UNIVERSITI TEKNIKAL MALAYSIA MELAKA

ENHANCED TRIANGLE GEOMETRY SHAPE FEATURES USING MULTI-STAGE FEATURE EXTRACTION FOR ARABIC HANDWRITTEN WORD RECOGNITION

NUR ATIKAH BINTI ARBAIN

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

DOCTOR OF PHILOSOPHY

2024

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Faculty of Information and Communication Technology

ENHANCED TRIANGLE GEOMETRY SHAPE FEATURES USING MULTI-STAGE FEATURE EXTRACTION FOR ARABIC HANDWRITTEN WORD RECOGNITION

Nur Atikah Binti Arbain

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Doctor of Philosophy

2024

ENHANCED TRIANGLE GEOMETRY SHAPE FEATURES USING MULTI-STAGE FEATURE EXTRACTION FOR ARABIC HANDWRITTEN WORD RECOGNITION

NUR ATIKAH BINTI ARBAIN

A thesis submitted
in fulfillment of the requirements for the degree of
Doctor of Philosophy

UTeM
اونيۏرسيتي تيکنيکل مليسيا ملاک
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2024

DEDICATION

To myself.

Nur Atikah Arbain

To my parents.

Arbain Basri

Siti Aisah Samlan

To my family.

Mohd Nazrien Zaraini

Burairah

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

ABSTRACT

Handwritten word is one of the current manuscript studies that has been intensively researched for many years because it is difficult to identify the styles, patterns, and signatures, especially the Arabic handwriting word. The Arabic letters have special unique characteristic such as dotted letters which can be divided into one dot, two dots and three dots (ش ,ي ,ظ), the writing direction and certain letters in Arabic script exhibit shape variations when connected to adjacent letters within a word. The Arabic word contains more than one character, which makes it quite a challenge due to the difficulties in recognizing the letter connections, the position-dependent letter shape, and the different writing style, which has led many researchers to separate each Arabic letter from the word, including the dots of the Arabic letters for the purpose of recognition. The Arabic dot letters are often considered as noisy features due to their irregular position on word which resulted in their removal during the preprocessing process, which may lead to incorrect data being extracted during feature extraction. The effect of eliminating dots in Arabic recognition is important because dots are used to distinguish Arabic characters that have a similar letter, such as (ب, ت, ث). This problem can be addressed by applying a global technique that considers the entire image of the Arabic word for feature extraction. Among the image features extracted, feature-based shape has the potential to recognize Arabic word handwriting based on geometric shapes. The triangle geometry method is applied, where the sides of the length of the triangle can be used as feature-based shape by using the appearances of triangle geometry to represent the Arabic word handwriting. Moreover, the multi-stage feature extraction is employed to extract various features such as angles and points of the triangle geometry shape. This study aims to propose the enhanced triangle geometry shape features for Arabic word handwriting recognition using multi-stage feature extraction. The objectives include to propose the triangle geometry shape features to represent the Arabic handwritten word images, to propose a method named Multi-Stage Feature Extraction of Triangle Geometry Shape Appearance Method (MFeTSA) to extract the triangle geometry shape features, and to improve prior straight-line problem solution by using proposed technique of Triangle Coordinate Point Rearrangement (TCPR). A dataset of Arabic (AHDB) and Iran (Iranshahr) is used for this study. The experimental results are obtained by comparing the classification accuracy results with Support Vector Machine, with promising results with high accuracy for both datasets AHDB (76.3825%) and Iranshahr (63.1563%). Statistical validation is used to check the quality of the proposed method, whether it is better than previous methods that use the same triangle geometry method to extract features from triangle geometry. Independent samples t-test is used to validate the proposed method, where the result based on the significant *p*-values shows that the proposed method is effective for both datasets of Iranshahr and AHDB.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

PENAMBAHBAIKAN FTUR BENTUK GEOMETRI SEGITIGA MENGGUNAKAN PENGEKSTRAKAN FITUR PELBAGAI PERINGKAT UNTUK PENGECAMAN TULISAN TANGAN ARAB

ABSTRAK

Kajian mengenai tulisan tangan merupakan salah satu bidang kajian manuskrip semasa yang telah dikaji secara intensif selama bertahun-tahun kerana ia sukar kenal pasti gaya, corak, dan tandatangan, terutamanya tulisan tangan Arab. Huruf Arab mempunyai ciri-ciri unik khas seperti huruf bergelung yang boleh dibahagikan kepada satu titik, dua titik, dan tiga titik (ش ,ي ,ظ), *arah penulisan, dan beberapa huruf tertentu dalam aksara Arab menunjukkan variasi bentuk apabila disambungkan dengan huruf bersebelahan dalam satu perkataan. Perkataan Arab mengandungi lebih daripada satu huruf, yang menjadikannya satu cabaran besar disebabkan kesulitan dalam mengenal pasti hubungan huruf, bentuk huruf bergantung pada kedudukan, dan gaya penulisan yang berbeza, yang telah mendorong ramai penyelidik untuk memisahkan setiap huruf Arab dari perkataan, termasuk titik-titik huruf Arab untuk tujuan pengiktirafan. Huruf titik Arab sering dianggap sebagai ciri-ciri bising kerana kedudukan yang tidak tetap pada perkataan yang mengakibatkan penghapusan semasa proses pra pemprosesan, yang mungkin mengakibatkan data yang salah diekstrak semasa mengekstrak ciri. Kesan penghapusan titik dalam pengiktirafan Arab adalah penting kerana titik digunakan untuk membezakan aksara Arab yang mempunyai huruf yang serupa, seperti* (ب, ت, ث). *Masalah ini boleh diatasi dengan menerapkan teknik global yang mempertimbangkan keseluruhan imej perkataan Arab untuk ekstrak ciri. Di antara ciri imej yang diekstrak, bentuk berdasarkan ciri mempunyai potensi untuk mengenal pasti tulisan tangan Arab berdasarkan bentuk geometri. Kaedah geometri segitiga digunakan, di mana sisi panjang segitiga boleh digunakan sebagai bentuk berasaskan ciri dengan menggunakan penampilan geometri segitiga untuk mewakili tulisan tangan Arab. Selain itu, pengeluaran ciri berperingkat digunakan untuk mengekstrak pelbagai ciri seperti sudut dan titik bentuk geometri segitiga. Kajian ini bertujuan untuk mengusulkan ciri-ciri bentuk geometri segitiga yang ditingkatkan untuk pengiktirafan tulisan tangan perkataan Arab menggunakan pengeluaran ciri berperingkat. Objektif termasuk mengusulkan ciri bentuk geometri segitiga untuk mewakili imej perkataan tulisan tangan Arab, mengusulkan kaedah yang dinamakan Kaedah Pengeluaran Ciri Berperingkat Bentuk Geometri Segitiga (MFeTSA) untuk mengekstrak ciri bentuk geometri segitiga, dan meningkatkan penyelesaian masalah garis lurus terdahulu dengan menggunakan teknik yang dicadangkan iaitu Teknik Penyatuan Semula Titik Koordinat Segitiga (TCPR). Data set Arab (AHDB) dan Iran (Iranshahr) digunakan untuk kajian ini. Hasil eksperimen diperoleh dengan membandingkan hasil ketepatan klasifikasi dengan Mesin Vektor Sokongan, dengan hasil yang memberangsangkan dengan ketepatan yang tinggi untuk kedua-dua data set AHDB (76.3825%) dan Iranshahr (63.1563%). Pengesahan statistik menggunakan ujian t-sampel independen untuk menilai kualiti kaedah yang dicadangkan, menunjukkan kesan yang berkesan pada kedua-dua data set Iranshahr dan AHDB dengan hasil yang signifikan pada nilai p.*

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

ACKNOWLEDGEMENT

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ

First and foremost, I would like to thank and praise Allah the Almighty, my Creator, my Sustainer, for everything I received since the beginning of my life. I would like to extend my appreciation to the Universiti Teknikal Malaysia Melaka (UTeM) for providing the research platform. Thank you also to the Malaysian Ministry of Higher Education (MOHE) and Zamalah Scholarship under UTeM for the financial assistance.

My utmost appreciation goes to my main supervisor, Assoc. Prof. Ts. Dr. Mohd Sanusi Azmi for all his support, advice, and inspiration. His constant patience for guiding and providing priceless insights will forever be remembered. Also, to my co-supervisor, Prof. Dr. Azah Kamilah Draman@Muda, who constantly supported my journey. My special thanks go to my supportive friends from Computational Intelligence and Technology (CIT) for all the help and support I received from them especially Amirul Ramzani Radzid, Eisy Humaira Abdul Azziz, Liew Siaw Hong, and Azrina Tahir during my Ph.D. journey.

Last but not least, from the bottom of my heart a gratitude to my beloved husband, Mohd Nazrien Zaraini, for his encouragement and who has been the pillar of strength in all my endeavors. My eternal love also to my daughter, Burairah, for enlightening my life. I would also like to thank my beloved parents (Arbain Basri and Siti Aisah Samlan) for their endless support, love, and prayers. Finally, thank you to all the individual(s) who had provided me with the assistance, support, and inspiration to embark on my study.

اونيۏرسيتي تيکنيکل مليسيا ملاک
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

TABLE OF CONTENTS

		PAGES
DECLARATION		
APPROVAL		
DEDICATION		
ABSTRACT		**i**
ABSTRAK		**ii**
ACKNOWLEDGEMENT		**iii**
TABLE OF CONTENTS		**iv**
LIST OF TABLES		**vii**
LIST OF FIGURES		**x**
LIST OF ABBREVIATIONS		**xiii**
LIST OF SYMBOLS		**xiv**
LIST OF APPENDICES		**xv**
LIST OF PUBLICATIONS		**xvi**
CHAPTER		
1.	**INTRODUCTION**	**1**
	1.1 Introduction	1
	1.2 Research Background	5
	1.3 Problem Statement	7
	1.4 Research Question	13
	1.5 Research Objective	14
	1.6 Hypothesis Statement	17
	1.7 Research Scope	17
	1.8 Research Significance	18
	1.9 Thesis Organization	18
	1.10 Summary	21
2.	**LITERATURE REVIEW**	**22**
	2.1 Introduction	22
	2.2 Characteristics of Arabic Language	23
	2.2.1 Arabic Words	28
	2.2.2 Arabic Writing Direction	29
	2.3 Challenges in Arabic Handwritten Word Recognition	29
	2.4 Analysis Phases of Arabic Handwritten Word Recognition	36
	2.4.1 Pre-Processing	36
	2.4.2 Feature Extraction	41
	2.4.3 Classification	52
	2.4.4 Validation	60
	2.5 Triangle Geometry Method in Digit Recognition	63
	2.5.1 Determination of Triangle Shape Points	64
	2.5.2 Triangle Features Formation	70
	2.6 Multi-Stage Feature Extraction in Image Recognition	75
	2.7 Problems in Prior Arabic Handwritten Word Research	79

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2.7.1 Arabic Dataset Problem 80
2.7.2 Existing Proposed Technique Problem 81
2.8 Summary 83

3. METHODOLOGY 85
3.1 Introduction 85
3.2 Research Framework 86
3.2.1 Investigation Phase 86
3.2.2 Implementation Phase 91
3.3 Dataset Collection 92
3.3.1 Arabic Handwritten Database (AHDB) 92
3.3.2 Testing and Training Data For AHDB 94
3.3.3 Iranshahr Database 96
3.3.4 Sorting File For Iranshahr 97
3.3.5 Renaming Iranshahr Images 98
3.4 Pre-Processing 99
3.4.1 Binarization Process 101
3.4.2 Blank Space Elimination 103
3.4.3 Image Size Normalization 106
3.5 Proposed Multi-Stage Feature Extraction Method 108
3.6 Analysis of Proposed Technique, Classification and Validation 109
3.7 Operational Framework 110
3.8 Experimental Design 112
3.8.1 Experiment I 112
3.8.2 Experiment II 112
3.8.3 Experiment III 113
3.8.4 Experiment IV 113
3.8.5 Experiment V 114
3.9 Research Tools 114
3.10 Summary 116

4. CONSTRUCTION OF TRIANGLE GEOMETRY 117
4.1 Introduction 117
4.2 Identify Types of Triangle Geometry Shapes 117
4.2.1 Identify Possible Generated Triangle Geometry Shape 118
4.3 Feature-Based Appearance of Triangle Shape 125
4.4 Phase I 126
4.4.1 Analyze Potential Appearance of Triangle Shape 127
4.5 Phase II 135
4.5.1 Select Triangle Properties 137
4.6 Summary 141

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

5. MULTI-STAGE FEATURE EXTRACTION OF TRIANGLE GEOMETRY SHAPE APPEARANCE **143**
5.1 Introduction 143
5.2 Overview of Multi-Stage Feature Extraction 143
5.3 Feature Extraction Stage I 145
5.4 Feature Extraction Stage II 148
5.5 Summary 150

6. TRIANGLE COORDINATE POINT REARRANGEMENT TECHNIQUE **151**
6.1 Introduction 151
6.2 Detection of Straight-Line Problem 151
6.2.1 Locate Angle Problem 153
6.3 Overview of Triangle Coordinate Point Rearrangement (TCPR) 156
6.3.1 Inequality Rules 157
6.3.2 Find Range of Value 158
6.3.3 Proposed Coordinate Rearrangement 159
6.4 Result of Proposed Technique (TCPR) 163
6.5 Summary 164

7. RESULT AND DISCUSSION **165**
7.1 Introduction 165
7.2 Performance Analysis 166
7.3 Statistical Validation 175
7.4 Summary 181

8. CONCLUSION AND FUTURE WORKS **182**
8.1 Introduction
8.2 Research Summary 182
8.3 Research Findings 184
8.3.1 To propose the triangle geometry shape features to represent the Arabic handwritten word images. 184
8.3.2 To propose a method named Multi-Stage Feature Extraction of Triangle Geometry Shape Appearance Methof (MFeTSA) to extract triangle geometry shape features. 185
8.3.3 To improve prior straight-line problem solution by using proposed technique of Triangle Coordinate Point Rearrangement (TCPR). 186
8.4 Research Contributions 187
8.5 Conclusion 187
8.6 Research Limitations 189
8.7 Future Works 190

REFERENCES **191**
APPENDICES **201**

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF TABLES

TABLE	TITLE	PAGE
Table 1.1	Summary of RP, RQ and RO	15
Table 1.2	Description of thesis organization	19
Table 2.1	Arabic letters	23
Table 2.2	Persian letters (Khosravi and Chalechale 2022)	25
Table 2.3	Arabic letters based on dotted group	27
Table 2.4	Example of changes on Arabic letters in word	28
Table 2.5	Example of Arabic scripted in two types of writing forms	29
Table 2.6	Summary of pre-processing in Arabic handwritten word recognition	37
Table 2.7	Summary of features representation method and image feature	42
Table 2.8	Summary of recognition technique used by related researchers	44
Table 2.9	The table of confusion matrix for basis measure	53
Table 2.10	Summary of classifier used by related researchers	57
Table 2.11	Summary of MCAR, MAR and NMAR by Sivakani and Ansari (2020)	59
Table 2.12	Position of coordinate y for Point C (Azmi, 2013)	66
Table 2.13	Algorithm of Point Process technique (Azmi, 2013)	69
Table 2.14	Description of triangle features based on Figure 2.12	71
Table 2.15	Algorithm of Scalene technique (Azmi, 2013)	71
Table 2.16	Description of zoning method	75
Table 3.1	Summary of investigation phase	88
Table 3.2	Classification of accuracy from various methods	91
Table 3.3	Description of AHDB dataset	93

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Table 3.4	Description of total distribution for validation purpose	95
Table 3.5	Description of Iranshahr dataset	97
Table 3.6	Algorithm of renaming filename process	99
Table 3.7	An algorithm for Blank Space Elimination (BSE)	105
Table 3.8	Accuracy result of image size normalization	107
Table 3.9	An algorithm for image size normalization	107
Table 3.10	Description of software used for this research	115
Table 3.11	Description of hardware used for this research	116
Table 4.1	Description of measurement angle used in Experiment I	118
Table 4.2	Description of conditions in Figure 4.2	121
Table 4.3	An algorithm of ITriS used in Experiment I	123
Table 4.4	Description of sub-types of potential appearance for acute scalene triangle	129
Table 4.5	Position of y-coordinates for sub-types	130
Table 4.6	Distribution based on number of triangle shape are formed	134
Table 4.7	Description of triangle shape properties	135
Table 4.8	The accuracy result using grid search algorithm	139
Table 4.9	The accuracy result using Random Forest and *k*-NN	140
Table 6.1	An algorithm of proposed coordinate rearrangement (TCPR)	128
Table 6.2	Classification accuracy based on SVM and Random Forest (RF)	163
Table 7.1	Best recognition rate achieved using grid search algorithm	166
Table 7.2	AHDB dataset result based on cost (c) = 1.0 and gamma (γ) = 0.0	167
Table 7.3	Iranshahr dataset result based on cost (c) = 1.0 and gamma (γ) = 0.0	168

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Table 7.4	Results based on SVM kernel functions using different value of cost (c) and gamma (γ)	169
Table 7.5	Result of classification accuracy using SVM from WEKA tool	170
Table 7.6	Result of accuracy based on *k*-fold cross-validation technique	172
Table 7.7	Comparison performance analysis-based classification accuracy	173
Table 7.8	Normality test results	176
Table 7.9	Homogeneity of variances results for AHDB dataset	177
Table 7.10	Result based mean for Kruskal-Wallis *H* test	177
Table 7.11	Result test statistics for AHDB dataset	177
Table 7.12	Result of group statistics using independent sample *t*-test	178
Table 7.13	Result of Levene test for equality of variances	178
Table 7.14	Result using independent sample *t*-test	179
Table 7.15	Result of initial report using comparing means test	180
Table 7.16	Result of ANOVA using eta squared	180
Table 7.17	Result of measures of association	181

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF FIGURES

FIGURE	TITLE	PAGE
Figure 1.1	Example of combined character for Arabic handwriting	8
Figure 1.2	Example of various diacritics (dots) used for same Arabic body letter	9
Figure 1.3	Example of various dots in Arabic character	9
Figure 2.1	Related work based on dataset of handwritten word recognition	35
Figure 2.2	Pre-processing in Arabic word handwriting recognition	37
Figure 2.3	Binarization process of document page (Tahir et al., 2018)	39
Figure 2.4	Sinusoidal plane result adapted by a Gaussian Kernel (Elzobi et al., 2014)	45
Figure 2.5	The proposed block diagram system from Jayech et al. (2016)'s work	46
Figure 2.6	Separated characters from Salem (2017)'s work	47
Figure 2.7	Word image extraction process (Fazel et al., 2016)	48
Figure 2.8	The image process using HOG method (Kadhm and Hassan 2016)	49
Figure 2.9	Feature extraction block diagram (Tavoli et al., 2018)	50
Figure 2.10	An overview of recognition step in Arani et al. (2019)'s work	51
Figure 2.11	Types of triangle geometry shape	64
Figure 2.12	Positions of Point A, Point B and Point C by Azmi (2013)	66
Figure 2.13	Example of triangle shape image based on location of Point C in Azmi (2013)	67
Figure 2.14	Point A (*Bucu A*) and Point B (*Bucu B*) by Azmi (2013)	68
Figure 2.15	Division of zoning method by Azmi (2013)	74
Figure 2.16	Single-stage CNN structure (SS-ConvNets) by Wu et al. (2016)	77
Figure 2.17	Multi-stage CNN structure (MS-ConvNets) by Wu et al. (2016)	78

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Figure 2.18	An overview of related problems in prior Arabic word handwriting	79
Figure 3.1	Summary of Arabic Handwritten Word Recognition framework	87
Figure 3.2	Summary of implementation phase	92
Figure 3.3	Example image of عبد from AHDB dataset	93
Figure 3.4	Example of 'gayr' and 'tm' image, respectively	94
Figure 3.5	Process of AHDB testing and training data distribution Process	95
Figure 3.6	of validation distribution	96
Figure 3.7	Example of Iranshahr images	96
Figure 3.8	Process of sorting Iranshahr image file	98
Figure 3.9	Process of renaming filename	98
Figure 3.10	An overview of pre-processing process	100
Figure 3.11	Example of binary image	101
Figure 3.12	Binarization process	101
Figure 3.13	Binarization result based on various threshold values for AHDB dataset	102
Figure 3.14	Process elimination step by step	103
Figure 3.15	Four process blank space elimination	104
Figure 3.16	Output image after applied BSE technique	104
Figure 3.17	Proposed Multi-Stage Feature Extraction Method	109
Figure 3.18	Flow chart of operational framework	109
Figure 4.1	Experiment I summary process	119
Figure 4.2	Flow chart of Experiment I	120
Figure 4.3	Output of Experiment I	122
Figure 4.4	Development of feature-based appearance of triangle geometry shape	125
Figure 4.5	The position of denoted points of triangle geometry shape	126

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Figure 4.6	Flow chart of Experiment II	127
Figure 4.7	Binary image from AHDB dataset	128
Figure 4.8	Sub-type of shape Type 1	131
Figure 4.9	Sub-type of shape Type 2	131
Figure 4.10	Sub-type of shape Type 3	132
Figure 4.11	Sub-type of shape Type 4	132
Figure 4.12	Sub-type of shape Type 5	133
Figure 4.13	Sub-type of shape Type 6	133
Figure 4.14	Output of analysis for Experiment II based on bar chart graph	134
Figure 4.15	Triangle geometry shape based on three lengths of sides	136
Figure 4.16	Triangle geometry shape based on two length of sides and one angle	136
Figure 4.17	Triangle geometry shape based on one length of side and two angles	137
Figure 4.18	Process of triangle properties selection	138
Figure 5.1	The multi-stage feature extraction of this research	144
Figure 5.2	Example of three generated points	146
Figure 5.3	Illustration of triangle points on AHDB dataset	147
Figure 5.4	Three length of sides on binary image of AHDB dataset	148
Figure 5.5	Position of Point A1, B1 and C1 used in this study	149
Figure 6.1	Illustration of straight-line problems	152
Figure 6.2	In detail illustration of straight-line problem in Figure 6.1(d)	154
Figure 6.3	Flow chart of Experiment V	155
Figure 6.4	Illustration of subword image from AHDB dataset	158
Figure 6.5	The Arabic word handwriting image is divided into six zones	158
Figure 6.6	Before and after applied TCPR	162

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF ABBREVIATIONS

AHDB	-	Arabic Handwritten Database
BSE	-	Blank Space Elimination
k-NN	-	*k*-Nearest Neighbor
ITriS	-	Identify Triangle Shape Technique
LIBSVM	-	Library of Support Vector Machine
MSA	-	Mohd Sanusi Azmi (2013)
MFeTSA	-	Multi-Stage Feature Extraction of Triangle Geometry Shape Appearance Method
RF	-	Random Forest
Ri	-	Resize image technique
SVM	-	Support Vector Machine
TCPR	-	Triangle Coordinate Point Rearrangement Technique

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF SYMBOLS

°	-	Degree angle of triangle geometry shape
c	-	Value of cost used for parameter in LibSVM
γ	-	Value of gamma used for parameter in LibSVM
τ	-	Threshold value

UTeM
اونيۏرسيتي تيکنيکل مليسيا ملاك
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
Appendix A	Grid Search Results for SVM Classification Accuracy	201

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF PUBLICATIONS

The followings are the list of publications related to the work on this thesis:

N. A. Arbain, M. S. Azmi, A. K. Muda, A. R. Radzid, K. Omar and N. A. Emran, 2019. "Digital Jawi Palaeography: Studies from the Perspective of Computer Science" *Journal of Advanced Computing Technology and Application (JACTA),* vol. 1, no 1, pp. 24-29.

N. A. Arbain, M. S. Azmi, A. K. Muda, A. R. Radzid, K. Omar and N. A. Emran, 2019. "State of the Art in Digital Paleography" *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE),* vol. 8, no 1.3, pp. 246-250.

N. A. Arbain, M. S. Azmi, A. K. Muda, A. R. Radzid and A. Tahir, 2019. "A Review of Triangle Geometry Features in Object Recognition" *2019 IEEE 9th Symposium on Computer Applications and Industrial Electronics (ISCAIE),* pp. 254-258.

N. A. Arbain, M. S. Azmi, A. K. Muda, A. R. Radzid and A. Tahir, 2018. "Triangle Shape Feature Based on Selected Centroid for Arabic Subword Handwriting" *International Journal of Advanced Computer Science and Applications (IJACSA),* vol. 9, no 12, pp. 169-173.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

CHAPTER 1

INTRODUCTION

1.1 Introduction

Prior to the widespread integration of technology, individuals utilized paper for written communication and task execution. Despite technological advancements, writing persisted, albeit with the assistance of applications. The uniqueness of individuals' handwriting, characterized by distinct styles, patterns, and signatures, has persisted. To facilitate the seamless transition of handwritten content into digital formats, handwriting recognition has been introduced. This technology empowers computers to understand human handwriting, thereby enhancing accessibility and processing capabilities for digitized information.

Handwriting recognition serves as a crucial technology in translating human handwriting into algorithms comprehensible by computers. This capability plays a pivotal role in the domain of document analysis and recognition. A notable area of ongoing research involves the intricacies of handwritten word recognition, a field that has garnered significant attention over the years. This emphasis stems from the inherent challenges associated with discerning diverse styles, patterns, and signatures present in handwritten word manuscripts.

A specific focus within this research realm is directed towards handwritten Arabic words. The distinct characteristics inherent in the styles and shapes of Arabic letters pose a big challenge in the recognition process. Consequently, understanding and accurately identifying handwritten Arabic words necessitate sophisticated techniques and ongoing advancements in handwriting recognition technology. Arabic handwritten word recognition

is utilized in various applications such as optical character recognition, language translation and digital annotation to convert handwritten Arabic text into digital format for easier processing and analysis. Besides that, the Arabic handwritten word recognition finds application across diverse domains and caters to a broad range of users. The primary target users include researchers and practitioners engaged in document digitization, historical preservation, and linguistic studies within the Arabic-speaking world.

Systems for recognizing handwritten Arabic words have improved drastically over the years. In past decades, many databases of handwritten words (Hull, 1994, M. Pechwitz et al., 2002, Fernando et al., 2003, Al-Ma'adeed et al., 2004) have been developed that can be used as benchmarks for training and testing data for word recognition. Arabic handwritten documents are currently available in many resources in physical form and on the Internet, which poses a challenge for word recognition.

Languages such as Farsi and Urdu use Arabic characters in their writing, which makes the task more difficult due to the different words used, the strength, and the order of the writing (Khosravi and Chalechale, 2022, Rasheed et al., 2022). In relation, numerous researchers (Salem, 2017, Tavoli et al., 2018, Hamida et al., 2020, Kaur and Kumar, 2021, Zohrevand and Imani, 2021, Eltay et al., 2022, Khosravi and Chalechale, 2022, Korichi et al., 2022) have worked hard to overcome this difficulty by developing more competent algorithms and improving existing detection methods and systems.

In a general context, words consist of multiple characters and form integral components of sentences by breaking down individual words. This process involves decomposing words to understand their composition within the structure of a sentence. The word is isolated during the segmentation process, where it is processed from mapping by page, by line and finally the word was isolated from sentence for each of line. In the Arabic

language, words may comprise multiple interconnected components, as exemplified by the term "taskheel" (تشكيل). This characteristic introduces complexities in recognizing individual Arabic letters due to variations in their shapes at the beginning, middle, or end of words. Distinguishing isolated Arabic letters with diverse contextual forms, encompassing initial, middle, and final configurations, including dots, poses challenges in handwritten words due to deviations from precise and standardized writing, as observed in printed texts.

Apart from their distinctive contextual forms, certain Arabic letters sharing similar basic shapes are differentiated by the presence of dots, with variations in the number and placement of dots either above or below the letters. The specific arrangement of dots on Arabic letters is recognized as a physical trait that is frequently misconstrued as noise, particularly when extracting features from Arabic word images, owing to irregular positions in word writing. Moreover, certain Arabic characters, such as "ت", "ب" and "ث" share identical shapes, potentially causing inaccuracies in discerning the meaning of Arabic word scripts when dots are absent or isolated. This situation can further result in the extraction of erroneous data during feature extraction, leading to inaccurate outcomes in classification if the features extracted convey incorrect information from the segmented image.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

In earlier studies by Davoudi et al. (2016) and Jayech et al. (2016), segmentation approach was utilized by isolating individual Arabic characters within words, including the isolation of dots from the Arabic characters during the segmentation process. This approach aimed to mitigate potential challenges arising from the adjacency of characters within the same Arabic word. In addition, the poor images of handwritten words, incorrect letter spacing, noise, and poor paper quality further make difficult the recognition of subwords. Furthermore, the difficult connection between the hardware tools and the subsequent

operations such as scanning and binarization poses an additional challenge for the recognition of handwritten Arabic words (Khosravi and Chalechale, 2022).

Therefore, image processing plays an important role in recognizing Arabic handwritten word images. The image processing steps such as pre-processing, segmentation, feature extraction, classification, and validation are applied to process the Arabic handwritten word images by decomposing it into meaningful features that contain valuable information from the Arabic handwritten word image. There are many methods for handwritten words recognition, such as convolutional neural networks, semi-incremental recognition, incremental recognition, line and word segmentation, zoning, and ensemble, which can be used to generate features (Rosyda and Purboyo, 2018). In addition to pre-processing, feature extraction is also an important step in the word recognition process, which is used to extract features from Arabic handwritten word images.

The statistical features often employed by researchers (Fazel et al., 2016, Al-Nuzaili et al., 2017, Tavoli et al., 2018) to recognize Arabic handwritten word images. This is due to the characteristics of Arabic handwritten word images, which can be recognized on the basis of contours and regions which is known as shape representation techniques. The shape features using geometric shapes such as straight lines in different directions have been proposed by Tavoli et al. (2018) for Arabic/Persian handwritten word images, which provide promising results in terms of classification accuracy. However, the method proposed by Tavoli et al. (2018) generates a large size of length for the extracted features, resulting in a longer process.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA