bayesian formula. that means this type of probability seeks for additional information by amending the prior probability.

3. Joint Probability

It is called multiplication formula too since it is the probability of the product of two arbitrary events (or cross-events) (Dong and Shang, 2011).

3.3.3.1 The Naïve Bayesian Probabilistic Model

In this type of classifier, the probability model is the conditional model that can be mathematically expressed in probability function (3.8) (Pop, 2006).

$$P(C|F_1, F_2, \dots, F_n) \quad (3.8)$$

Where over a dependent class variable C with a small set of outcomes (classes) that are conditional on many feature variables represented by vector (3.9).

$$(F_1, F_2, \dots, F_n) \quad (3.9)$$

Now, using Baeyes' Theorem will yields in Equation (3.10).

$$P(C|F_1, F_2, \dots, F_n) = \frac{P(C)P(F_1, F_2, \dots, F_n|C)}{P(F_1, F_2, \dots, F_n)} \quad (3.10)$$

The mathematical formula in (3.10) can be plainly understood as in (3.11).

$$\textit{Posterior} = \frac{\textit{Prior} \times \textit{Likelihood}}{\textit{Evidence}} \quad (3.11)$$

Practically speaking, the numerator is equivalent to the joint probability model that is given by function probability in (3.12).

$$P(C, F_1, F_2, \dots, F_n) \quad (3.12)$$

By repeated application of the basic definition of conditional probability, Probability (3.12) can be re-written as in (3.13).

$$\begin{aligned} &= P(C)P(F_1, F_2, \dots, F_n | C) \quad (3.13) \\ &= P(C)P(F_1|C)P(F_2, \dots, F_n | C, F_1) \\ &= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3, \dots, F_n | C, F_1, F_2) \\ &= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2)P(F_4, \dots, F_n | C, F_1, F_2, F_3) \\ &= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2) \dots P(F_n | C, F_1, F_2, F_3, \dots, F_{n-1}) \end{aligned}$$

Now, the assumption of “naïve “conditional independence plays its role as following:

Assuming that each feature vector(F_i) is conditionally independent of every other feature vector (F_j) in the dataset such that $j \neq i$, then you get (3.14).

$$P(F_i|C, F_j) = P(F_i|C) \quad (3.14)$$

For each $j \neq i$, therefore the joint probability model can be expressed as follows.

$$P(C, F_1, \dots, F_n) = P(C)P(F_1|C)P(F_2|C)P(F_3|C) \quad (3.15)$$

Under the assumptions in Equation (3.15) above, the conditional distribution over the class variable C can be expressed mathematically as in Equation (3.16).

$$P(C|F_1, \dots, F_n) = \frac{1}{Z} P(C) \prod_{i=1}^n P(F_i|C) \quad (3.16)$$

Where Z represent the evidence that we talked about in explanation of the Bayesian theorem in terms of simple English language, which is a scaling factor depends only on F_1, \dots, F_n , thus, it becomes a constant if the values of these features variable are known.

The Equations above represent a manageable model of a certain problem, where we can factor it in to a so-called *class prior* $P(C)$ and independent probability distribution $P(F_i|C)$. Then the Classifiers function as in (3.17).

$$\text{classify } (f_1, \dots, f_n) = (\text{argmax } P)C = c \left(\prod_{i=1}^n P(F_i = f_i | C = c) \right) \quad (3.17)$$

3.3.3.2 Model Parameter Estimation

Model parameters represented by: class prior probability and the feature probability distribution can be approximated with the concept of relative frequency evaluated based on the training dataset. These are considered the likelihood estimates of the probabilities.

To estimate the parameters of features variables probability distribution, we first assume a particular distribution for the features depending on the training dataset too. One of the most used distributions is the Normal distribution or as so-called Gaussian distribution. In this research, we first assume that the features vectors of our training dataset is normally distributed.

Figure (3.34) shows the probability density function of the Gaussian distribution function given by the following in Equation: (Ribeiro, 2004).

$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)} \quad (3.18)$$

Where, Here, μ is the mean of the distribution .The parameter σ is its standard deviation.

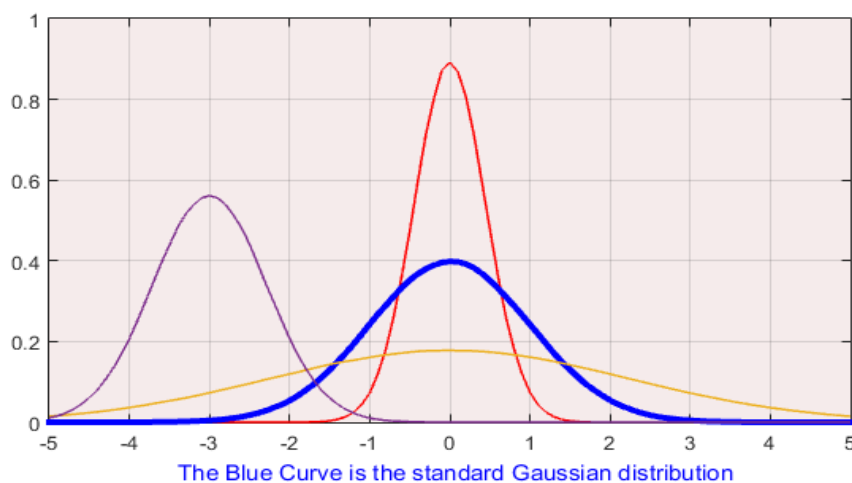


Figure (3.34): Normal Distribution Function (Bowman, et al., 1997)

However, the Gaussian distribution is commonly used with continuous random variables, and in our case, we have discrete valued feature variables. Thus, Gaussian distribution is not the optimal choice, as we will see in the experimental results. (Ribeiro, 2004).

To overcome this issue, we proposed Kernel distribution estimator. Kernel distribution works as following:

Let (x_1, x_2, \dots, x_n) be an independent and identically distributed sample drawn from some distribution with an unknown density function f . We are interested in estimating the shape of this function f . Then, its kernel density estimator is given by (Zambom, & Dias, 2012) as in (3.19).

$$\begin{aligned} f_h^\wedge &= \frac{1}{n} \sum_{i=1}^n K^h(x - x_i) \\ &= \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) \end{aligned} \quad (3.19)$$

Where: $K(\cdot)$ is the kernel which is a non-negative function that has unit integration and zero mean.

h : is the bandwidth of the kernel function with subscript(h) so the kernel is called

scaled kernel if : $K_h(x) = \frac{1}{hK(\frac{x}{h})}$

Kernel function could be one of the well-known distribution functions such as Uniform, biweight(Quartic), normal, Triangular and other. Figure (3.35) shows one probability density function shape that estimated by Kernel distribution.

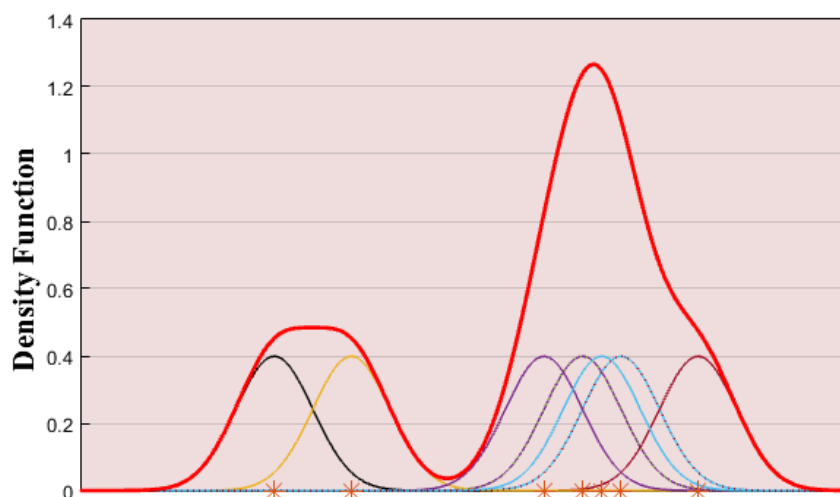


Figure (3.35): Kernel distribution function (Ribeiro, 2004)

3.3.4 Stage (IV): Character Classification

This is the heavy section of this chapter where we present the core of processing of BBACOR system.

After pre-processing, feature extraction and validation are completed, and since Naïve Bayesian classifier gives us the green sign to use the extracted features in the processes of classification via back propagation network, then we can now begin training and testing phases of our proposed AOCR system.

3.3.4.1 Back Propagation

Because of its high importance and naïve simplicity, we have used back propagation ANN in the handwritten Arabic character recognition through the features vectors for the characters and their desired labels. In order to be able to run the algorithm, we first present the neural network operation in general, and then we present the back propagation algorithm in particular. Figure (3.36) illustrates the methodology in which each neuron in MLP (multi-layer perceptron) processes the coming information.

In the first phase, a neuron of the j^{th} layer receives the activation (stimuli) from neurons of the $(j - 1)^{\text{th}}$ layer, namely, the x vector: $[x_1^{l-1}, x_2^{l-1}, x_3^{l-1}, \dots, x_i^{l-1}]$.

Then, each input component of the x vector is first multiplied by the corresponding weight parameter.

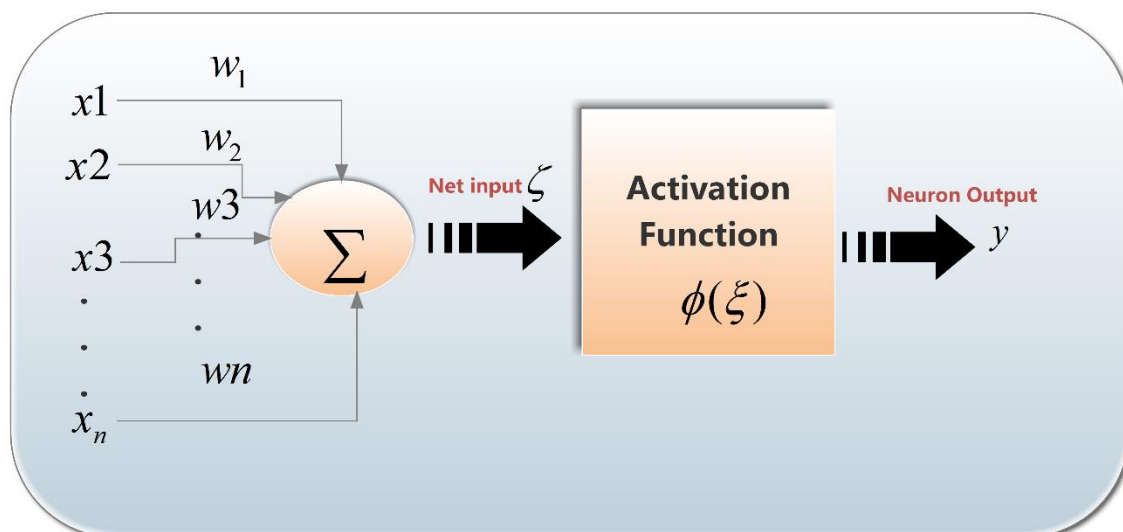


Figure (3.36): Information processing by the i^{th} neuron of the j^{th} layer

Then, they added together to produce the ‘net input’ or a weighted sum (ζ). This net input is passed through a neuron transfer (activation) function ($\phi(\cdot)$) to attain the final output of the neuron. This neuron output y_i^l can, in turn, become an activation for neurons in the next layer.

3.3.4.1.1 Transfer Function

The transfer function of neurons ($\sigma(\cdot)$) can be either, linear or nonlinear function depending of the type of neural network and the nature of problem.

The most commonly used transfer function is the sigmoid function, which we have used in our research, and it is given by the following mathematical formula in (3.20) (Zhang, 2000).

$$\sigma(\gamma) = \frac{1}{(1 + e^{-\gamma})} \quad (3.20)$$

$$\sigma(\gamma) \rightarrow \begin{cases} 1 & \text{as } \gamma \rightarrow +\infty \\ 0 & \text{as } \gamma \rightarrow -\infty \end{cases}$$

where, γ is the net input of neuron.

As shown in Figure (3.37), the sigmoid function has a smooth switching property exemplified by:

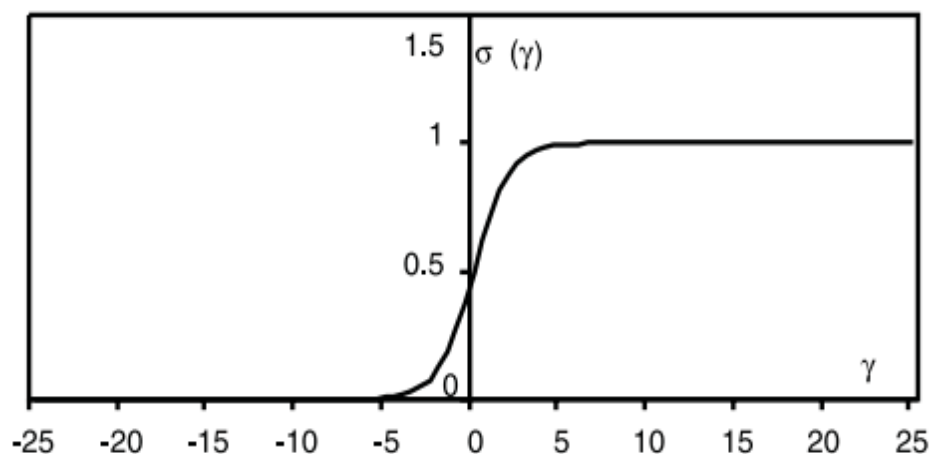


Figure (3.37): Sigmoid function (Zhang, 2000)

3.3.4.2 Back Propagation Training Algorithm

Back propagation can be viewed as a gradient descent technique used to train the weights in a multi-layer perceptron until it reaches to the minimum Mean Square Error (MSE) between the actual output of multi-layer perceptron and the target (desired output).

For a given problem, let us define a set of training vector, where for every vector component $\mathbf{x} \in \mathbf{X}$, there is a corresponding output vector $\mathbf{d} \in \mathbf{D}$, where \mathbf{D} is the set of desired (target) outputs corresponding to the training vectors in \mathbf{X} .

Let E_i , the instantaneous error of the i^{th} desired output vector, defined as (Ruck, et al., 1992).

$$E_i = \frac{1}{2} (\mathbf{d}_i - \mathbf{y}_i)^T (\mathbf{d}_i - \mathbf{y}_i) \quad (3.21)$$

$$E_i = \sum_{k=1}^N (\mathbf{d}_{k,i} - \mathbf{y}_{k,i})^2 \quad (3.22)$$

where, $\mathbf{d}_{k,i}$ the k^{th} component of the i^{th} desired (target) output vector \mathbf{d}_i , and $\mathbf{y}_{k,i}$ is the k^{th} component of the actual output vector \mathbf{y}_k when the i^{th} training instance \mathbf{x}_i is used as an input to the multilayer perceptron.

The total error E_T can be defined as in (3.23) (Ruck, et al., 1992).

$$E_T = \sum_{i=1}^I E_i \quad (3.23)$$

where I is the number of elements in \mathbf{X} and E_T is a function of training set and the weights in the neurons.

Now, the back propagation learning rule is defined as in (3.24).

$$\mathbf{w}_{ij}^{(t+1)} = \mathbf{w}_{ij}^{(t)} - \epsilon \frac{\partial E}{\partial \mathbf{w}_{ij}}(\mathbf{t}) \quad (3.24)$$

where ϵ is the learning rate, which considered the scaling factor of the error gradient, (\mathbf{t}) is the current iteration and $\frac{\partial E}{\partial \mathbf{w}_{ij}}(\mathbf{t})$ is the partial derivative of the error function with respect to the corresponding weight.

Figure (3.39), lists the major steps of the back-propagation algorithm implementation that was illustrated in the well-known paper of (Lippmann, 1987).

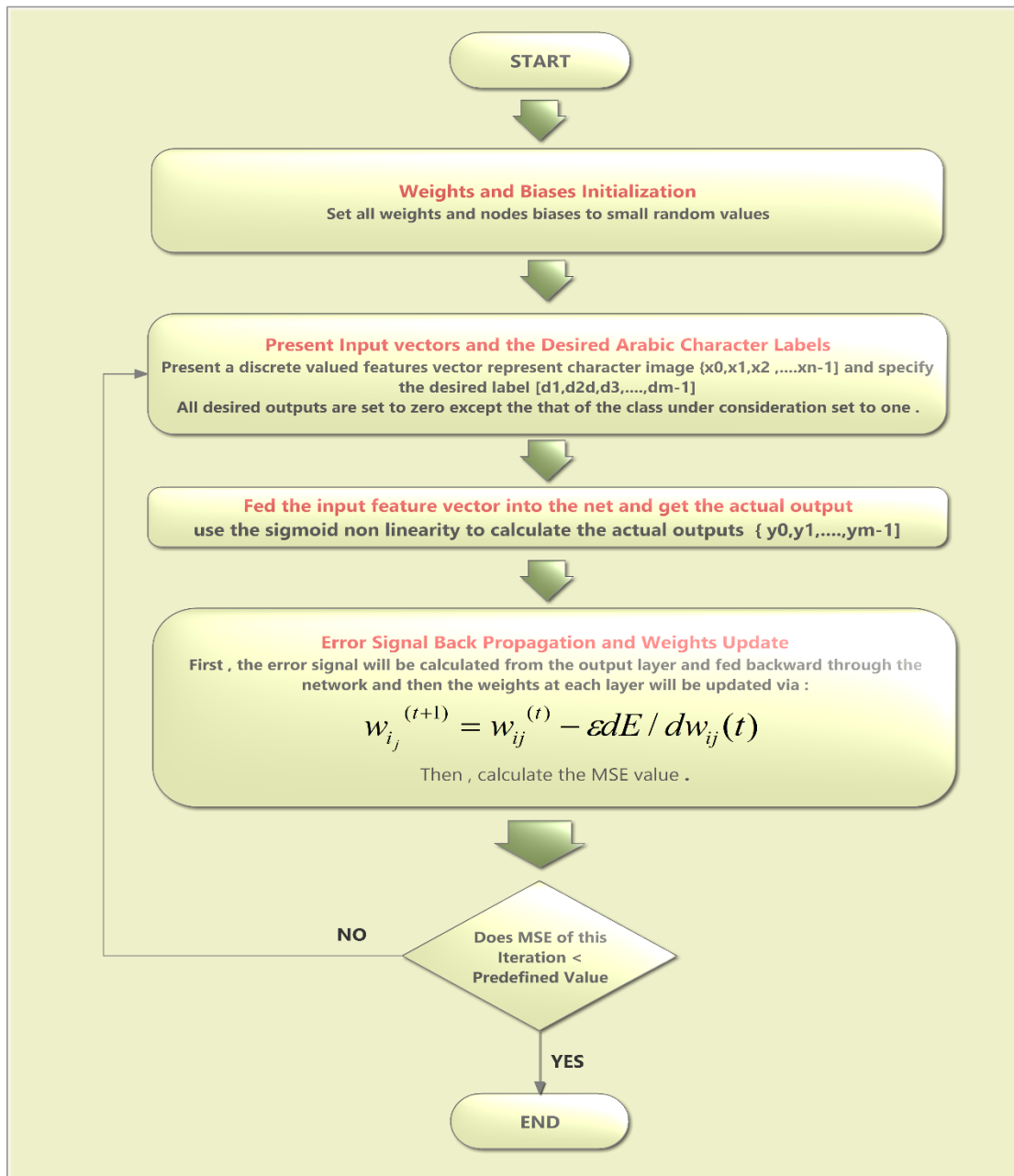


Figure (3.38): The Back Propagation Training Algorithm

3.3.4.3 Proposed BBAOCR System Training Phase

In this section, we will train the back propagation neural network algorithm that elaborated in previous section, which in addition to feature extraction techniques consists the heart of our proposed BBAOCR system. In this section, we will use the back propagation algorithm to build our BBAOCR system.

The training phase consists of two major phases: weights setting and Parameter tuning. Firstly, the back propagation ANN must be trained via the training dataset until it reach our optimal pre-specified MSE, which in turn leads to optimal weights.

In the parameter-tuning phase, the number of iterations, learning rate, and momentum value, initial weights, number of layers, number of neurons (nodes) at each layer parameters are set through tuning process, which lead to final network structure.

Then the optimal weights that obtained in training phase will be used in the testing phase of Back propagation network. The following steps summarize the training algorithm:

1. Load Training Database of features vectors.
2. Run Back propagation algorithm training as explained in the following steps: (Lippmann, 1987).

Step 1: Initialization phase:

In this phase, the weight w_{jk} for the k^{th} node and j^{th} layer will be initialized to random values between -0.5 and 0.5 for hidden layers and random values between -0.3333 and 0.333 for output layer .As explained in our MATLAB code:

```
for i=1:layersLength-2
weights{i} = [0.025 - 0.5.*rand(layers(i+1),layers(i)+1); zeros(1,layers(i)+1)];
end
weights{end}= 0.0333 - 0.333.*rand(layers(end),layers(end-1)+1);
```

Step 2: Present Input and Desired Output.

Present an input vector x_0 where $x_0 = [x_1, x_2, \dots, x_{N-1}]$ and specify the desired output $d_0 = [d_1, d_2, \dots, d_{N-1}]$

Step3: Calculate Actual Outputs

Run the selected input/desired output pattern through the network and evaluate the summation output y_{jk} (net input) for each layer from the hidden layer j through L , see Equation (3.25).

$$y_{jk} = \sum_{i=0}^N (x_{j-1}, w_{jk}) \quad (3.25)$$

Where N : is the number of node inputs except the offsets (bias).

The output of activation function is as in Equation (3.26).

$$x_{jk} = f(y_{jk}) = \frac{1}{1 + \exp(-ay_{jk})} \quad (3.26)$$

where a : slope of sigmoid function and in our case $a = 1$.

Step 4: Back-propagate the error signals through the network.

First, calculate the error signal for the output layer L in the Equation (3.27).

$$e_L = f'(y)(O - x) \quad (3.27)$$

where O For output layer L .

The error signal for the interior hidden layer (j) is as in (3.28).

$$\delta_j = f'(y_j) \sum_j (e_{j+1} * w_{j+1}) \quad (3.28)$$

where $f'(y_j)$ is the first derivative of the activation function $f(y_j)$. Then you get (3.29).

$$\frac{\partial E}{\partial w} (j-1) = - \sum_j^k \delta_{j-1} * x_{j-1} \quad (3.29)$$

where $\frac{\partial E}{\partial w} (j-1)$ represents the gradient of error function E with respect to weight w .

Step 5: Weights Update

In this step, the weights will be updated via (3.30).

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} - \epsilon \frac{\partial E}{\partial w_{ij}}(t) \quad (3.30)$$

Step 6: Checking Mean Square Error

check if MSE reach its desired value, if ‘YES’ then quit the algorithm and if ‘No’, then return back to step 2 of this algorithm.

3.3.4.4 Proposed BBAOCR Testing Phase

After training phase yields the optimal weights and optimal network structure through weights and parameters tuning, the back propagation neural network algorithm will be used to test our proposed BBAOCR system with its optimal weights and network structure. The testing phase consist essentially of the following steps:

1. Load Testing Dataset (Dataset of features vectors).
2. Pass the optimal weights to the trained back propagation ANN and run the test feature vectors (that corresponding to character images) through it.
3. Finally, run *Accuracy Function* to evaluate the recognition accuracy performance of BBAOCR system.

Chapter Four

Experimental Results

4.1 Experimental Setup

This chapter presents discussion of experimental results obtained from testing the proposed Off-line Arabic handwritten isolated character recognition system mentioned in chapter three. Our system is implemented using MATLAB 2015a as a powerful integrated development environment and windows 8 platform with Intel Core i3 2Due CPU 2.10 GHz with RAM 4.0GB.

4.2 Training and Testing Data Sets Statistics

As best elaborated in Chapter three, the proposed model is tested using (Center for Pattern Recognition and Machine Intelligence) CENPARMI well known dataset (Alamri , Sadri, Suen & Nobile, 2008) where it composed of two basic groups; one for training phase and the other is for testing phase. The neural network is trained using the CENPRMI labelled dataset until we reach the desired MSE, then the neural network tested using the test dataset.

The original dataset was divided further into three-sub datasets: training, testing and validation datasets. In our research, we use the training set that composed of (198) images for each letter and then merged the testing and validation datasets into one dataset: testing dataset that composed of (134) images for each letter. Figure (4.1) shows the statistical of both training and testing dataset.

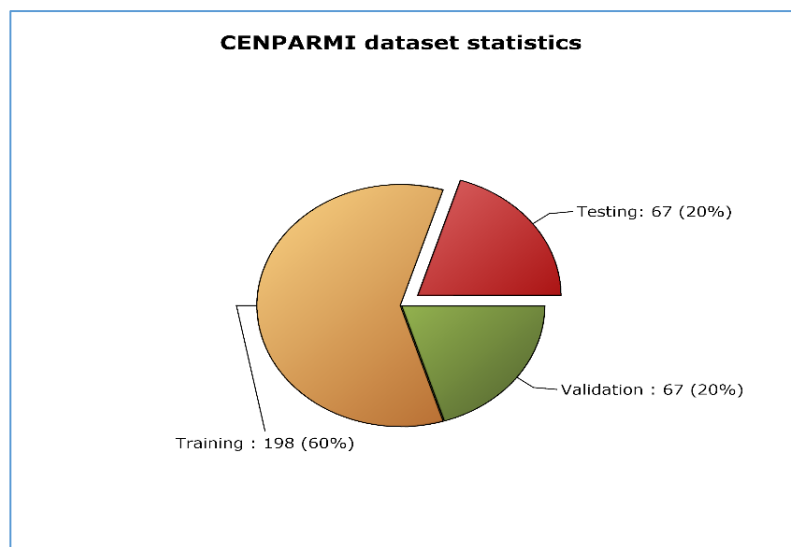


Figure (4.1): The Statistics of Training, Testing and Validation

Our dataset consists of Arabic isolated characters as illustrated in Figure (4.2),

أ	بـ	تـ	ثـ	م	نـ	ح
د	ذ	ر	ز	س	عـ	طـ
ضـ	ظـ	عـ	غـ	فـ	قـ	
كـ	لـ	مـ	نـ	هـ	وـ	يـ
			ةـ	وـ	اـ	ةـ

Figure (4.2): Offline Handwritten Arabic Isolated Letters

4.3 Experimental Results of Feature Validation Engine: Bayesian Classifier

As mentioned in chapter three, Naïve Bayesian Classifier has been used to validate the features that have been extracted using directional and regional features techniques. Naïve Bayesian classifier has been built using two probability density functions in sake of performance enhancement: Gaussian and Kernel distributions.

At first, we have built the Naïve Bayesian classifier based on Gaussian distribution ,however it yields very weak classification accuracy reached down to (40)%.

This lead us to use different distribution function, namely, Kernel distribution where it considered a natural response since this distribution used when the probability density function of specific dataset is unknown.

In our research, we try first to investigate the nature of the distribution of the features vectors that were extracted using two approaches and then concatenated in one vector. Naïve Bayesian classifier with kernel distribution was able to classify Arabic characters in accuracy reached up to (70 %). Which means that two third of the dataset was classified and recognized correctly, that is a powerful indicator that the features dataset that was used in the recognition phase via back propagation artificial neural network is valid.

4.4 Experimental Results of Classification Engine: Back Propagation

Multi-layer back propagation ANN is considered the classification engine of our proposed BBAOCR system, which consists of two major dependent phases: *Training* and *Testing* phase. As was illustrated in the chapter three, In the *Training phase*, the optimal weights, network parameters and network structure that will be fed as inputs to *Testing* phase.

4.4.1 Experimental Results of the BPANN: Training Phase

In the training stage of our Back propagation multilayer neural network handwritten recognition system, we have achieved a MSE of (**0.0205**) after (**1700**) epochs with overall recognition rate of (**91.32%**).

The topology and the parameters of the Back propagation neural network are listed in Table (4.1).

Table (4.1): The Parameters and Topology of Back Propagation ANN: Training Phase

Parameters	Value ('Description')
Layers	[140 70 42 5]
Performance	MSE
Transfer function	$f(y_k) = \frac{1}{1+exp(-ay_k)}$, where a = 1
No. of epochs	1700

The number of hidden neurons of the Back propagation neural network have been chosen to be high enough to model our problem at hand but at the same time not too high to avoid over-fitting.

Therefore, the hidden layers and the neurons (nodes) of each one has been chosen to have optimal performance in the testing phase. Although we have achieved an overall recognition rate reached up to (91.32%), هـ (Ha) letter is recognized in (100%). Most of Arabic characters are recognized between (80% to 95%) recognition accuracy. The recognition rates of characters in the Training phase are presented in Table (4.2).

Table (4.2): Achieved Recognition Rates for Arabic Characters in Training Phase

Character	Recognition Percentage	Character	Recognition Percentage	Character	Recognition Percentage
أ (Alif)	96.28%	ش (Sheen)	98.56%	ن (Noon)	98.58%
ب (Baa)	94.80%	ص (Saad)	94.39%	هـ (Ha)	100%
ت (Taa)	89.36%	ض (Daad)	88.10%	هـ (Ha2)	89.96%
ث (Thaa)	95.04%	ط (Ta)	95.80%	و (Waaw)	87.06%
ج (Jeem)	93.86%	ظ (Tha)	97.16%	و (Waaw2)	86.71%
ح (Haa)	93.68%	ع (Ayn)	82.90%	ء (Hamza)	89.38%
خ (Kha)	89.32%	غ (Ghayn)	76.21%	ي (Yaa)	93.17%
د (Daal)	85.13%	ف (Faa)	86.62%	ي (Yaa2)	88.89%
ذ (Thaal)	93.36%	ق (Gaaf)	88.85%		
ر (Raa)	85.77%	ك (Kaaf)	89.32%		
ز (Zaay)	90.28%	ل (Laam)	99.28%		
س (Seen)	87.41	م (Meem)	96.49%		

As shown in Table (4.2), the character Ha (هـ) has the highest recognition rate whereas the character Ghayn (غ) has the lowest one. However (15) characters have recognition accuracy above (90%).

Selecting proper parameters in the process of back propagation ANN building is principal effect and has direct impact on the output results in the Testing phase. Which includes number of layers and number of neurons per each layer. This process of tuning end up with the optimal back propagation ANN structure that leads to the optimal weights as output of training phase.

During the Training phase, we have tried to attain the optimal performance through initial weights, weights, and network parameters and structure tuning. Figure (4.3) illustrates the epochs tuning during Training phase.

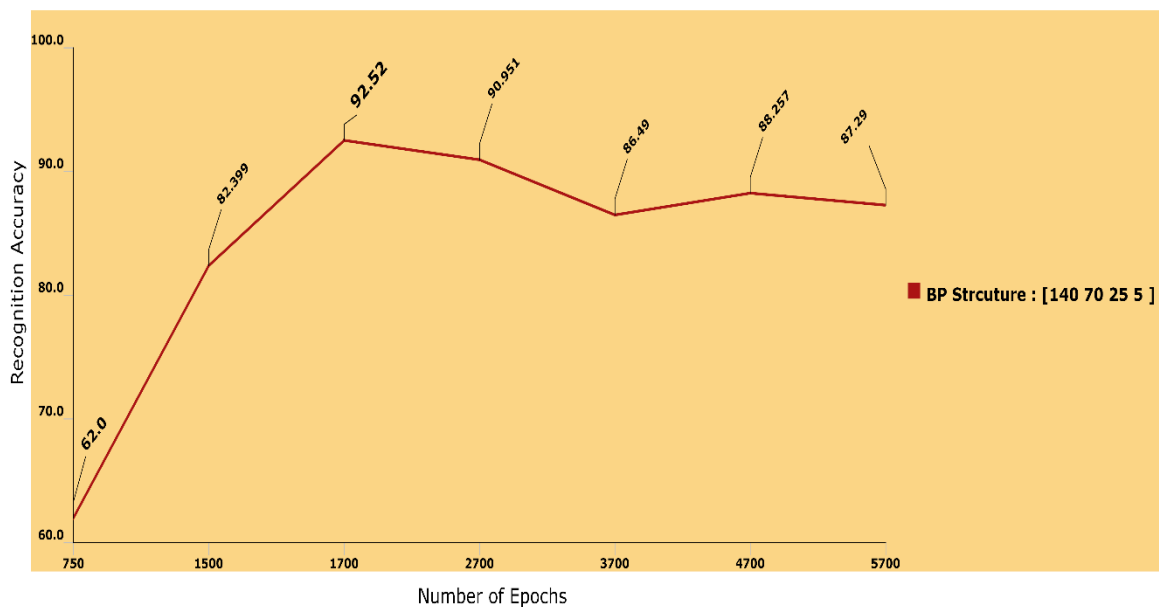


Figure (4.3): Epochs Tuning

As shown in Figure (4.3), the increasing in the number of epochs used to train the BPANN does not yield further enhancement in the network performance , which is one of the principal facts that is well known about BP ANN in particular and ANNs in general. The best performance represented in least MSE is achieved after (1700) iterations.

In the same procedure, we have tuned the structure of the back propagation ANN in order to achieve optimal performance as shown in Figure (4.4)

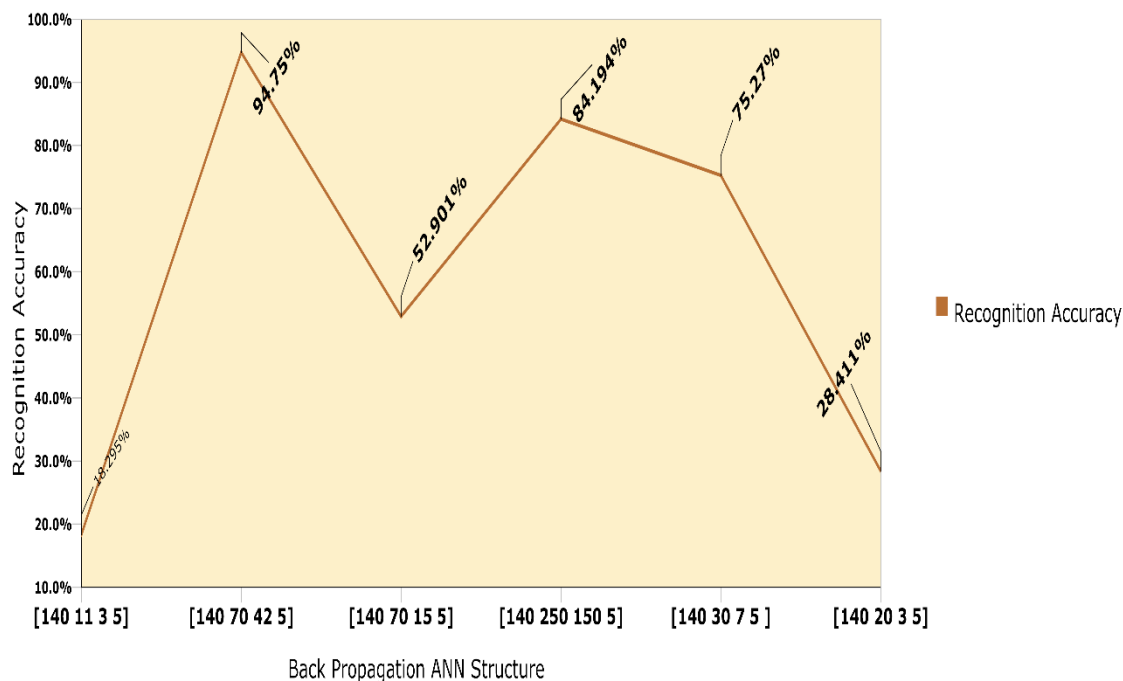


Figure (4.4): Back Propagation ANN Structure (Number of neurons /Layer) Tuning

Figure (4.4) illustrates the effect of Back propagation ANN structure tuning on the overall recognition accuracy of our proposed OCR system, where we used four layers: Input, two hidden and one output layers, where the first hidden consists of (70) neurons whereas the other hidden layer consists of (42) neurons and (5) neurons at the output layer. This structure yields the optimal weights, which leads to the best classification performance at testing phase reaches up to **(94.75 %)**.

4.4.2 Experimental Results of the BPANN: Testing Phase

After training phase has been performed and the optimal weights of Back Propagation Artificial Neural Network have been obtained. Then these optimal weights are fed into testing phase of our proposed BBAOCR system. Our proposed BBAOCR has achieved low MSE of (0.0126) associated with overall recognition accuracy reached up to (94.75%) and (100 %) for many characters. The recognition rates achieved by our proposed BBAOCR system for each character are presented in Table (4.3).

Table (4.3): Achieved Recognition Rates for Arabic Characters in Testing Phase

Character	Recognition Percentage	Character	Recognition Percentage	Character	Recognition Percentage
أ (Alif)	100%	ش (Sheen)	100 %	ن (Noon)	100%
ب (Baa)	96.72 %	ص (Saad)	100%	هـ (Ha)	100%
ت (Taa)	90.32 %	ض (Daad)	86.89%	هـ (Ha2)	95.08%
ث (Thaa)	96.77 %	ط (Ta)	96.88%	و (Waaw)	89.06%
ج (Jeem)	98.46 %	ظ (Tha)	100%	و (Waaw2)	96.88%
ح (Haa)	95.08 %	ع (Ayn)	93.44%	ء (Hamza)	92.06%
خ (Kha)	96.83 %	غ (Ghayn)	73.77%	ي (Yaa)	98.39%
د (Daal)	93.44 %	ف (Faa)	93.44%	ي (Yaa2)	96.77%
ذ (Thaal)	95.31 %	ق (Gaaf)	100%		
ر (Raa)	87.30 %	ك (Kaaf)	93.85%		
ز (Zaay)	90.32 %	ل (Laam)	100%		
س (Seen)	84.38%	م (Meem)	100%		

As shown in table (4.3), the achieved results are very promising where the achieved overall accuracy reached up to (94.75 %) for all letters although we get success (recognition) rate reached up to (100%) for some letters as highlighted in Table (4.3).

Figure (4.5) depicted a graphical comparison of Training and Testing phases of our proposed OCR system in terms of recognition accuracy, number of Arabic characters that are fully recognized (namely, with recognition rate of 100%), number of Arabic

characters that have recognition rates above (**90%**), and the number of Arabic characters that have recognition rates less than (**90%**).

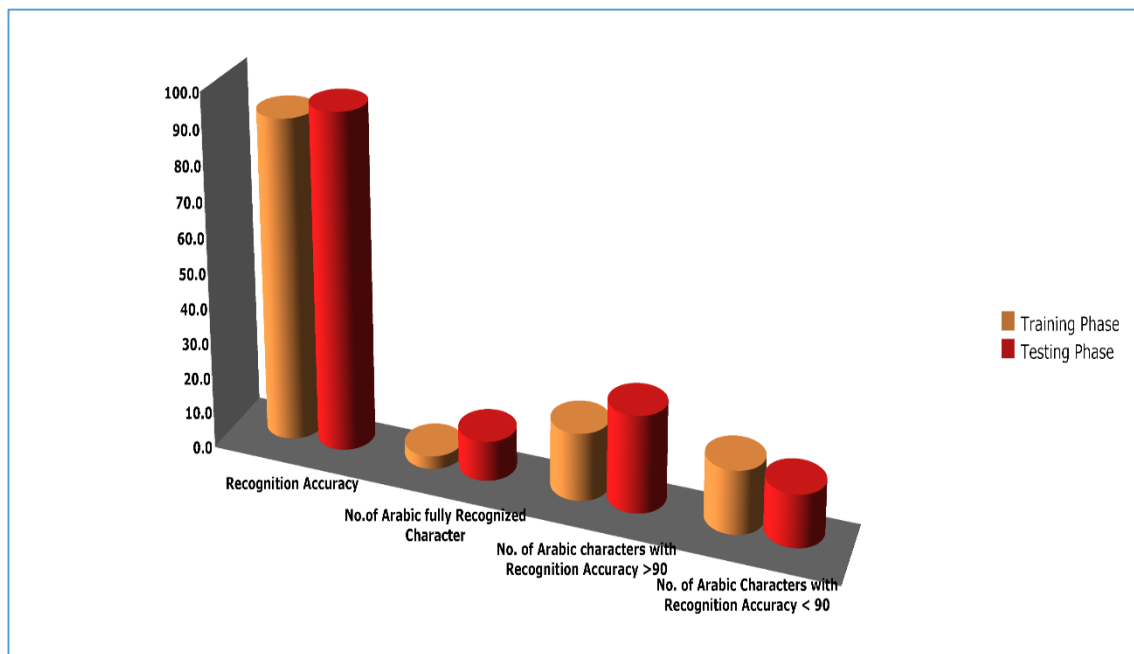


Figure (4.5): Recognition Accuracy Comparison: Training and Testing Phases of our

As noted in Figure (4.5), the accuracy at training phase is less than that achieved in testing phase of our OCR system, and at the same time the number of fully recognized Arabic characters, which indicate, that the weights that obtained in the training phase was optimal to be used in the testing phase.

In addition, the number of characters that are recognized with recognition accuracy above (**90%**) more in testing phase rather than that achieved in training phase, which proves the high stability of our back propagation neural network structure and the effectiveness of the feature extraction techniques that we used .

4.5 Comparisons of BBAOCR system to Sahlol's OCR System

In this research, the process of combination of two sets of features, namely, direct and regional features yielded high recognition accuracy of handwritten Arabic characters. On the other hand, the utilization of back propagation artificial neural network has a dramatic effect on the classification results, which prove the powerfulness of these types of artificial neural networks.

After examining the recognition accuracy for each character, we found that the recognition rate is between (100%) for the easiest characters as Saad(ص), Noon(ن), Meem(م), and (73.77%) and (84.38%) for Ghayn(غ), and Seen(س), respectively, where Ghayn(غ) is considered one of the hardest recognized characters.

However, we achieved (100%) and (93.44%) for Gaaf(ق) and Ayn(ع) respectively although these Arabic characters are categorized as the hardest recognized characters in Arabic Language.

In contrary to what achieved by (Shalol, et.al, 2014A), where they got an accuracy rates of (61%) and (66%) for Gaaf(ق) and Ayn(ع) respectively, which is considered far from the results that we achieved for the same characters.

(Shalol, et.al, 2014A) used the same dataset (CENPARMI) that we used during our research; however, we have achieved better results either in terms of overall recognition accuracy or in terms of accuracy rate for each character.

Figure (4.6) shows a comparison between the experimental results that achieved by our proposed BBAOCR system and that proposed by (Shalol, et.al, 2014A). Where we first establish a comparison in terms of **fully** recognized letters in our proposed system and its corresponding recognition rates by (Shalol, et.al, 2014A) OCR system.

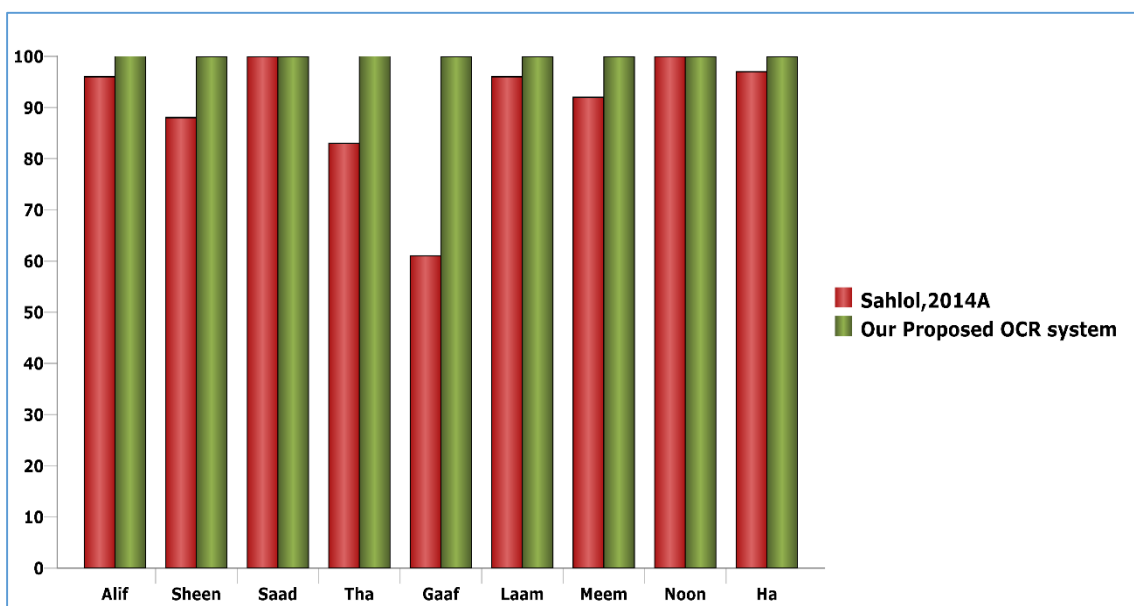


Figure (4.6): Accuracy Comparison of our proposed OCR system to proposed by (Sahlol, 2014A)

Where shows our fully recognized characters to that recognized by (Sahlol, 2014A)

Whereas Figure (4.7) shows the experimental results comparisons that achieved by our proposed OCR system and that proposed by (Shalol, et.al, 2014A) if the fully recognition is in side of (Shalol, et.al, 2014A).

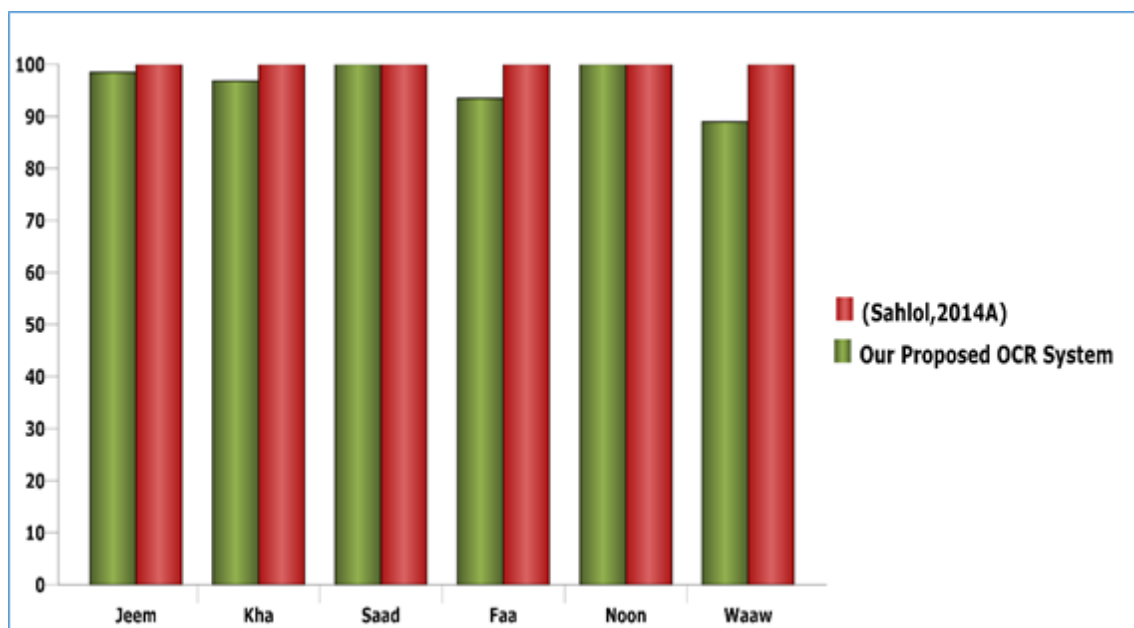


Figure (4.7): Accuracy Comparison of our proposed OCR system to proposed by (Sahlol, 2014A)

Where shows our achieved accuracy rates of characters that fully recognized by (Sahlol, 2014A)

As explained in Figures (4.6) and (4.7), we have achieved more **fully** recognized handwritten Arabic characters in compared to that achieved by (Shalol, et.al, 2014A) utilizing same dataset (CENPARMI).

Even in case of characters that fully recognized by (Shalol, et.al, 2014A), we achieved recognition accuracy above (**90%**).

Since both systems (our proposed OCR system and that proposed by Shalol, et.al, 2014A) were built on same dataset, our experimental results that have the superiority which has a strong indicator that our feature extraction techniques associated with back propagation ANN as a classification technique proved to be more efficient and robust than that proposed by (Shalol, et.al, 2014A). Tables (4.4) and (4.5) list the lowest ten recognized characters registered by our proposed system and that registered by (Shalol, et.al, 2014A) system respectively.

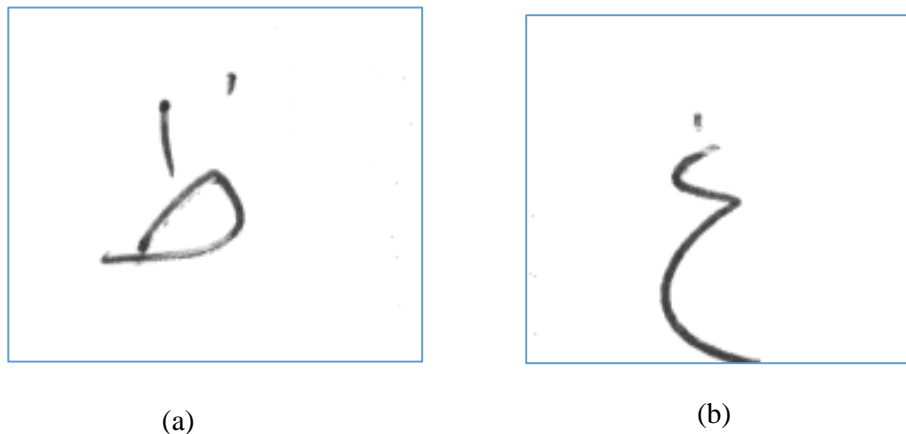
Table (4.4): Ten worst Characters recognized by (Shalol, et.al, 2014A)

Character	Recognition Accuracy
Gaaf(ق)	61%
Ayn(ع)	66%
Taa(ت)	69%
Daad(ض)	72%
Thaa(ث)	76%
Seen(س)	78%
Zaay(ز)	80%
Kaaf(ك)	81%
Daal (د)	83%
Tha(ظ)	83%

Table (4.5): Ten worst Characters recognized by our proposed

Character	Recognition Accuracy
Ghayn(غ)	73.77%
Seen(س)	84.38%
Daad(ض)	86.89%
Raa(ر)	87.30 %
Waaw(و)	89.06%
Taa(ت)	90.32 %
Zaay(ز)	90.32 %
Hamza(ء)	92.06%
Faa(ف)	93.44%
Daal(د)	93.44 %

The lowest recognition rate (worst cases) that registered by our proposed BBAOCR system is (73.77%) for **Ghayn(غ)** character, where **Ghayn(غ)** tend to be recognized as **Tha(ظ)** for ten times since there is some substantial similarities between the skeleton zones of both characters as illustrated in Figure (4.8) and Figure (4.9) respectively.

**Figure (4.8): (a) Tha (ظ) Character (b) غ (Ghayn) Character**

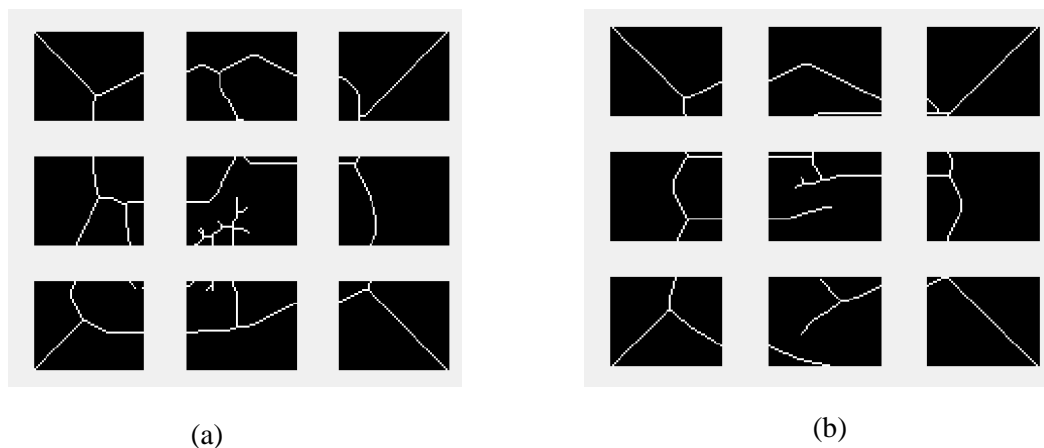


Figure (4.9): 3x3 zones of (a) Tha (ظ) Character (b) Ghayn (غ) Character

As noted in Figure (4.9), there is a notable similarity between the skeletons of both **Ghayn(غ)** and **Tha(ظ)** characters with a slight difference in zone 22, which explains the low recognition rate that registered for **Ghayn(غ)** character where ten of **Ghayn(غ)** characters were misrecognized as **Tha(ظ)** characters due to high skeleton-zones similarity.

One of the major reasons that lead to relatively low recognition rates for **Zaay(ز)** character is reside in CENPARMI database itself. Many of **Zaay(ز)** characters were written almost as **Thaal(ذ)** characters which hard to recognized even by a human experts as shown in Figure (4.10).

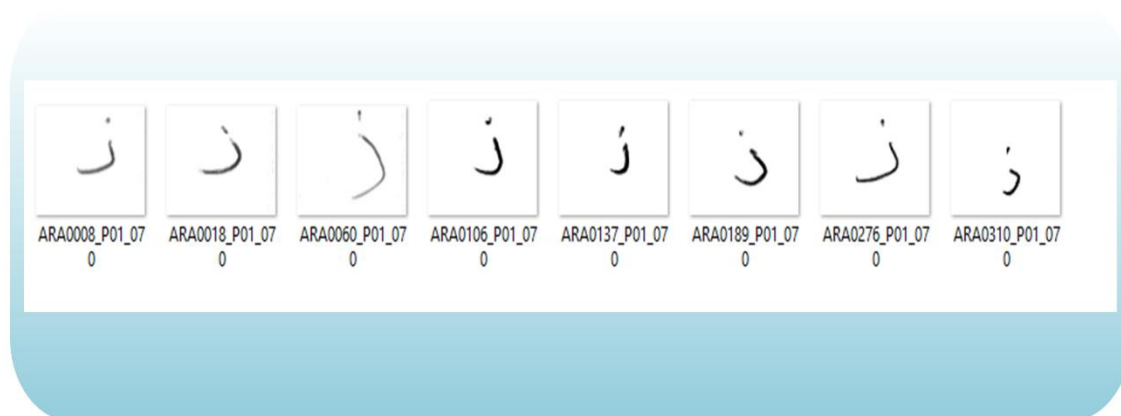


Figure (4.10): Zaay (ز) Characters written as Thaal (ذ) Characters

As illustrated in Figure (4.10), a snapshot of database is taken with the titles of the letters images which shows **Zaay** (ز) characters written exactly as **Thaal** (ث) character. This is can be considered one of the limitations of this research.

Moreover, either CENPARMI or other Arabic Characters databases that can be created currently or in the future, we cannot control the way that used by writers or the stroke of the pen during the process of character drawing especially in Arabic language which contains many characters that distinguished only by dot.

Another issue arises when some writers draw **Sheen**(ش) character in Al-Roqa'a writing style, this lead to misrecognition **Daad**(ض) character with **Sheen**(ش) character since the skeleton zones of **Sheen**(ش) character drawn in Al-Roqa'a writing style and **Daad**(ض) character almost have the same skeleton shape which explains the relatively low recognition rate of **Daad** (ض) character whereas **Sheen**(ش) character is fully recognized with (100%) recognition accuracy as illustrated in Figure (4.11).

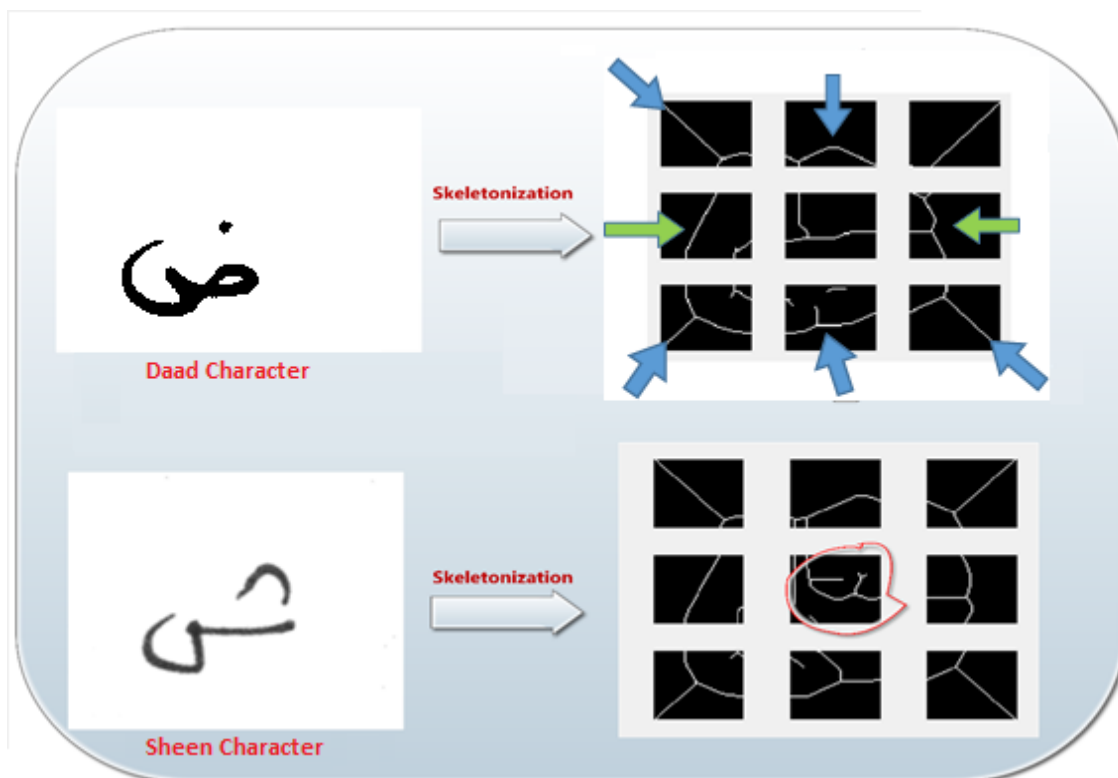


Figure (4.11): The skeleton zones of Daad (ض) and Sheen (ش) Characters where Sheen (ش) written in Al-Ruqa'a Style

As elaborated in Figure (4.11), **Daad**(ض) and **Sheen**(ش) character if it written in Al-Roqa'a style, they will almost have same segments types in the different zones except the centre zone(zone22), this explain the misrecognition of **Daad**(ض) character with **Sheen**(ش) character for more than six times. Figure (4.12) shows more examples that represent this case.

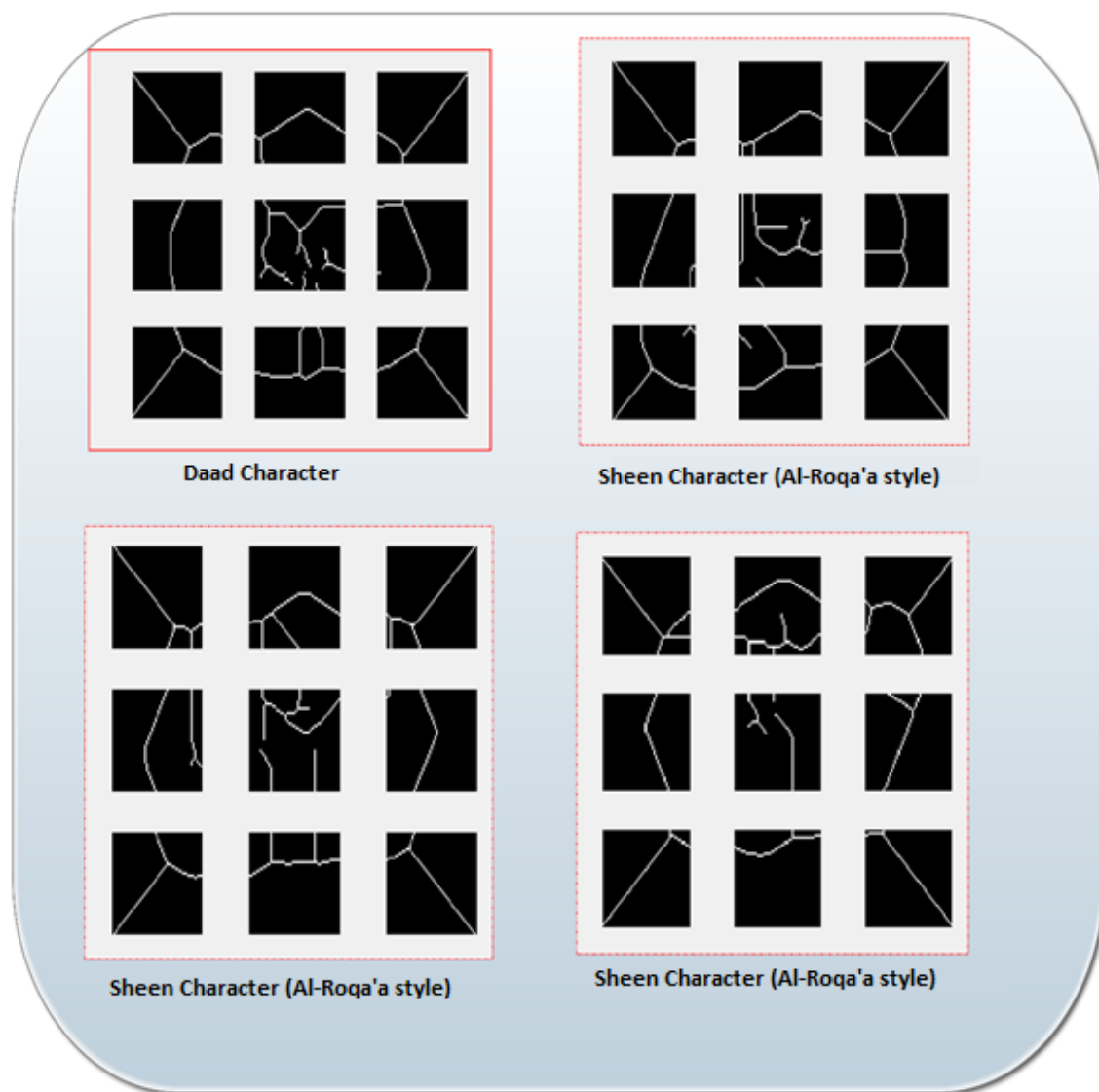


Figure (4.12): The skeleton zones of Daad (ض) and three Sheen (ش) Characters

On the other hand, in case of the OCR system that proposed by (Shalol, et.al, 2014A) the lowest recognition rate start with (61%) and end up with (83%). In case of (Shalol, et.al, 2014A) the ten letters that have the lowest recognition rates are those letters that have semi-looping or drawn with loops with some writing styles, where there is a high similarity with Arabic letters that have real looping. Moreover, the feature extraction techniques that proposed by (Shalol, et.al, 2014A) bear a strong weakness in recognizing the secondary components due to the statistical feature extraction techniques represented by the connected components identification in the Arabic letters. This feature extraction technique distinguish Arabic letters among those characters that do not contain connected components as ((Haa ح), Daal(د), Raa(ر), Seen(س), Saad(ص)) and others that contain connected components (Alif(أ), Kha(خ), Thaal(ث), Gaaf(ق), Yaa(ي)). However, many of similar Arabic characters have similar body but different connected components, so the knowledge of whether the letter have connected component or not is not adequate to recognized character.

Ignoring the classification technique that have been used in our thesis or that have been used by (Shalol, et.al, 2014A). The high recognition accuracies that have been registered by ten lowest recognized characters are considered an obvious indicator of the high efficiency of the feature extraction techniques that we have applied in our proposed BBAOCR system comparing to what used by (Shalol, et.al, 2014A).

4.6 Comparisons of BBAOCR system to other AOCR systems

This section is dedicated to compare our experimental results with other existing optical character recognition techniques and systems. Where we first compare the performance of the proposed BBAOCR system with other systems that built on the same data sets, namely, CENPARMI. Then we compare our experimental results with that achieved by other systems that used different dataset for Off-line Arabic handwritten isolated characters and built on different classification algorithms and techniques.

4.6.1 Comparison of BBAOCR System and other AOCR Systems Based on CENPRIM Dataset

(Shalol, et.al, 2014) has a pioneering work in the field of isolated Arabic character recognition represented by two works published in a couple of papers (Shalol, et.al, 2014A) and (Shalol, et.al, 2014B) where they built an AOCR systems based on CENPARMI database.

However, they have utilized two broadly different algorithms: one of these systems was built using back propagation ANN, whereas the other one was built using SVM and KNN algorithms separately.

A comparison of our achieved experimental results compared to that achieved by (Shalol, et.al, 2014A) and (Shalol, et.al, 2014B) in terms of recognition accuracy is elaborated graphically in Figure (4.13).

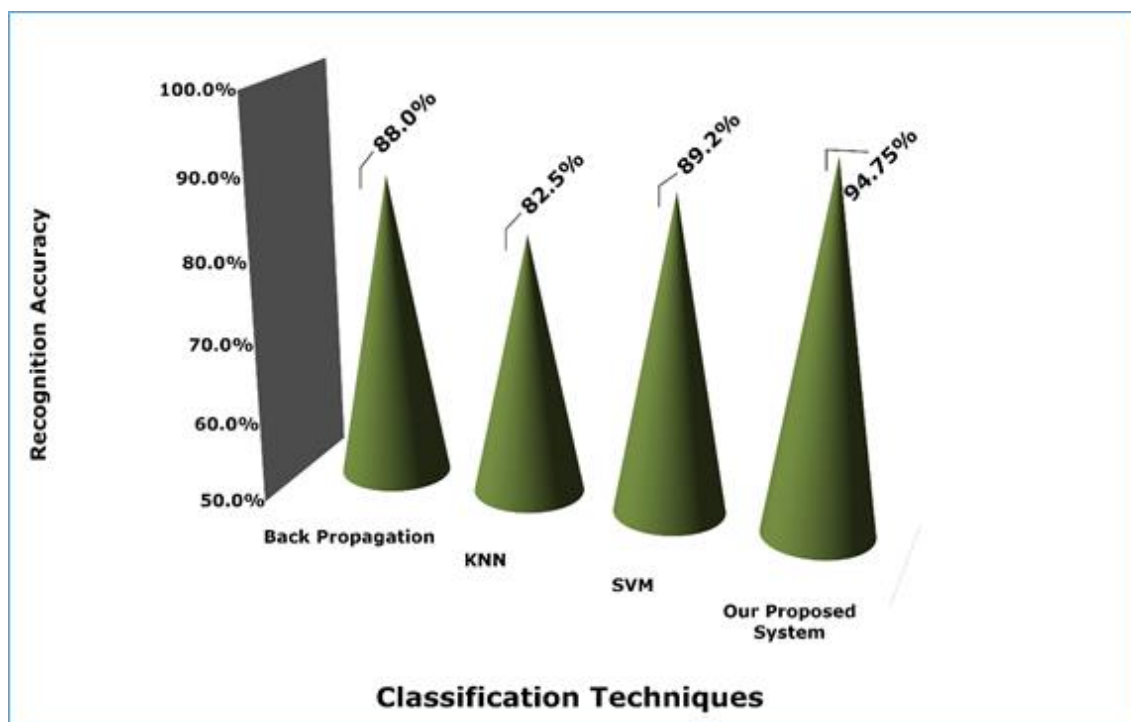
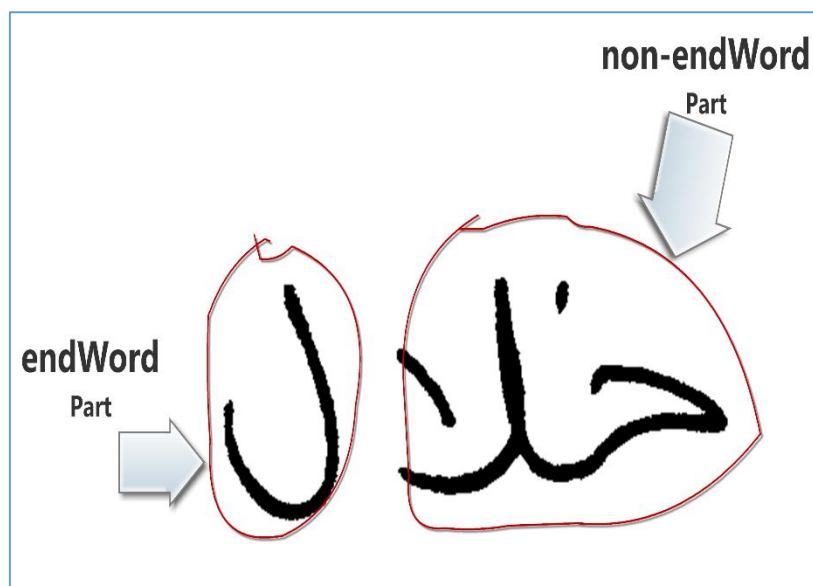


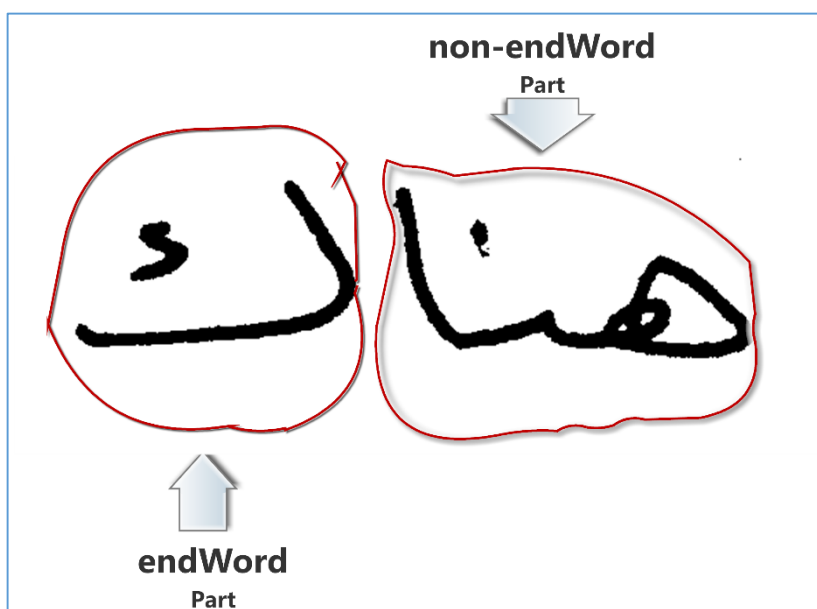
Figure (4.13): Recognition Accuracy Comparison: Our Proposed BBAOCR system to other AOCR systems.

As shown in Figure (4.13) our proposed system outperforms all techniques that proposed by (Shalol, et.al, 2014A) and (Shalol, et.al, 2014B) which are considered state-of-arts achieved recognition accuracy in the field of Off-line handwritten Arabic characters. This is due to the novel feature-extraction techniques that we have exploited in association of powerful data mining algorithm: back propagation ANN.

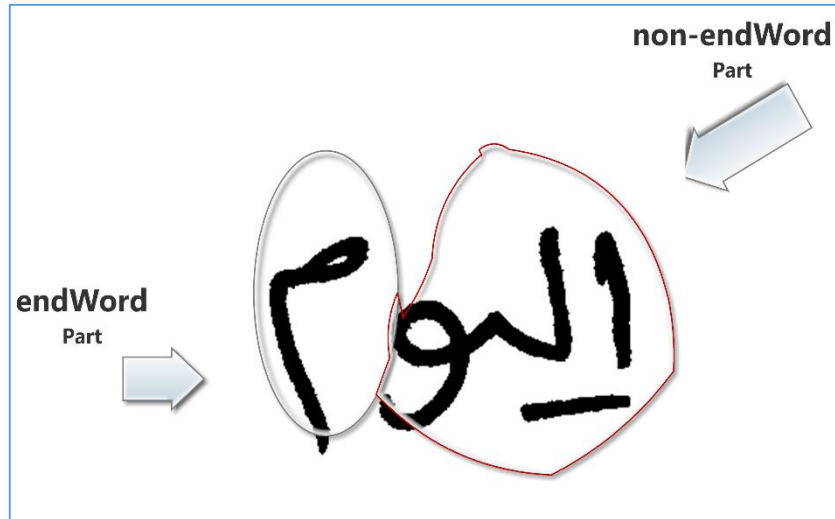
(Jamal, 2015) suggested a novel handwritten segmented Arabic word recognition algorithm as a research core for his PhD thesis. Up on this algorithm, the word consists of major classes: end-Word and non-end-Word, the algorithm actually can be viewed as two-class problem: one for the non-end-word recognition and the other one is end-word recognition. Figure (4.14) elaborated these concepts.



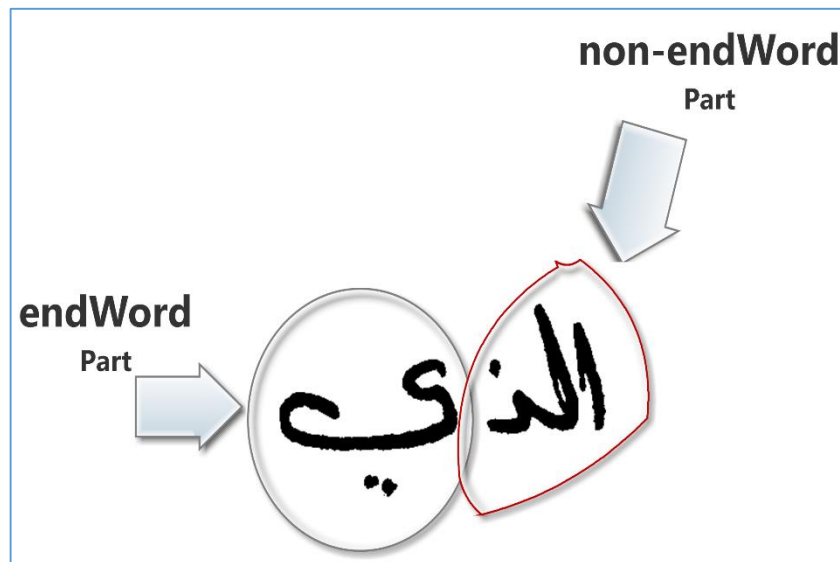
(a)



(b)



(c)



(d)

Figure (4.14): (a) the word Khelal (خلال) composed of two classes: non-end word Khel (خلا) and end-word Laam (ل).

(b) The word Hunak (هناك) composed of two classes: non-end word Huna (هنا) and end-word Kaaf (ك).

(c) The word Alyaom (اليوم) composed of two classes: non-end word Alyao (اليو) and end-word Maam (م).

(d) The word Allthi (الذي) composed of two classes: non-end word Allth (الذ) and end-word Yaa (ي).

Where we note that, the end-word recognition is equivalent to isolated letters recognition, which represents the core idea of our thesis.

Accordingly, (Jamal, 2015) had divided the experimental results into two main groups, the experimental results of the non-end-word recognition, and the other group is the experimental results that obtained for the end-word recognition.

He achieved promising recognition rate reached up to (90.88%). Figure (4.15) shows a graphical comparison of the experimental result that got by (Shalol, et.al, 2014A,B), (Jamal, 2015) and our proposed system.

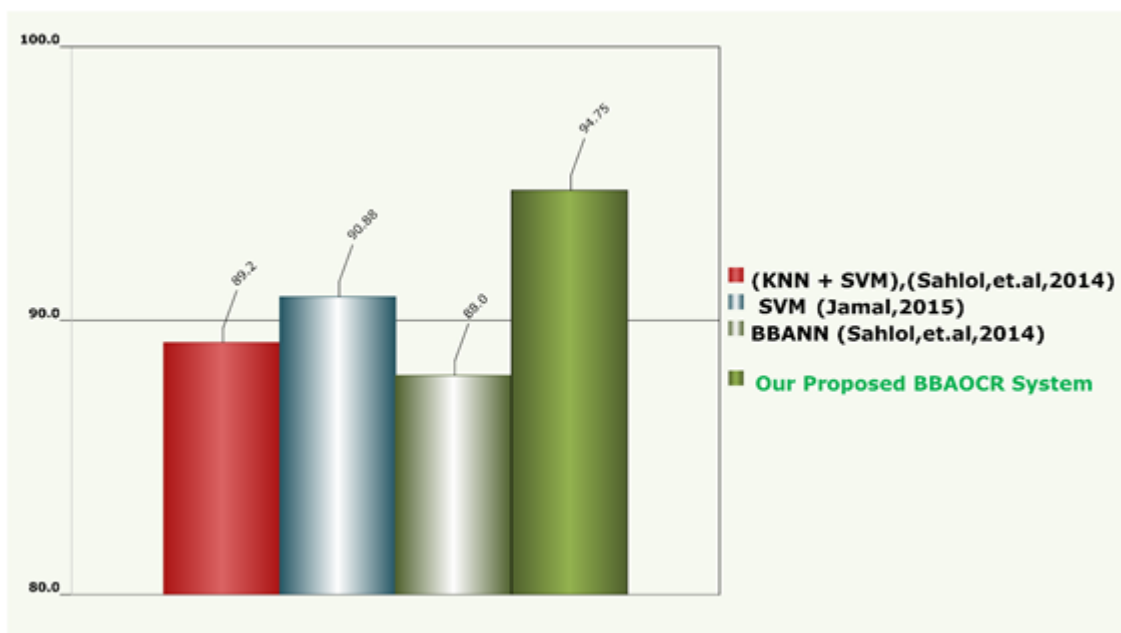


Figure (4.15): Graphical Comparison between (Jamal, 2015) and (Sahlolo, 2014) AOCR systems and our proposed BBAOCR system.

Even though (Jamal, 2015) has achieved high recognition performance, we have reached better performance where we have reached a recognition accuracy up to (94.75 %) which is considered a promising in comparison to existing AOCR schemes.

A deep insight in the experimental results discussed above will reveal that our proposed BBAOCR system can be efficiently utilized as a major part of the system that

proposed by (Jamal, 2015) to recognized end-word classes. This will enhance the overall handwritten word recognition accuracy of (Jamal, 2015) system.

4.6.2 Comparison of BBAOCR System and other AOCRs

Since many of available AOCR systems were built and implemented using different datasets other than CENPARMI and based on different techniques and algorithms other than BBANN, in this section we compare our achieved performance to other systems that built using different datasets and approaches.

Figure (4.16) illustrates a graphical comparison between our proposed BBAOCR system and other up-to-date AOCR systems that used different features extraction and classification techniques.

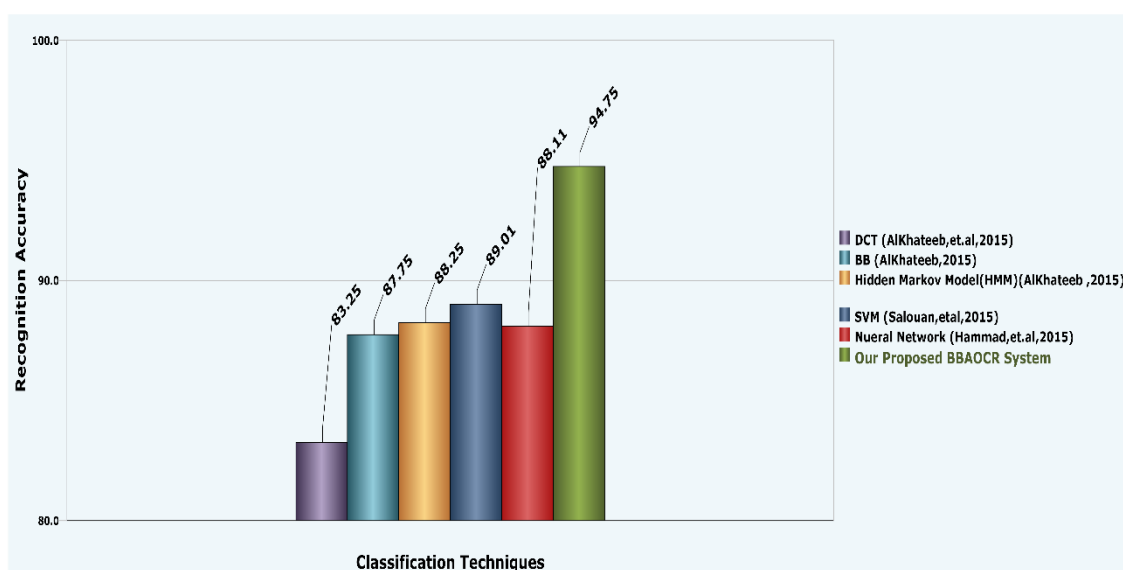


Figure (4.16): Graphical Comparison between other up-to-date AOCR systems and our Proposed BBAOCR

As shown in Figure (4.16), our proposed system shows its superiority in compared with other OCR systems due to high effective feature extraction and validation techniques that have been employed in our proposed system.

On the other hand, we also outperformed other optical character recognition techniques and systems proposed in last years.

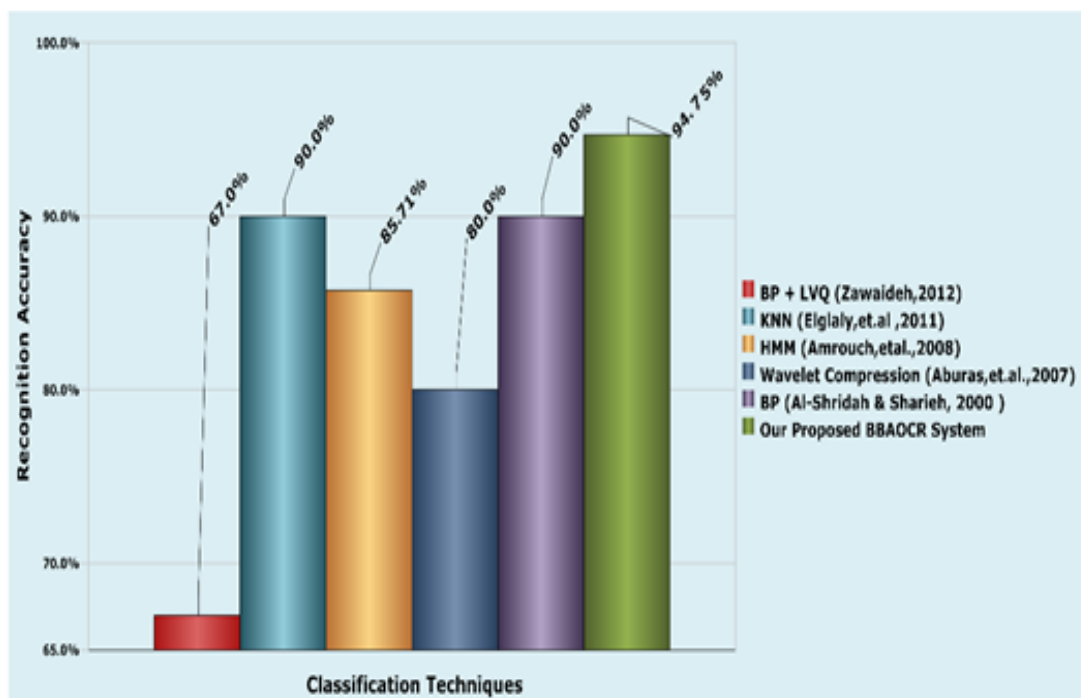


Figure (4.17): Graphical Comparison between other AOCR systems

However, a few efficient OCR systems proposed, implemented and has achieved high recognition performance comparable to that has achieved by our proposed BBAOCR system as best illustrated in Figure (4.18).

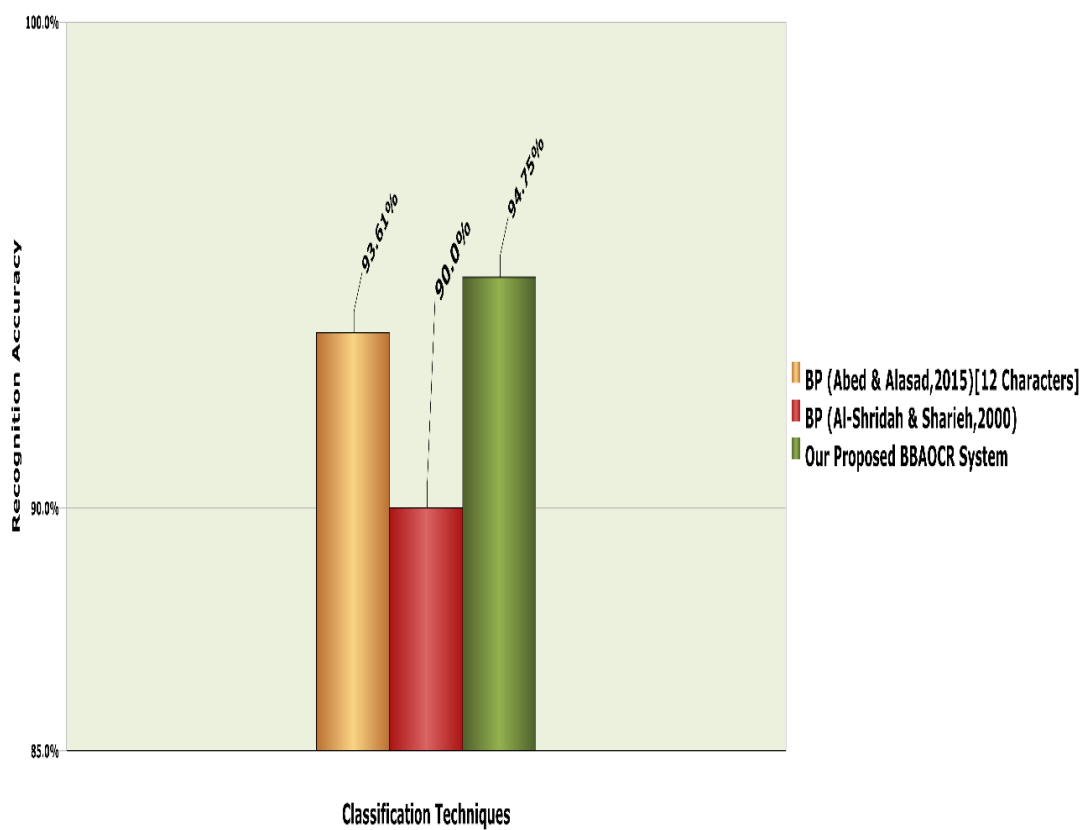


Figure (4.18): Graphical Comparison of BBAOCR system to High-performance AOCR Systems

Chapter Five

Conclusions and Future work

This major purpose of this thesis is to design and implement a novel Arabic handwritten character recognition system that can overcome the low recognition accuracy problem of currently available Arabic handwritten optical isolated character recognition systems and to enhance the accuracy recognition of the existing handwritten Arabic text recognition systems.

5.1 Conclusions

- An Off-line Arabic Handwritten isolated recognition system based on novel feature extraction techniques and powerful machine learning algorithm: Back Propagation artificial neural network is proposed in this thesis. This research proves that the proposed system BBAOCR provide better recognition accuracy rather than that achieved by other approaches.
- The proposed system BBAOCR achieved **(94.75%)** overall recognition accuracy and a recognition rate of **(100%)** for **(Alif(أ), Sheen(ش), Saad(ص), Tha(ظ), Gaaf(ق), Laam(ل), Meem(م), Noon(ن), Ha(هـ))** which is considered a promising results in the field of Arabic handwritten recognition.
- The character image zoning into horizontal, vertical and 3x3 square zones enabled us to exploit the high capabilities of the feature extraction techniques in optimal way where the zoning techniques cove the tinny details of the character curve.

- Even though the novel Features extraction techniques that have been used in this thesis are originally designed to be used for English letters, it was considered one of the major reasons of our high performance that we have achieved in this thesis which proved by two tools: Naïve Bayesian and Back Propagation ANN classifiers.
- The feature validation using Bayesian Network provide us with a powerful indicator for the applicability and efficiency of the novel features that have been used in our thesis.
- The Back Propagation Neural network proves its high recognition capability, which can be considered the core engine of our proposed AOCR.
- Despite of the computational complexity of this system, it is suitable for real time applications because the run time is acceptable where (0.366214) seconds, which means (0.01144) second for each character

5.2 Recommendations and Future Work

In aim of concluding our thesis, the researcher can recommend the following ideas:

- It is an open project covering a broad range of subjects and there are a lot of extended possibilities, namely, the researcher can use any of data mining or machine learning techniques in association with Back propagation in a hybrid manner in order to enhance the overall recognition accuracy of the proposed AOCR system where the system will be: hybrid AOCR system.
- We used the Z-score normalization techniques during our AOCR system implementation, however for further investigation, we recommend using the other type of normalization, such as min-max normalization or decimal normalization technique.
- Further theoretical analysis needed to find further optimality in choosing the number of layers, and the number of neurons per layer to get better optimal performance for the back propagation artificial neural network.
- Since the handwritten Indian and Arabic digits are much easier than the isolated handwritten Arabic characters, we recommend using our proposed BBAOCR system to recognize it and we expect higher performance rather than that achieved in case of isolated Arabic characters. In addition, we can use our proposed BBAOCR system for isolated printed Arabic character recognition and we expect higher recognition accuracy than that we have achieved in case of handwritten Arabic characters due to its smoothest forms of printed characters.

- Depending on the high performance that has been achieved in this thesis. We recommend using another classifier in association with the novel feature extraction techniques that have been used in this thesis, such as SVM, KNN, Fuzzy Logic, and Hidden Markove Models (HMM), Self-Organized Map (SOM) network or other types of artificial neural networks.

References

A. Rosenfeld and A. C. Kak (1982). Digital Picture Processing, *Academic Press Inc. New York*.

Abed, Alasad (2015). High Accuracy Arabic Handwritten Characters Recognition Using Error Back Propagation Artificial Neural Networks. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 6(2), 145-152.

Abed, A. P. M. A. Improving Handwritten Isolated Arabic characters Recognition with Particle Swarm Optimization Algorithm. *Diyala Journal for pure sciences*, 10(2), 18-34.

Aburas, A. A., & Rehiel, S. A. (2008). JPEG for Arabic Handwritten Character Recognition: *Add a Dimension of Application. INTECH Open Access Publisher*.

Aburas, A.A. and Rehiel, S. M. A (2007). Off-line Omni-style Handwriting Arabic Character Recognition System Based on Wavelet Compression, *ARISER*, 3(4), 123-135.

Amrouch, M., Elyassa, M., Rachidi, A., & Mammass, D. (2008). Off-line arabic handwritten characters recognition based on a hidden markov models. *Springer Berlin Heidelberg In Image and Signal*, 447-454.

Abuzaraida, M. A., Zeki, A. M & Zeki, A. M. (2013). Recognition Techniques for Online Arabic Handwriting Recognition Systems. *IEEE .In Advanced Computer Science Applications and Technologies (ACSAT), International Conference on* 518-523.

Alamri, H., Sadri, J., Suen, C. Y., & Nobile, N. (2008). A novel comprehensive database for Arabic off-line handwriting recognition. *In Proceedings of 11th International Conference on Frontiers in Handwriting Recognition, ICFHR*, 8, 664-669.

Albashiti, A. H. & Tamimi, H. (2012). A Lexicon Based Offline Arabic Handwritten Recognition Using Naive Bayesian Classifier with Gaussian. *Distribution. Image*, 15(16), 17.

Alginahi, Y. M. (2013). A survey on Arabic character segmentation. *Springer. International Journal on Document Analysis and Recognition (IJDAR)*, 16(2), 105-126.

Alijla, B. and Kwaik, K. (2012) OIAHCR: Online Isolated Arabic Handwritten Character Recognition Using Neural Network, *The International Arab Journal of Information Technology*, 9(4), 343-351.

Alkhateeb, J. H. (2015). OFF-LINE ARABIC HANDWRITTEN ISOLATED CHARACTER RECOGNITION. *International Journal of Engineering Science and Technology (IJEST)*, 7(7), 251-257.

Al-Ohali, Y., Cheriet, M., & Suen, C. (2003). Databases for recognition of handwritten Arabic cheques. *Pattern Recognition*, 36(1), 111-121.

Al-Sharaidah, M. A. (2000). Recognition of Handwritten Arabic Characters via Neural Networks (**Doctoral dissertation, M. Sc. Thesis, Jordan University**).

Amara, N. E. B., Mazhoud, O., Bouzrara, N., & Ellouze, N. (2005). ARABASE: A Relational Database for Arabic OCR Systems. *Int. Arab J. Inf. Technol.*, 2(4), 259-266.

Araki, N., Okuzaki, M., Konishi, Y., & Ishigaki, H. (2008). A statistical approach for handwritten character recognition using bayesian filter. *IEEE.In Innovative Computing Information and Control, 2008. ICICIC'08. 3rd International Conference on*, 194-194.

Asebriy, Z., Bencharef, O., Raghay, S., & Chihab, Y. (2014, November). Comparative systems of handwriting Arabic character recognition. *IEEE. In Complex Systems (WCCS), 2014 Second World Conference on* (90-93).

Bansal, S., Garg, M., & Kumar, M. (2014). A Technique for Offline Handwritten Character Recognition. *IJCAT International Journal of Computing and Technology*, 1(2), 210-215.

Bahashwan, M. A., & Bakar, A. (2014). A database of Arabic handwritten characters. *IEEE .In Control System, Computing and Engineering (ICCSCE), 2014 IEEE International Conference on.* 632-635).

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. *Elsevier. Journal of microbiological methods*, 43(1), 3-31.

Bhatia, N (2014). Optical Character Recognition Techniques: A Review, *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(5), 1219- 1223.

Blumenstein, M., Verma, B., & Basli, H. (2003, August). A novel feature extraction technique for the recognition of segmented handwritten characters. *IEEE. In Document Analysis and Recognition, Proceedings. Seventh International Conference on* 137-141.

Dileep, D. (2012). A feature extraction technique based on character geometry for character recognition. *Cornell Univ. Libr*, 1-4.

Dong, T., & Shang, W. (2011). Identification of Sensitive Information Based on Improved Naive Bayesian Classifier. In Computational Sciences and Optimization (CSO), *IEEE. 2011 Fourth International Joint Conference on*, 816-820.

El Qacimy, B., Hammouch, A., & Kerroum, M. A. (2015). A review of feature extraction techniques for handwritten Arabic text recognition. *IEEE. In Electrical and Information Technologies (ICEIT), 2015 International Conference on* (241-245).

Elglaly, Y., & Quek, F. (2011). Isolated Handwritten Arabic Character Recognition using Multilayer Perceptron and K Nearest Neighbor Classifiers. *filebox.vt.edu*.

Geist, J., Wilkinson, R. A., Burges, C., Creecy, R., Hammond, B., Hull, J. & Wilson, C. (1992). *First Census Optical Character Recognition System Conference. as a NISTIR*.

Gonzalez , R. C. and Woods, R. E. (1993). Digital Image Processing, *Addison-Wesley, Reading, Massachussets*.

Guyon, I., Haralick, R. M., Hull, J. J., & Phillips, I. T. (1997). Data sets for OCR and document image understanding research. *Handbook of character recognition and document image analysis*, 779-799.

Hammad, N. H., & Elhafiz, M. A (2015). Divide And Conquer Method For Arabic Character Recognition. *IOSR Journal of Engineering (IOSRJEN)*,5(4), 36-41.

Hemalatha, I., Varma, G. S., & Govardhan, A. Social network analysis and mining using machine learning techniques.

Hecht-Nielsen, R., (1990). Neurocomputing. Addison-Wesley, *Reading, MA*.

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. **IEEE. Computer**, (3), 31-44.

Jain, A., Nandakumar, K., & Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern recognition*, 38(12), 2270-2285.

Jamal, A. (2015). *End-Shape Analysis for Automatic Segmentation of Arabic Handwritten Texts* (Doctoral dissertation, Concordia University).

Jiang, L., Zhang, H., & Cai, Z. (2009). A novel Bayes model: Hidden naive Bayes. *Knowledge and Data Engineering, IEEE Transactions on*, 21(10), 1361-1371.

Kavallieratou, E., Liolios, N., Koutsogeorgos, E., Fakotakis, N., & Kokkinakis, G. (2001). The GRUHD database of Greek unconstrained handwriting. *IEEE. In Document Analysis and Recognition. Proceedings. Sixth International Conference on*, 561-565.

Khurma, N., Ahmed, M., & Ward, R. (1999). A new comprehensive database of handwritten Arabic words, numbers, and signatures used for OCR testing. *In Electrical and Computer Engineering, IEEE Canadian Conference on*, 2, 766-768.

Lawgali A. (2015). A Survey on Arabic Character Recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 8(2), 401-426.

Lawgali, A., Angelova, M., & Bouridane, A. (2013). HACDB: Handwritten Arabic characters database for automatic character recognition. *In Visual Information Processing (EUVIP), 4th European Workshop on*, 255-259.

Lawgali, A., Angelova, M., & Bouridane, A. (2014). A Framework for Arabic Handwritten Recognition Based on Segmentation. *International Journal of Hybrid Information Technology*, 7(5), 413-428.

Li, H., Doermann, D., & Kia, O. (2000). Automatic text detection and tracking in digital video. *Image Processing, IEEE Transactions on*, 9(1), 147-156.

Lim, J. S. (1990). Two-dimensional signal and image processing. *Englewood Cliffs, NJ, Prentice Hall, New Jersey, 07458*.

Lippmann, R. P. (1987). An Introduction to Computing with Neural Nets. *IEEE ASSP Magazine*, 4(2), 4-22.

Mani, N., & Srinivasan, B. (1997). Application of artificial neural network model for optical character recognition. *IEEE. In Systems, Man, and Cybernetics, Computational Cybernetics and Simulation., International Conference on* (3), 2517-2520.

Marti, U. V., & Bunke, H. (2002). The IAM-database: an English sentence database for offline handwriting recognition. *Springer International Journal on Document Analysis and Recognition*, 5(1), 39-46.

Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica, IEEE Transactions on Systems, Man, and Cybernetics* 9(1), 62–66.

Pechwitz, M., Maddouri, S. S., Märgner, V., Ellouze, N., & Amiri, H. (2002). IFN/ENIT-database of handwritten Arabic words. *In Proc. of CIFED*, (2), 127-136.

Pop, I. (2006). An approach of the Naive Bayes classifier for the document classification. *General Mathematics*, 14(4), 135-138.

Rashad, M., & Semary, N. A. (2014). Isolated Printed Arabic Character Recognition Using KNN and Random Forest Tree Classifiers. *Springer International publishing .In Advanced Machine Learning Technologies and Applications* 11-17.

Ribeiro, M. I. (2004). Gaussian probability density functions: Properties and error characterization. **Institute for Systems and Robotics, Lisboa, Portugal.**

Ruck, D., Rogers, S., Kabrisky, M., Maybeck, P., and Oxley, M. (1992), Comparative Analysis of Backpropagation and the Extended Kalman Filter for Training Multilayer Perceptrons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(6), 686-691.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation. *CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE*.

Sahlol, A. T., Suen, C. Y., Basyouni, M. R., & Sallam, A. A. (2014 A). A Proposed OCR Algorithm for the Recognition of Handwritten Arabic Characters. *Journal of Pattern Recognition and Intelligent Systems*, 2(1), 90-104.

Sahlol, A. T., Suen, C. Y., Elbasyoni, M. R., & Sallam, A. A. (2014 B). Investigating of Preprocessing Techniques and Novel Features in Recognition of Handwritten Arabic Characters. *Springer International Publishing. In Artificial Neural Networks in Pattern Recognition*, 264-276).

Sahlol, A., & Suen, C. (2014). A Novel Method for the Recognition of Isolated Handwritten Arabic Characters. *arXiv preprint arXiv:1402.6650*.

Schalkoff, R. J. (1997). Artificial neural networks. *ACM.McGraw-Hill Higher Education*.

Shah, P., Karamchandani, S., Nadkar, T., Gulechha, N., Koli, K., & Lad, K. (2009). OCR-based chassis-number recognition using artificial neural networks. *IEEE In Vehicular Electronics and Safety (ICVES), International Conference on* 31-34.

Song, Y., Zhu, Y., Zheng, E., Tao, F., & Wang, Q. (2014). Classifier Selection for Locomotion Mode Recognition Using Wearable Capacitive Sensing Systems. *Springer International Publishing. In Robot Intelligence Technology and Applications2*, 763-774.

Sossa-Azuela, J. H., Cuevas-Jiménez, E. V., & Zaldivar-Navarro, D. (2010). Computation of the Euler number of a binary image composed of hexagonal cells. *Journal of applied research and technology*, 8(3), 340-351.

Tlemsani, R., & Benyettou, A. (2012). Arabic on line characters recognition using improved dynamic Bayesian Networks. *IEEE.In Multimedia Computing and Systems*.

Vijaykumar B , Vikramkumar & Trilochan(2014). Bayes and Naive-Bayes Classifier,*arXiv:1404.0933*.

Wu, J., Pan, S., Zhu, X., Cai, Z., Zhang, P., & Zhang, C. (2015). Self-adaptive attribute weighting for Naive Bayes classification. *Expert Systems with Applications*, 42(3), 1487-1502.

Wu, V., Manmatha, R., & Riseman, E. M. (1999). Textfinder: An automatic system to detect and recognize text in images. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11), 1224-1229.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 algorithms in data mining. *Springer. Knowledge and Information Systems*, 14(1), 1-37.

Zambom, A. Z., & Dias, R. (2012). A review of Kernel density estimation with applications to econometrics. *arXiv preprint arXiv:1212.2812*.

Zenzo, S. D., Cinque, L. and Levivaldi, S. (1996). Runbased algorithms for binary image. *IEEE Transactions on PAMI.analysis and processing*, 18(1), 83-89.

Zhang, Q. J. and Gupta, K. C. (2000), Neural Networks for RF and Microwave Design. *Artech House*.

Zupan, J., & Gasteiger, J. (1993). Neural networks for chemists: an introduction. *ACM. John Wiley & Sons, Inc*.

Rouhani, S., & Ravasan, A. Z. (2013). ERP success prediction: An artificial neural network approach. *Scientia Iranica*, 20(3), 992-1001.

Salouan, R., Safi, S., & Bouikhalene, B. (2015). Handwritten Arabic Characters Recognition Using Methods Based on Racah, Gegenbauer, Hahn, Tchebychev and

Orthogonal Fourier-Mellin Moments. *International Journal of Advanced Science and Technology*, 78, 13-28.

Zawaideh ,F.H.(2012) Arabic Handwritten Character Recognition Using Modified Multi Neural Network, *Journal of Emerging Trends in Computing and Information Sciences*, 3, (7), 1021- 1026.

Bowman, A. W., and A. Azzalini. *Applied Smoothing Techniques for Data Analysis*. *New York: Oxford University Press Inc.*, 1997.

Appendix

UNITED STATES POSTAL SERVICE® **POSTAL MONEY ORDER**

Serial Number	Year, Month, Day	Post Office	U.S. Dollars and Cents
22218318450	2015-05-20	483120	\$300.00
Pay to: CenParmi Concordia University		Amount	THREE HUNDRED DOLLARS & 00c *****
Pay to	Mr. Nicola Nobile Concordia University CenParmi		Clerk
Address	1455 De Maisonneuve Blvd	From	Zeinab Alabdali
	West, Suite EV 003.403 Montreal, Canada	Address	33172 Shrewsbury Dr
Memo	Ahmed Subhi Abdalkafar		MI, 48310

© 2008 United States Postal Service. All Rights Reserved. SEE REVERSE WARNING • NEGOTIABLE ONLY IN THE U.S. AND POSSESSIONS

00000800 22218318450

Ms. Zeinab Alabdali
33172 Shrewsbury Dr.
Sterling Hts., MI 48310-6422


UNITED STATES POSTAL SERVICE®

1000 00299 00101381-03

U.S. POSTAGE
PAID
STERLING HEIGHT, MI
48311
MAY 20, 15
AMOUNT
\$0.71

TO: Mr. Nicola Nobile
concordia University CenParmi
1455 De Maisonneuve Blvd
West, Suite EV 003.403
Montreal, Quebec, Canada, H3G 1M8

U.S. FIRST CLASS PERMIT NO. 4831 STERLING HEIGHTS, MI

* ARCHIVE DOC		WPX	DHL EXPRESS
Not to be attached to Package			
From:	CONCORDIA UNIVERSITY CENP ARMI	Contact: 5148482424X7950	Origin: YUL
	MR. NICOLA NOBILE - RESEARCH MANAGE 1515 SAINT CATHERINE WEST EV3.403 OR EV3.180		
	MONTREAL H3G 2W1 Canada		
To:	AHMED SUBHI ABED ALGHAFOUR AHMED SUBHI ABED ALGHAFOUR ABDULLAH GHOSHEH ST.	Phone no:	
amman			
Jordan			
Product:	48 EXPRESS WORLDWIDE	Features/Services:	C
Ref Code:	Account 956624909	Shipment Weight: 0.5 lb	Pieces
		Shipping Date: 2015-06-10	1
Content:	DVD for language		
			InsuredValue

```
% This script file reads and show images

% Clean workspace ...
clear all;
close all;
clc;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Select Training and Test Datasets
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
|
dirIndex = [mainFolders.isdir];
subDirs = {mainFolders(dirIndex).name};
features3x3 = []; % To be filled with features ...

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% LABELING
% Define map container .
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

KeySet = {'Alif' , 'Baa' , 'Taa', 'Thaa', 'Jeem', 'Haa', 'Kha', 'Daal', 'Thaal', 'Raa', ...
```