

Arabic Calligraphy Classification using Triangle Model for Digital Jawi Paleography Analysis

Mohd Sanusi Azmi

Faculty of Information Communication and Technology
Universiti Teknikal Malaysia Melaka, Malaysia
Melaka, Malaysia
sanusi@utem.edu.my

Mohammad Faidzul Nasrudin

Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Malaysia
mfn@ftsm.ukm.my

Khairuddin Omar

Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Malaysia
mfn@ftsm.ukm.my

Azah Kamilah Muda

Faculty of Information Communication and Technology
Universiti Teknikal Malaysia Melaka, Malaysia
Melaka, Malaysia
azah@utem.edu.my

Azizi Abdullah

Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Malaysia
azizi@ftsm.ukm.my

Abstract— Calligraphy classification of the ancient manuscripts gives useful information to paleographers. Researches on digital paleography using calligraphy are done on the manuscripts to identify unidentified place of origin, number of writers, and the date of ancient manuscripts. Information that are used are features from characters, tangent value and features known as Grey-Level Co-occurrence Matrix (GLCM). For Digital Jawi Paleography, a novel technique is proposed based on the triangle. This technique defines three important coordinates in the image of each character and translates it into triangle geometry form. The features are extracted from the triangle to represent the Jawi (Arabic writing in Malay language) characters. Experiments have been conducted using seven Unsupervised Machine Learning (UML) algorithms and one Supervised Machine Learning (SML). This stage focuses on the accuracy of Arabic calligraphy classification. Hence, the model and test data are Arabic calligraphy letters taken from calligraphy books. The number of model is 711 for the UML and 1019 for the SML. Twelve features are extracted from the formed triangles used.

Keywords— Paleography, Calligraphy, Jawi, Arabic, Triangle Model, Features Extraction

I. INTRODUCTION

The calligraphies applied in ancient manuscripts contain useful information for paleography research. Based on that, date, number of writers, place of origin and originality of manuscripts can be known [1-4].

From the previous researches done in [3], [5], [6], there are two approaches to identify calligraphy in the ancient manuscript. The first digital paleography was done in [7] named System for Paleography Inspection (SPI) using the local approach. The same approach was also used in [3], [8], whereas [6] used global approach.

The SPI system for Latin manuscripts that was developed for University of Pisa used centroid and tangent distance values. The values are taken from individual characters from 37 manuscripts book and were clustered using Dendrogram diagram [7][9]. However, the SPI development were not completed and applied in the University of Pisa [9]. Due to the lack of features in [7], [6] introduced global approach using twelve Haralick's features. The Haralick's features are also known as Grey-Level Co-occurrence Matrix (GLCM). The GLCM is a statistical feature. The GLCM features gives significant result but requires the same dimension of size for each image for modelling and testing [8]. This is because the GLCM will do multiplication of matrices.

Besides Roman calligraphy, the digital paleography research are also done in Hebrew and Malay [3], [4]. [3] used local features based on the space from selected image. However, the objects were only two Hebrew characters, and no justification is stated. The test conducted in [3] is also insufficient because only 14 documents were chosen with only twenty Aleph and Lamed chosen from each document.

For the Malay digital paleography, research is being made in [4] and [8]. In [4], a framework for digital Jawi

paleography is produced. The framework covers calligraphy and illumination usually found in Malay ancient manuscripts. In the framework, features can be global, local and hybrid. Those features may be applied for the Digital Jawi Paleography.

This paper is an enhancement of the previous model in [8]. Amendments has been made and test that have been conducted using eight machine learning algorithms discussed in the experimental result.

II. PRE-PROCESSING

The proposed method is based on further enhancement from [10] and [8]. Some improvements have been made in order to identify the style of Arabic calligraphy applied in the Malay ancient manuscripts. The steps involved in the stage are:

- Data collection
- Segmentation of Jawi characters
- Image categorization
- Binarization
- Features Extraction and proposed method






The Features Extraction and Proposed method will be detailed here because it is a novel technique for Digital Jawi Paleography that is currently researched in Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi.

A. Data collection

In [11], images are taken from Kekwa and test images are from Terengganu Inscribed Stone and the manuscript Hikayat Merong Mahawangsa. The images are categorized based on the six models of Triangle Blocks proposed in the paper.

For this paper, the images from the previous research were retained but are enriched with images taken from calligraphy books and expert. The books that are used to model the calligraphy characters are [12-19]. The sample of images for model and test are shown in Table 1.

TABLE I
SAMPLE OF ARABIC CALLIGRAPHY

	Arabic Calligraphy for character Ba, Ta and Tsa	Type of Arabic Calligraphy
		Diwani
		Riqah
		Thuluth
		Nasakh
		farisi

B. Segmentation of Jawi Character for the Model and Test images

The segmentation process of characters is done manually. This process is based on the works done in [3] and [20]. There are 1411 images that have been segmented and clustered into five categorizes suitable to Jawi ancient manuscripts [21] and [11]. The categories are i. Nasakh, ii. Thuluth, iii. Riqah, iv. Diwani and v. farisi. The number of images for the model is 722.

C. Binarization

The model and test of Arabic characters are automatically binarized using Otsu's method. Previously, the threshold is fixed as either 127 or 180 in [8], [2] and [11]. In this paper, amendment has been made by applying Otsu's method. The method dynamically chooses the discriminant threshold based on the foreground and background of image. So, the threshold value is more precise for images from various sources of Jawi ancient manuscripts. The purpose of this process is to remove noise and prepare images for the proposed features selection.











D. Image Categorization

Based on [12-14], [16], [18], some of the Arabic and Jawi characters share the same shape but differ in the presence or absence of diacritics and location of diacritics. Besides, most of the images have nearly the same shape. Table II below shows some of the characters with the same shape labeled 'A' and 'B' in capital form and nearly same shape as 'a' and 'b' in lower form. These groups are used in the testing phase.

The classification of Arabic calligraphy is different compared to Optical Arabic Character Recognition (OACR). The classification of Arabic calligraphy identifies the calligraphy applied in manuscripts. Whereas, the OACR identifies the characters in images.

Some of the calligraphy books such as [12-19] provide Arabic characters with no diacritics. The characters with the same shape are represented with no diacritics. Thus, the model and test images in this paper follow the approach presented in the books. In TABLE II shows images with and without diacritics. Characters that share the same shape are presented in one group. The images are grouped in Roman alphabet in capital letter form and images that near to the shape will be grouped with the same roman alphabet but in the lower form.

TABLE II
ARABIC CALLIGRAPHY WITH THE SAME SHAPE

Group	Character	Sample Images	
		Before	After
A	Ba		
	Ta		
	tsa		
a	Fa		
	Kaf		

	nun	ن	ن
	nun	ن	ن
B	sin	س	س
	shin	ش	س
b	Sad	ص	ص
	Dhad	ض	ص

E. Feature Extraction and proposed method

The feature extraction is based on [8]. Six features were selected from the triangle model that proposed in [8]. However, the model has been improved by dividing each image into four parts as shown in Fig. 1

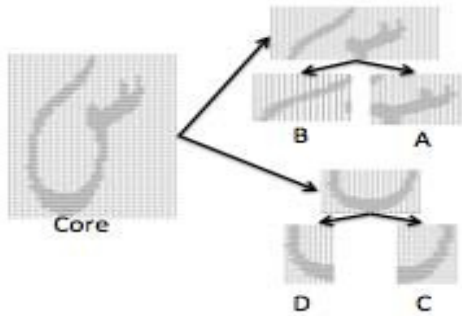


Fig. 1 Segregation of Isolated Characters to Four Parts

In Fig. 1, the image labeled as Core is divided into four parts labeled as A, B, C and D. The features from Core is made in reference to [8]. The same types of features are extracted from each of the parts.

In this paper, features are extracted from Core and Core with A. Part B, C and D are left for future researches. The experimental result in this paper is discussed in Experimental Result.

i. Extracting three important points.

Images that are freed from noises and diacritics are ready for the process of extracting three important coordinates [8]. The coordinates extracted are explained in Table III.

TABLE III
LOCATION OF TRIANGLE POINTS

Point	Label	Connected points	Angle	Points Location
Point 1	A	b and c	A	
Point 2	C	a and b	C	
Point 3	B	a and c	B	

Point A and B are taken from the centroid from the right and the left of the centroid image C as shown in TABLE III. These points are used to form a triangle. The features shown in TABLE III below will be extracted from each triangle.

ii. Extracting sub-points from four parts

After the features for Core are extracted. The images are divided into four parts. The features in part A are extracted with the same criteria for Point 1 to Point 3 in Table III. As mentioned earlier, this paper only considers the Core and part A. The location of points in part A is as shown in figure Fig. 2 below.

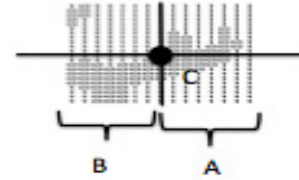


Fig. 2 Three Points Selection for the Part A

The details about the features are explained in Table IV. The features that are checked with \checkmark are used for part Core and A.

TABLE IV
FEATURES FROM THE TRIANGLE

No.	Feature Name	Features for classification	Description
1	a	X	Length from B(x,y) to C(x,y). Used for calculating ratio a/b, c/a and angles of triangle
2	b	X	Length from A(x,y) to C(x,y). Used for calculating ratio a/b, b/c and angles of triangle
3	c	X	Length from A(x,y) to B(x,y). Used for calculating ratio b/c, c/a and angles of triangle.
4	c/a	\checkmark	Ratio c to a
5	a/b	\checkmark	Ratio a to b
6	b/c	\checkmark	Ratio b to c
7	A	\checkmark	Angle of A
8	B	\checkmark	Angle of B
9	C	\checkmark	Angle of C

III. EXPERIMENTAL RESULT

There are two types of testing that have been conducted. The types of testing are Unsupervised Machine Learning and Supervised Machine Learning.

For the Unsupervised Machine Learning, seven algorithms have been used in order to classify the classification of Arabic calligraphy. The results are shown in Table V below.

TABLE V
RESULT FOR UNSUPERVISED MACHINE LEARNING

Test Image	Unsupervised Algorithms	Correct Group and Rank		Best Distance		Type
		Core+ A	A	Core+ A	Core	
	Echudian	1,3	1, 4	0.8012	0.3040	Diwani
	Manhattan	1, 3, 8	1,4	1.9340	0.5180	Diwani

Diwani	Chebyshev	1,4,6	1,5	0.5690	0.2440	Diwani
	Minkowski	1,4,7	1,3	1.2557	3.2526	Diwani
	Sorenson	1,3,8	1,4	0.0027	0.0014	Diwani
	Correlation	The correct Group and Calligraphy are not listed in top ten out of 711				
	Angular Separation	1,4	1,5	0.9999 96	0.9999 97	Diwani
	Ecludian	Not Listed	10		1.5525	Farisi
	Manhattan	Not Listed	Not Listed	-	-	-
C	Chebyshev	Not Listed	10	-	1.2590	Farisi
	Minkowski	1,8	5,8	0.0029 8	0.0104 7	Farisi
	Sorenson	6	10	0.6123	0.0069	Farisi
	Correlation	6	5,9	1.0397 E-5	6.6896 E-4	
	Angular Separation	6	6,9	0.3046	0.8468	Farisi

The result in Table VII shows that the features from *Core* with *part A* give good result based on the rank. However, the accuracy is better for *part A*.

For the Supervised Machine Learning, testing has been conducted by using Support Vector Machine (SVM). There are 1019 model images where 244 are *Diwani*, 187 *Farisi*, 198 *Nasakh*, 183 *riq'ah* and 207 *Thuluth*. The result of this testing is as shown in Table VII.

TABLE VI
CLASSIFICATION RESULT USING SVM

Image	Number of Image	Result
Diwani	29	69.697%
Farisi	25	12.9032%
Riqah	27	14.2857%
Thuluth	27	25.9259%

For the supervised machine learning using SVM, the classification for the class *Diwani* give very significant result. However, for the other classes the result still need to be improved. This is due to the characters in the class *Diwani* are uniquely different from other classes. Thus, the discriminant features need to be found out in order to improve the classification.

IV. FUTURE EXPANSION

For the future expansion, part B, C and D need to be implemented and tested using Supervised Machine Learning and Unsupervised Machine Learning in order to find discriminant features that can uniquely differentiate each calligraphy class well. The categorization of characters needs to be expanded to other characters and give some point/score to the successful classification with the same group of characters or to the nearest group. This may help to increase the successful classification.

ACKNOWLEDGMENT

The authors would like to thank Tuan Haji Hamdan Abdul Rahman for his consultation on Malay language, Jawi and also the software K-Jawi and Arabic fonts that are used as a standard of calligraphy in Malaysia. Also thanks to Ustaz Mohd Asrak Othman from National Association of

Calligraphy Malaysia, for the consultation and validation of the dataset used in this research. Many thanks also to the Pattern Recognition Group under Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia for providing an excellent space for research and facilities.

REFERENCES

- [1] J. Just, "The didactics of Palaeography," 2009.
- [2] K. Omar, M. S. Azmi, S. N. Syekh Abdullah, A. Abdullah, and M. F. Nasrudin, "Framework of Jawi Digital Paleography: A Preliminary Work," in *2nd International Conference on Mathematical Sciences*, 2010, p. 5.
- [3] I. B. Yosef, K. Kedem, I. Dinstein, M. Beit-arie, and E. Engel, "Classification of Hebrew Calligraphic Handwriting Styles: Preliminary Results," in *Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)*, 2004, no. ii, p. 299.
- [4] M. S. Azmi, K. Omar, M. Faidzul, N. Azah, K. Muda, and A. Abdullah, "Digital Paleography: Using the Digital Representation of Jawi Manuscripts to Support Paleographic Analysis," 2011, no. June.
- [5] and G. Z. Aiolfi, F., M. Simi, D. Sona, A. Sperduti, A. Starita, "SPI: A System for Palaeographic Inspection," vol. 4, pp. 34-38, 1999.
- [6] I. Moalla, a M. Alimi, F. Lebourgeois, and H. Emptoz, "Image Analysis for Palaeography Inspection," *Second International Conference on Document Image Analysis for Libraries (DIAL'06)*, pp. 303-311, 2006.
- [7] F. Aiolfi et al., "SPI: A System for Paleographic Inspections." 1999.
- [8] M. S. Azmi, K. Omar, M. Faidzul, N. Khadijah, and W. Mohd, "Arabic Calligraphy Identification for Digital Jawi Paleography using Triangle Blocks," *Science And Technology*, no. July, 2011.
- [9] A. Ciula, "Digital palaeography: using the digital representation of medieval script to support palaeographic analysis," vol. 1, no. Spring, pp. 1-31, 2005.
- [10] K. Omar, M. Sanusi, and A. Razak, "Batu Bersurat Terengganu: Perspektif Geometri Segitiga," in *Seminar Batu Bersurat Piagam Terengganu*, 2011.
- [11] M. S. Azmi, K. Omar, M. Faidzul, N. Khadijah, and W. Mohd, "Arabic Calligraphy Identification for Digital Jawi Paleography using Triangle Blocks," in *2011 International Conference on Electrical Engineering and Informatics*, 2011, no. July, pp. 1714-1718.
- [12] C. Anwar HR, *Dasar dasar Pokok Seni Kaligrafi*. Jombang: Lintas Media.
- [13] M. Haji Abdul Rahman, *Qaidah Seni Khat*, 1st ed. Kuala Lumpur: Pustaka Nikmah, 1993, pp. 1-96.
- [14] M. Haji Awang, *Panduan Menulis dan Kaedah Mengajar Khat Nasakh*, 1st ed. Selangor: Dewan Bahasa dan Pustaka, 2009, pp. 1-156.
- [15] N. Kustiawan, *Koleksi Aneka Seni Kaligrafi*, 1st ed. Surabaya: Pustaka Media, 2005, pp. 1-112.
- [16] M. Munir, *Mengenal Kaidah Kaligrafi Al Quran*, 1st ed. Semarang: Binawan-Semarang, 2004, pp. 1-271.

- [17] M. Munir, *Mengenal Kaidah Kaligrafi Al-Quran*, 1st ed. Semarang: Binawan, 2004, pp. 1-144.
- [18] D. Noerzaman, *Khalligrafi dan Tahsinul-Khat*, 4th ed. Bandung, Indonesia: Penerbit Pustaka, 2002, pp. 1-48.
- [19] M. A. Syaifulloh, *No Title*. Jombang: Lintas Media.
- [20] M. F. Nasrudin, "Pengecaman Tulisan Tangan Jawi Luar Talian Menggunakan Jelmaan Surih," Universiti Kebangsaan Malaysia.
- [21] K. Omar, "Pengecaman Tulisan Tangan Teks Jawi Menggunakan Pengkelas Multiaras," Universiti Putra Malaysia, 2000.