2011 International Conference on Electrical Engineering and Informatics
17-19 July 2011, Bandung, Indonesia

Arabic Calligraphy Identification for Digital Jawi Paleography using Triangle Blocks

Mohd Sanusi Azmi [#1], Khairuddin Omar [*2], Mohammad Faidzul Nasrudin[*3]Khadijah Wan Mohd Ghazali[#4], Azizi Abdullah[*5]

[1,4]*Faculty of Information Communication and Technology*
Universiti Teknikal Malaysia Melaka, Malaysia
[1]sanusi@utem.edu.my
[4]khadijah@utem.edu.my

[2,3,5] *Faculty of Information Science and Technology*
Universiti Kebangsaan Malaysia, Bangi, Malaysia
[2]ko@ftsm.ukm.my
[3]mfn@ftsm.ukm.my
[5]azizi@ftsm.ukm.my

Abstract—**Digital Jawi Paleography is a field of research that helps paleographers to identify authors, origin and date of Jawi manuscripts. This research is important because of the existence of a huge amount of Malay manuscripts with unidentified authors, origin and date. Most researches in the area are for Roman and Hebrew text, whereas researches for Jawi text have just begun recently. In this paper, a novel technique is proposed in order to identify types of Arabic calligraphy in Malay ancient manuscripts that were written in Jawi. The novel technique is based on the triangle blocks that were adapted from scalene triangle. Twenty-one features have been extracted from the triangle blocks.**

Keywords— **Paleography, Jawi, Features Extraction**

I. INTRODUCTION

Malay ancient manuscripts portray a lot of information implicitly or explicitly. This is proved by [1]. The work identified three unknown authors who wrote the book "Kitab Undang-Undang Melayu" in the middle of 18th century by identifying the type of calligraphy applied in the ancient book. This is also supported by [2]. From the research conducted in [3], four types of arabic calligraphy were found in "Hikayat Merong Mahawangsa". Calligraphy type identification applied to the manuscripts is a subset of paleography study. Research done in [4] for Roman ancient manuscripts and [5] for Hebrew ancient manuscripts used type of calligraphy in order to identify the origin and dating of manuscripts. The researches fall under digital paleography researches [4-6]. The purpose of the research is to assist paleographers in order to identify the origin and dating of manuscripts [3-5]. The first digital paleography was developed in University of Pisa, Italy by a team coordinated by Prof. Alessandro Sperduti and Prof. Antonina Starita [7,8]. In 2004, a preliminary research on paleography for Hebrew styles of writing in order to identify the origin and dating was conducted [5]. Two years later, [4] introduced the global approach in order to identify the style of calligraphy applied in the Roman ancient manuscripts. On the other hand, digital Jawi paleography is still at an early stage. The first Digital Jawi Paleography framework has been introduced in [6].

The motivation for Digital Jawi Paleography research is the huge number of Malay manuscripts that exists. In Malaysia alone the number is around 7789 and many others are located outside Malaysia such as in British Library, UK and Leiden University, Holland [6,9].

In order to do research in Digital Jawi Paleography, research conducted by [4,5,7] has been studied. In [7], statistical approach was used. Centroid and tangent values were used to cluster Roman alphabets collected from 37 manuscript books into Dendogram diagrams. Due to the lack of features in [7], [4] introduced global approach by using twelve Haralick's features. The haralick's features are also known as Grey-Level Co-occurrence Matrix (GLCM). For Hebrew digital paleography, [5] used local features from only five selected Hebrew alphabets. For feature extractions and testing, only two alphabets were chosen. No justification is stated on the selection of the few alphabets. [5] introduced a technique that is based on the space from selected alphabets. The test conducted in [5] is also insufficient because only 14 documents were chosen with only twenty Aleph and Lamed chosen from each document.

Based on the huge number of manuscripts located in Malaysia, Malay Archipelago and other countries as well as potential amendments to the digital paleography, we propose a novel technique to identify the style of writing in Malay ancient manuscript in order to give significant input to the study of paleography. The technique is Triangular Block

Model. The relationship between triangular blocks with Arabic calligraphy and is written in [10]

II. PRE-PROCESSING

We suggest a new technique that is based on the Scalene Triangle. The processes that take place in our proposed algorithms are:

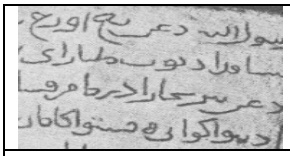
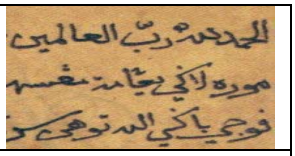
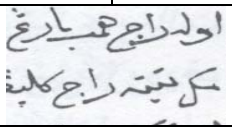
i. Data Collection,
ii. Pre-processing,
iii. Features extraction and Proposed method
iv. Testing

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

The data collected are from Malay ancients manuscripts and an inscribed stone. The manuscripts that were chosen are *Hikayat Merong Mahawangsa* and unknown manuscripts from National Library of Malaysia. The inscribed stone that was chosen is the inscribed stone of Terengganu. The examples of data that were used are in Table 1 below.

TABLE I
SAMPLE MALAY ANCIENT MANUSCRIPTS

	
The Inscribed Stone of Terengganu	Hikayat Merong Mahawangsa- page1
	
Hikayat Merong Mahawangsa- page30	

B. Pre-processing

Images from the stated Malay ancients manuscripts and inscribed stone are manually segmented into twenty alphabets in random representation. The segmented Jawi alphabets are in initial, middle, end or isolated state. This manual segmentation is done based on [5].

A default threshold value of 127 was given to the segmented alphabets/images except for some images with lower quality that required a threshold value of 180. The process converted RGB to binary color model as shown in Table 2.

TABLE II
THRESHOLD PROCESS

Original Image	Threshold value 127	Threshold value 180
		
		

After segmentation process was done, images were cleaned from noise pixels. As the quality of ancient manuscripts and inscribed stone normally get degraded with time, the images can be unclear and smeared [5,11]. After this process, the images are ready for features extraction process.

C. Features extraction and proposed method

The features extraction in this research is using Triangular Block Model that is adapted from Scalene Triangle. The processes are divided into three sub processes.

- Selection of three important points.
- Formation of type of blocks.
- Features Extraction based on Scalene Triangle.

21 features are extracted from Scalene Triangle including the type of blocks. These features will be used in order to identify the Arabic calligraphy that was applied in the manuscripts and inscribed stone. The details are discussed in the Proposed Method subtopic.

D. Testing.

Testing is conducted using forty segmented images of the Inscribed Stone of Terengganu, twenty images from the first page of *Hikayat Merong Mahawangsa*, twenty images from page 1 and 30. The result from the test data will be compared with result from standard calligraphies taken from k-Jawi software from Ministry Of Unity, Culture, Arts, Heritage and Tuan Haji Hamdan Abdul Rahman, the Malay language expert for the Dewan Bahasa dan Pustaka, Malaysia.

III. PROPOSED METHOD

A. Selection of three important points.

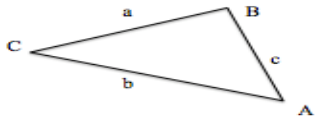
The three important points can be extracted from specific coordinates after images are freed from any unwanted noises. The three points selected are based on black pixels from the image. The selection of three points are chosen and labeled based on Table 3 below.

TABLE III
LOCATION OF THREE IMPORTANT POINTS

Point	Location of pixel	Label in Triangle	Line connected from points	Angle
Point 1	The first black pixel on the right image	A	b and c	A
Point 2	The first black pixel on the left image	C	a and b	C
Point 3	Centroid of image	B	a and c	B

The selected points, labels and lines connected to A, B and C and sample output is explained in Table 4 below.

TABLE IV
COORDINATE OF THREE POINTS.

Triangle	Image	Point
 *This triangle does not represent image that represented in column image*		Point 1 x : 21 y : 17 Point 2 x : 3 y : 27 Point 3 x : 13 y : 23

B. Formation of the type of Triangle Blocks

From research that has been conducted on three selected points, it is found that, if A is fixed to the right and C to the left whereas B is always in between A and C, it is possible that the triangle geometry models can be formed in six blocks. Fig. 1 below shows the possible triangle blocks.

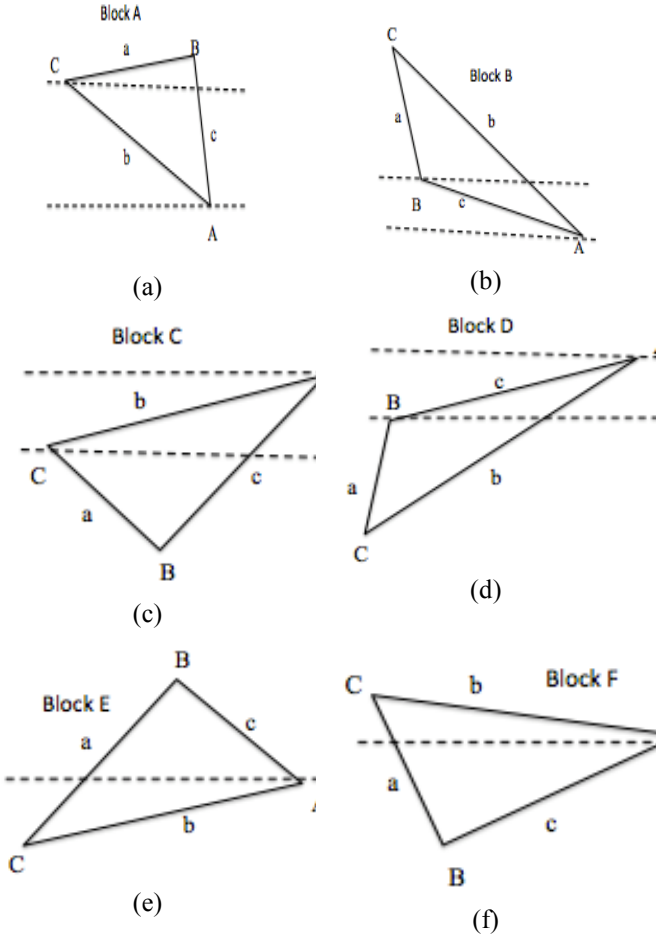


Fig. 1 Triangle Block Models

The triangle types formed from Fig. 1 above are based on the y-coordinate for points A, B and C as shown in Table 5.

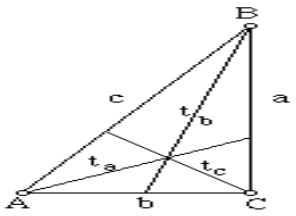
TABLE II
POSITION OF COORDINATE Y

Shape	Coordinate y
A	$yA \geq yC \geq yB$
B	$yA \geq yB \geq yC$
C	$yA \leq yC \leq yB$
D	$yA \leq yB \leq yC$
E	$yA \geq yB \leq yC$
F	$yA \geq yC \leq yB$

C. Features Extraction

After the six triangle blocks are formed, the next process is features extraction from the triangle blocks that are based on the three selected points on the image. There are twenty-one features extracted including the proposed triangle block. Table 6 shows features extracted from triangle blocks that are proposed in Fig. 1.

TABLE IIII
FEATURES FROM SCALENE TRIANGLE

Number	Feature Name	Description
1	Triangle Blocks	Block A, B, C, D, E and F
2	a	Length from B(x,y) to C(x,y)
3	b	Length from A(x,y) to C(x,y)
4	c	Length from A(x,y) to B(x,y)
5	Ratio c/a	Ratio c to a
6	A	Angle of A
7	B	Angle of B
8	C	Angle of C
9	Angle of A to B	Magnitude of A
10	Area	Area
11	Angle of Bisector side a (ta)	
12	Angle of Bisector side b (tb)	
13	Angle of Bisector side c (tc)	
14	Median of side a	*ma*
15	Median of side b	*mb*
16	Median of side c	*mc*
17	Altitude of side a	*ha*
18	Altitude of side b	*hb*
19	Altitude of side c	*hc*
20	Circumscribed Circle Radius	*R*

21	Inscribed Circle Radius	*r*

IV. Experimental Result

In the testing phase, segmented images selected from the inscribed stone and Malay ancient manuscripts will be identified.

The segmented images will be tested with the proposed method that has been developed. Next, result generated from the proposed method will be evaluated with the existing result from five types of calligraphy that are taken from k-Jawi software from Ministry Of Unity, Culture, Arts, Heritage and Tuan Haji Hamdan Abdul Rahman the Malay language expert for the Dewan Bahasa dan Pustaka, Malaysia.

Table 7 shows the result of the number of occurrences of every type of Arabic calligraphy in the inscribed stone of Terengganu. From forty segmented images that were tested, eighteen of them do not comply with standard Arabic calligraphy and the rest show that Arabic calligraphies exist in the inscribed stone. Block D is the highest occurrence in the inscribed stone, based on the result in Table 8.

TABLE IVI
NUMBER OF ARABIC CALLIGRAPHY IN FORTY SEGMENTED IMAGES IN THE INSCRIBED STONE OF TERENGGANU

Calligraphy	No. of occurrence
Diwani	18
Farisi	12
Nasakh	20
Riq'ah	16
Thuluth	16
Not comply	18

TABLE VI
NUMBER OF TRIANGLE BLOCKS IN INSCRIBED STONE OF TERENGGANU

Triangle Block	Number of occurrence
A	4
B	6
C	4
D	20
E	2
F	4

Table 9 and 10 below show result from the tests conducted to the two ancient Malay manuscripts. Twenty images each were extracted from page of from *Hikayat Merong Mahawangsa* on page 1 and 30.

Number of non-compliance from manuscript is lower compared to the inscribed stone.

TABLE IX
NUMBER OF ARABIC CALLIGRAPHY IN TESTED MALAY ANCIENT MANUSCRIPTS

Calligraphy	No. of occurence	Hikayat Merong Mahawangsa page1	Hikayat Merong Mahawangsa Page 30
Diwani	32	6	12
Farisi	25	8	8
Nasakh	31	8	9
Riq'ah	28	8	8
Thuluth	25	8	9
Not listed	13	6	5

TABLE VI
NUMBER OF PROPOSED TRIANGLE BLOCKS IN TESTED MANUSCRIPTS

Triangle Block	Hikayat Merong Mahawangsa page1	Hikayat Merong Mahawangsa Page 30	Total
A	4	2	7
B	5	1	7
C	1	1	4
D	8	15	38
E	2	1	3
F	0	0	1

V. Future Expansion

In this paper, the topic under study is the triangle blocks. The blocks extracted from the features extraction will be compared with triangle blocks extracted from the dataset based on the K-Jawi software. For the time being, only seven out of the twenty-one features were used to form the triangle blocks. This is because the novel technique is currently still under research in Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia. The next fourteen features including the current seven features will be used in the classification process. Based on the triangle blocks, the accuracy of type of Arabic calligraphy can only be determined by analysing extracted features by adapting scalene triangle. This research will be expanded by using the Perception Based Model that can exploit scalene triangle to be tested with Jawi ancient manuscript.

VI. Acknowledgment

Thank you to Tuan Haji Hamdan Abdul Rahman for his consultation on Malay language, Jawi and also software K-Jawi and Arabic fonts that is used as a standard of calligraphy in Malaysia. Not to forget 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] M.J. Abd. Rahman, *Teks undang-undang melayu pertengahan abad kelapan belas*, Kuala Lumpur: Dewan Bahasa dan Pustaka, 1994.

[2] M.S. Azmi, K. Omar, and A. Abdullah, "Perekayasaan Histogram Orientasi Kecerunan Mengesan Erotan dan Pencongan manuskript Merong Mahawangsa," *Jurnal Teknologi Maklumat & Multimedia*, vol. 2, 2005, pp. 63-79.

[3] K. Omar, M.S. Azmi, S.N. Syeikh Abdullah, A. Abdullah, and M.F. Nasrudin, "Framework of Jawi Digital Paleography: A Preliminar Work," *2nd International Conference on Mathematical Sciences*, Kuala Lumpur: Universit Kebangsaan Malaysia, 2010, p. 5.

[4] 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)*, 2006, pp. 303-311.

[5] I.B. Yosef, K. Kedem, I. Dinstein, M. Beit-arie, and E. Engel, "Classification of Hebrew Calligraphic Handwriting Styles : Preliminary Results," *Analysis*, 2004.

[6] K. Omar, M.S. Azmi, S.N. Syeikh Abdullah, M.F. Nasrudin, and A. Abdullah, "Kerangka Paleografi Jawi Digital : Satu Cadangan Awal," *Seminar*, 2010, pp. 1-14.

[7] and G.Z. Aiolli, F., M. Simi, D. Sona, A. Sperduti, A. Starita, "SPI: A System for Palaeographic Inspection," vol. 4, 1999, pp. 34-38.

[8] A. Ciula, "Digital palaeography : using the digital representation of medieval script to support palaeographic analysis," vol. 1, 2005, pp. 1-31.

[9] M.H. Rifin and A.N. Zainab, "Creating a Digital Library to Handle Malay Manuscripts Using Greenstone," *Image (Rochester, N.Y.)*, 2007, pp. 223-231.

[10] K. Omar, M. Sanusi, and A. Razak, "Batu Bersurat Terenganu : Perspektif Geometri Segitiga," *Seminar Batu Bersurat Piagam Terengganu*, Kuala Terengganu, Malaysia: Lembaga Muzium Negeri Terengganu, 2011.

[11] S.R. Yahya, S.N.H.S. Abdullah, K. Omar, M.S. Zakaria, and C.Y. Liong, "Review on image enhancement methods of old manuscript with the damaged background," *2009 International Conference on Electrical Engineering and Informatics*, Aug. 2009, pp. 62-67.