



Methodology for Multi-Stage Document Segmentation in Mushaf Al-Quran using Dominant Foreground: A Preliminary Work

Amirul Ramzani Radzid^{1*}, Mohd Sanusi Azmi¹, Intan Ermahani A. Jalil¹, Zahriah Othman¹, Azrina Tahir¹, Nur Atikah Arbain¹

¹Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia.

*amirulramzani@gmail.com

ABSTRACT

The vast number of images of manuscript with their heterogeneity contribute to big data research in the area. Manuscripts like the Mushaf of Al-Quran include ornaments and diacritics together with their main text. The shape and structure of the ornaments and diacritics in their different variety can show the origin of the manuscript. However, in digital image processing, ornaments are considered as foreground image and contribute noises for feature extraction and classification. On the other hand, the diacritics that exist in Mushaf Al-Quran cannot be removed because it is vital to enable readers to read it correctly. Thus, the ornaments have to be excluded whereas the diacritics must be remained. The process of removing ornaments and segmenting line by line requires multiphase segmentation based on physical dominant of the foreground image. There are many researches for segmentation especially for Latin and Arabic based handwritten documents. Unfortunately, the existing researches do not involve complicated structures and sensitive documents such as the Mushaf Al-Quran. In this research, segmentation of ornaments and text lines will be conducted. Novel multiphase segmentation using dominant foreground is proposed to solve the problems aforementioned. The proposed method focuses on segmenting through the four phases: decoration and text, text line, verse and sub-word. The algorithm is conjoined with multi-threading for parallel processing of big data. Through these phases, the heterogeneous ornaments will be excluded by identifying the ornaments structure and traversing the connected foreground pixels of the ornaments. For the text segmentation, neighbour of diacritics will be populated and computed based on the normality of existence. After the segmentation is performed, the evaluation of the segmentation output will be evaluated using expert judgment from Islamic scholar. Also, the segmented images will be evaluated using supervised machine learning in order to support the result and make the result comparable.

Key words: Line Segmentation; Multiphase Segmentation; Ornaments Extraction; Overlapped Segmentation; Text Segmentation.

1. INTRODUCTION

Document processing involve segmentation during pre-processing phase. This is a crucial step for Optical Character Recognition (OCR) and keyword spotting [1]. The problem encountered during text segmentation is the overlap-

ping in the between of two text line. Based on previous research, L.B. Melhem stated the limitation on their research finding was segmenting the overlap on the text line [2]. This overlapping occurs cause by interfering the diacritical marks or the stroke of the Arabic word. This issue is the problem that researcher going to solves in their future works.

Other than that, previous research exhibits inaccurate result when dealing with the large-scale dataset of page segmentation that has a degree of variability. In the previous studies [3][4][5][6][7][8] show inaccurate segmenting for multiform decoration frame, text line and verse on Mushaf Al-Quran text. The result from the previous study shows that the previous method cannot solve overlapping that cause by interfering of diacritical marks or stroke of the Arabic word.

Besides that, in the previous research [9][10][11][12][13][14][15] tends to mistakenly consider the unnecessary text (meaning of ayah) as part of ayah Al-Quran. It also mistakenly consider the necessary object (tashkil/end of verse) as part of decoration that leading to the incorrect interpretation of Al-Quran. The diacritical marks ownership are challenging process because misplaced or missing can change the meaning of ayah [2].

Therefore, the general manuscripts analysis method is not suitable to perform on Mushaf Al-Quran. This situation occurs regarding on non-uniform decoration heterogeneity, diacritic and tashkil, unnecessary text (meaning of ayah) [2]. To carry out document analysis and recognition for document Mushaf is the challenging task because Mushaf Al-Quran is not the same as ordinary document information varies on the different type of printed version.

2. RELATED WORK

There is a large number of historical manuscripts have been digitized and made available to the public [9]. It give an interest for research scholars to carry out to studies layout analysis more efficiently and in greater depth [16]. Table 1 show several document analysis that provided to produced groundtruth data on historical documents. Besides that, study on page segmentation has been done by Kai Chen et al. using DIVADIA dataset of historical document that classified page into several component as either periphery, background, text block, or decoration [9][3][10][11][12][13][14][15]. Apart from that, Table 2 show a competitions project on document analysis organized by International Conference of Document Analysis and Recognition (ICDAR) for robust reading competitions.

Table 1 and Table 2 show that researcher more focusing on historical manuscripts or scenery image. Unfortunately, there is a lack of research conducts in analysis the Mushaf Al-Quran document compared to others viz. historical (medieval documents) and record. Therefore this study will concentrate on the page segmentation of Mushaf Al-Quran text as a pillar of research contribution.

In previous study [2], a method has been proposed for page segmentation in removing illumination on Mushaf Al-Quran is using Binary Representation. Then, improved has been made [27]. However, the problem with previous study was inaccurate segmenting for multiform decoration frame as shown in Table 3.

Table 1: Analysis of Historical Documents








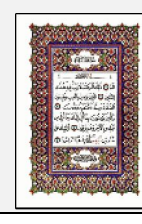
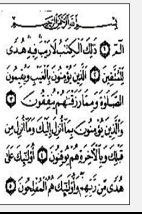
Research	Description
[17]	This study propose a new dataset and a ground-truthing methodology for layout analysis of historical documents with complex layouts. The dataset used in this study is DIVADIA. This dataset contains 120 pages from three historical document image collections.
[4]	This study proposed a web-based system to help users produce ground truth data for document images. The dataset used in this study is a degraded historical document images.
[5]	This study proposed a new XML-based page image representation framework that Records information on image characteristics (image borders, geometric distortions and corresponding corrections, binarisation etc.) in addition to layout structure and page content. The dataset used in this study is a public contemporary and historical ground-truthed datasets and in the ICDAR Page Segmentation competition series.
[6]	This study proposed a tools for visualizing and creating groundtruth and metadata. It reads and stores groundtruth metadata in XML format.

Table 2: Dataset on ICDAR

Specific application domains	Details	Datasets
2017 COCO-Text	Dataset consist of scene text detection and recognition, based on the largest scene text dataset currently available, based on real (as opposed to synthetic) scene imagery. Task : Text Localization, Cropped Word Recognition and End-To-End Recognition	COCO-Text dataset [7]
2017 De-TEXT	Dataset consist of biomedical figures. They propose semantic interpretation of biomedical figure mining in order to mining information from figures. Task: Text Localization (Text Detection), Cropped Text Blocks Recognition, End-to-End Recognition.	DeTEXT dataset [8]
2017 DOST	Dataset consist of omnidirectional video consists of video mode (Task: Localisation, End-to-end) and still image mode (Task: Localisation, Cropped word recognition, End-to-end).	DOST dataset [18]
2017 FSNS	Dataset consist of French Street Name Signs (FSNS). Task: end-to-end recognition on the Google FSNS dataset.	FSNS dataset [19]
2017 MLT	Dataset consist of multi-lingual scene text detection and script identification. Task: Multi-script text detection, Cropped Word Script identification, Joint text detection and script identification	MLT dataset [20]
2017 IEHHR	Dataset consist of historical handwritten records in order to extract information. Task: Extract information from the Historical Handwritten records.	IEHHR dataset [21]
2011-2015 Born-Digital Images	Dataset consist of digital images of electronic documents (Web and email) that consist of embed textual information. Task: Text Localization, Text Segmentation, Word Recognition, End-to-End.	Born-Digital Images dataset [22] [23]

2013-2015 Focused Scene Text	Dataset consist of real scenes in order to reading of text. Task: Text Localization, Text Segmentation, Word Recognition, End-to-End.	Focused Scene Text dataset [24] [23]
2013-2015 Text in Videos	Dataset consist of video sequences in order to localize and recognize text in the depicted scene. Task: Text Localisation, End to End.	Text in Videos dataset [25] [23]
2015 Incidental Scene Text	Dataset consist of real scene images in order to read the text. Task: Text Localization, Word Recognition, End-to-End.	Incidental Scene Text dataset [26]

Table 3: Text Extraction of Mushaf Al-Quran Pages

Source	Input Image	Result of Binary Representation [2]
Image of Al-Quran Al-Karim from Mawarsoft Digital Furqan 1.0 (Page 2)		Cannot be processed (Execution Error)
Image of Al-Quran Al-Karim from Mushaf Al-Madinah Quran Majeed (Page 1)		Cannot be processed (Execution Error)
Image of Al-Quran Al-Karim from KSU - Electronic Mosshaf (Page 1)		Cannot be processed (Execution Error)
Image of Al-Quran Al-Karim from Mushaf Al-Madinah Quran Majeed (Page 3)		
Image of Al-Quran Al-Karim from Mawarsoft Digital Furqan 1.0 (Page 4)		
Image of Al-Quran Al-Karim from Uthmani Script Mushaf (Page 2)		

In previous study [28], a method has been proposed for text line segmentation on Mushaf Al-Quran is using Binary Representation. However, the problem with previous study was inaccurate segmenting for text line as shown in Table 4 to Table 7.

Table 1: Result of Line Segmentation of Mushaf Al-Quran Rasm Uthmani Publish by Company S Abdul Majeed (page 6; row 11-13)

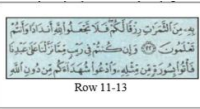
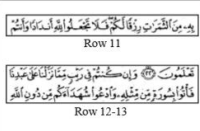
Input	
Result of Binary Representation [27] [28]	

Table 2: Result Of Line Segmentation on Muhaf Al-Quran from Mushaf Al-Madinah Quran Majeed (page 3; row 3-5)

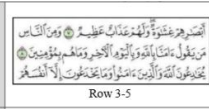
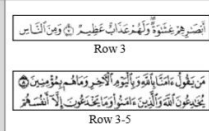
Input	
Result of Binary Representation [27] [28]	

Table 3: Result of Line Segmentation on Mushaf Al-Quran from Mushaf Al-Madinah Quran Majeed (page 3; row 6-8)

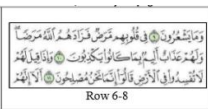
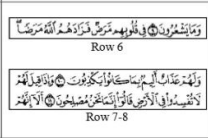
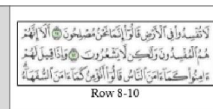
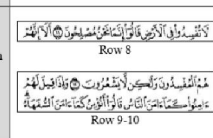
Input	
Result of Binary Representation [27] [28]	

Table 4: Result of Line Segmentation on Mushaf Al-Quran from Mushaf Al-Madinah Quran Majeed (page 3; row 8-10)

Input	
Result of Binary Representation [27] [28]	

Unfortunately, based on study there is a lack of research conducts in analysis the Mushaf Al-Quran document compared to others viz. historical (medieval documents) and record. Table 1 and table 2 show that researcher more focusing on historical manuscripts or scenery image. The general manuscripts analysis method on table 1 and table 2 have been investigated and not suitable to perform on Mushaf Al-Quran regarding on non-uniform decoration heterogeneity, diacritic and tashkil, unnecessary text (meaning of ayah) [2]. To carry out document analysis and recognition for document Mushaf is the challenging task because Mushaf Al-Quran is not the same as ordinary document information vary on different type of printed version. Besides that, prior research tends to mistakenly consider the unnecessary text (meaning of ayah) as part of ayah Al-Quran and also mistakenly consider the necessary object (tashkil / end of verse) as part of decoration that leading to the incorrect interpretation of Al-Quran. The diacritical marks ownership are challenging process because misplaced or missing can change the meaning of ayah [2].

Recent study related to foreground segmentation has been done by several researcher [29][30][31][32]. In 2018, a study by W. Yu *et al.* proposed a synthetic superpixel grouping mechanism to group the remainder SLIC superpixels into foreground or background until the whole superpixels are completely grouped [32]. Besides that, K. Chen *et al.* was proposed page segmentation for historical document images based on superpixel classification with

unsupervised feature learning [10]. Despite an outstanding result shown by recent research on its application, however regarding on Mushaf Al-Quran there will be lacking in accuracy. This is because superpixel cannot solve the problem with the diacritical marks ownership. Superpixel method will differentiate diacritical marks with its character. Result provided by [10] shown that the accuracy in pixel-labelling method (n=100k) much better than superpixel method (SLIC), however lack in runtime per image (T-min). The result from this study are shown as in Figure 1. Therefore, a novel dominant foreground method are proposed to solve the problems aforementioned.



Figure 1: Result from using superpixel method (image from K. Chen *et al.* 2016).

There is a recent study on Arabic text segmentation using area Voronoi diagrams by J.Ramdan *et al.* in 2016 [33]. In this study, segmentation is carried out by choosing appropriate sites bordering the Voronoi. The study address the issue where Voronoi Edges are effectively segmented fully neighbours however, it does not effectively segment partly neighbour. The result from this study indicated the diacritical marks will be separated from the character. The result from this study are shown as in Figure 2. Therefore, a novel dominant foreground method are proposed to solve the problems aforementioned.

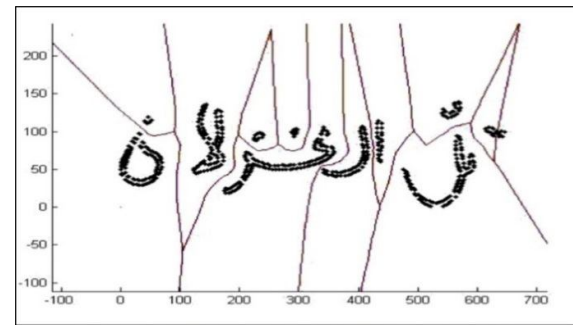


Figure 2: Result using Area-Voronoi Diagram (image from J.Ramdan *et al.* in 2016).

Besides that, the page segmentation have been investigated in the previous work [27][2][28] on Mushaf Al-Quran but producing inaccurate result when dealing with large-scale dataset that has degree of variability. Thus, this study proposed a novel method to tackle the issues that has been addressed. This proposed segmentation techniques has four phases which is (a) decoration and text, (b) text line, (c) verse and (d) sub-word on Mushaf Al-Quran. In order to

cope with speed, this algorithm conjointly with multithread for parallel processing on a big data. Fork and Join are implemented on the algorithm where study show a good result on performance in order to speed up for most program [34]. The multithread are made dynamically based on number of page and number of text line. Preprocessing for every phases start with converting image into binary images by using Otsu thresholding [35]. In binary images, the foreground image is known as '0' while background image is known as '1' [36].

3. METHODOLOGY

There are five phase for document segmentation which is (a) Theoretical study; (b) Data collection; (c) Page segmentation; (d) Feature extraction; and (e) Evaluation.

3.1. Theoretical study

At this phase, theories and approaches regarding to page segmentation are examined to investigate the comparisons of the current and suitable approaches, strengths, weaknesses and scopes of the page segmentation in mushaf Al-Quran text. This investigation is important to determine the gap and solutions for the research problems. Thus, a systematic method has been prepared in selecting suitable approaches and frameworks that are capable in providing valid answers to the research questions.

3.2. Data collection

Searching and selecting standard datasets. Investigate dataset issues. Develop local dataset to suit this research.

3.3. Page segmentation

This study is focusing on page segmenting using Dominant Foreground and conjointly with Fork and Join in order to speed up processing time for big data of mushaf Al-Quran with high accuracy. Multiphase in the research refer to different stage or phase during page segmentation process. This phase involves the four stages which are:

- 1) Decoration and text segmentation stage:
This stage is to identify multiform decoration frame on mushaf Al-Quran pages and extract only text of mushaf Al-Quran.
- 2) Text line segmentation stage:
This stage is to segment the text line on mushaf Al-Quran text to solve overlapping problem.
- 3) Verse segmentation stage:
This stage is to segment the verse on mushaf Al-Quran text by identifying the object end of verse (tashkil) correctly. Different mushaf have different pattern and shape of object.
- 4) Sub-word segmentation stage:
This stage is to segment sub-word on mushaf Al-Quran text by identifying the unoccupied space between word. It is challenging to identify unoccupied space for mushaf Al-Quran text because of the diacritical mark existence in arabic text of mushaf Al-Quran. This diacritical mark may cause misleading or erroneous to the sub-word segmentation result.

Segmentation phase are based on a novel dominant foreground method. Foreground in this study refer to detect the objects that do not belong to the background by comparing the current observation with previous references [31]. Foreground segmentation are widely used in many field such as static image and video sequence [29][30][31][32]. Study by W. Yu *et al.* used supervised image segmentation algorithm to the extract foreground [32]. A synthetic superpixel grouping mechanism is proposed to group the remainder SLIC superpixels into foreground or background until the whole superpixels are completely grouped. Unfortunately, this method cannot be applied to the page of Mushaf Al-Quran domain. This is because Arabic text on Mushaf Al-Quran contain important object which is diacritical marks that are placed with the arabic character but not connect to the character itself. By using the superpixel grouping, it will miscategorise the diacritical marks from the character itself. Dislocated diacritical marks will change the meaning of ayah. In order to segment the Mushaf Al-Quran text in precisely, algorithm must determine and categorize each of every single pixel. The pixels flow constantly spread heading to the dominant foreground. Dominant foreground refer to character or ornaments that has been determined by cluster point. The illustration of proposed method are shown in Figure 3. Red colour indicate the pixels (diacritical marks) flow constantly spread heading to the dominant foreground (character). Therefore, as a result of the proposed method segmentation multiphase of Mushaf Al-Quran will be accurate.

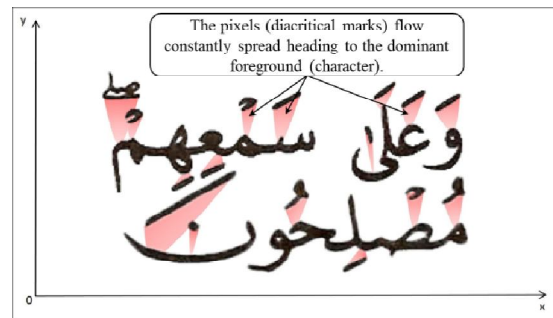


Figure 3: Illustration of dominant foreground method.

On the other hand, the algorithm to determine the single object is by identify its neighbouring pixel. The concept are illustrated as in Figure 4. As a result of pre-processing stage (Otsu thresholding) where the image will convert into binary form, the value of "0" indicated as foreground whereas the value "1" indicated as background per pixels. Afterwards, in order to cluster the neighbouring pixels around the pixel point of selection are shown as Figure 4. The outer point of pixel from selecting neighbour will not be cluster for example as in Figure 4.

3.4. Feature extraction

After the segmentation process completed, the feature extraction process will be performed to extract features using triangle geometry method. Triangle geometry features are proposed by M.S. Azmi [37] and improved by N.A. Arbain [38]. Study has been conducted that the triangle geometry features shows a better result for digit recognition as well as text recognition [36]. In this study, triangle geometry features apply dynamic multi-zoning divisions to solve the

issues on big size of features in order to tackle big data that causes the time taken is longer when processing data.

3.5. Evaluation

The proposed method will be evaluated in terms of accuracy and time in the Unsupervised Machine Learning (UML) environment. The classification process are UML with ranking measures that are widely used in information retrieval. The UML algorithm used is the Euclidean Distance Method (EDM) whereas the information retrieval measures used for the classification are Majority Voting (MV) and Mean Average Precision (MAP). Recent study by M.S Azmi [39] prove that the features from Triangle Model techniques with the UML and MAP techniques give better result compared to SML with Multi-layer Perceptrons and UML with MV.

1	2	3	4	5
1	1	1	1	1
6	7 (x-1,y+1)	8 (x,y+1)	9 (x+1,y+1)	10
1	1	1	0	1
11	12 (x-1,y)	13 (x,y)	14 (x+1,y)	15
1	1	0	1	0
16	17 (x-1,y-1)	18 (x,y-1)	19 (x+1,y-1)	20
1	1	1	1	1
21	22	23	24	25
1	1	1	1	0

Figure 4: Illustration of neighbouring pixels.

4. RESULT

This research paper are focusing on methodology of document segmentation of Mushaf Al-Quran. Hence, our preliminary result is on ornament removal the different shape of decoration on Mushaf Al-Quran page and text line segmentation. Table 8-12 shown a comparison result between exiting method and proposed method.

Table 8: Result Comparison for Ornament Removal in Mushaf Al-Quran

Source	Input Image	Result of Binary Representation [2]	Result of Proposed Method
Image of Al-Quran Al-Karim from Mawarsoft Digital Furqan 1.0 (Page 2)			

Image of Al-Quran Al-Karim from Mushaf Al-Madinah Quran Majeed (Page 1)			essed (Execution Error)
Image of Al-Quran Al-Karim from KSU Electronic Mosshaf (Page 1)			Cannot be processed (Execution Error)
Image of Al-Quran Al-Karim from Mushaf Al-Madinah Quran Majeed (Page 3)			
Image of Al-Quran Al-Karim from Mawarsoft Digital Furqan 1.0 (Page 4)			
Image of Al-Quran Al-Karim from Uthmani Script Mushaf (Page 2)			

Table 9: Result of Text Line Segmentation on Mushaf Al-Quran Rasm Uthmani Publish by Company S Abdul Majeed (Page 6)

Input	<p>يٰۤاَيُّهَا الَّذِيْنَ اٰمَنُوْا اٰتُوا زَكٰتَ رِزْقِكُمْ فَلَا تَجْعَلُوْا لِقٰوَالِكُمْ اَعْدٰۤا وَاَنْتُمْ تَعْلَمُوْنَ ﴿١١﴾ وَاِنْ كُنْتُمْ فِيْ رَيْبٍ مِّمَّا نَزَّلْنَا عَلٰى عَبْدِنَا فَأْتُوْا بِسُوْرَةٍ مِّنْ مِّثْلِهٖ ۗ وَاَدْعُوْا شُهَدٰۤاءَكُمْ مِّنْ دُوْنِ اللّٰهِ</p> <p>Row 11-13</p>
Result of Binary Representation (Melhem, 2015) [2][28]	<p>يٰۤاَيُّهَا الَّذِيْنَ اٰتُوا زَكٰتَ رِزْقِكُمْ فَلَا تَجْعَلُوْا لِقٰوَالِكُمْ اَعْدٰۤا وَاَنْتُمْ تَعْلَمُوْنَ ﴿١١﴾ وَاِنْ كُنْتُمْ فِيْ رَيْبٍ مِّمَّا نَزَّلْنَا عَلٰى عَبْدِنَا فَأْتُوْا بِسُوْرَةٍ مِّنْ مِّثْلِهٖ ۗ وَاَدْعُوْا شُهَدٰۤاءَكُمْ مِّنْ دُوْنِ اللّٰهِ</p> <p>Row 11</p> <p>يٰۤاَيُّهَا الَّذِيْنَ اٰتُوا زَكٰتَ رِزْقِكُمْ فَلَا تَجْعَلُوْا لِقٰوَالِكُمْ اَعْدٰۤا وَاَنْتُمْ تَعْلَمُوْنَ ﴿١١﴾ وَاِنْ كُنْتُمْ فِيْ رَيْبٍ مِّمَّا نَزَّلْنَا عَلٰى عَبْدِنَا فَأْتُوْا بِسُوْرَةٍ مِّنْ مِّثْلِهٖ ۗ وَاَدْعُوْا شُهَدٰۤاءَكُمْ مِّنْ دُوْنِ اللّٰهِ</p> <p>Row 12-13</p>
Result of Proposed Method	<p>يٰۤاَيُّهَا الَّذِيْنَ اٰتُوا زَكٰتَ رِزْقِكُمْ فَلَا تَجْعَلُوْا لِقٰوَالِكُمْ اَعْدٰۤا وَاَنْتُمْ تَعْلَمُوْنَ ﴿١١﴾ وَاِنْ كُنْتُمْ فِيْ رَيْبٍ مِّمَّا نَزَّلْنَا عَلٰى عَبْدِنَا فَأْتُوْا بِسُوْرَةٍ مِّنْ مِّثْلِهٖ ۗ وَاَدْعُوْا شُهَدٰۤاءَكُمْ مِّنْ دُوْنِ اللّٰهِ</p> <p>Row 11</p> <p>يٰۤاَيُّهَا الَّذِيْنَ اٰتُوا زَكٰتَ رِزْقِكُمْ فَلَا تَجْعَلُوْا لِقٰوَالِكُمْ اَعْدٰۤا وَاَنْتُمْ تَعْلَمُوْنَ ﴿١١﴾ وَاِنْ كُنْتُمْ فِيْ رَيْبٍ مِّمَّا نَزَّلْنَا عَلٰى عَبْدِنَا فَأْتُوْا بِسُوْرَةٍ مِّنْ مِّثْلِهٖ ۗ وَاَدْعُوْا شُهَدٰۤاءَكُمْ مِّنْ دُوْنِ اللّٰهِ</p> <p>Row 12</p> <p>يٰۤاَيُّهَا الَّذِيْنَ اٰتُوا زَكٰتَ رِزْقِكُمْ فَلَا تَجْعَلُوْا لِقٰوَالِكُمْ اَعْدٰۤا وَاَنْتُمْ تَعْلَمُوْنَ ﴿١١﴾ وَاِنْ كُنْتُمْ فِيْ رَيْبٍ مِّمَّا نَزَّلْنَا عَلٰى عَبْدِنَا فَأْتُوْا بِسُوْرَةٍ مِّنْ مِّثْلِهٖ ۗ وَاَدْعُوْا شُهَدٰۤاءَكُمْ مِّنْ دُوْنِ اللّٰهِ</p> <p>Row 13</p>

	<p>وَمَا يَشْعُرُوْنَ ﴿١٠﴾ فِيْ قُلُوْبِهِمْ مَّرَضٌ فَزَادَهُمُ اللّٰهُ مَرَضًا وَلَهُمْ عَذَابٌ اَلِيْمٌ يُمَاطِكُوْنَ اَنْۢوَابَهُمْ وَيَكُوْبُوْنَ ﴿١١﴾ وَاِذَا قِيْلَ لَهُمْ لَا تُفْسِدُوْا فِي الْاَرْضِ قَالُوْا اِنَّمَا نَحْنُ مُصْلِحُوْنَ ﴿١٢﴾ اَلَا اِنَّهُمْ</p> <p>Row 6-8</p>
Result of Binary Representation (Melhem, 2015) [2][28]	<p>وَمَا يَشْعُرُوْنَ ﴿١٠﴾ فِيْ قُلُوْبِهِمْ مَّرَضٌ فَزَادَهُمُ اللّٰهُ مَرَضًا</p> <p>Row 6</p> <p>وَلَهُمْ عَذَابٌ اَلِيْمٌ يُمَاطِكُوْنَ اَنْۢوَابَهُمْ وَيَكُوْبُوْنَ ﴿١١﴾ وَاِذَا قِيْلَ لَهُمْ لَا تُفْسِدُوْا فِي الْاَرْضِ قَالُوْا اِنَّمَا نَحْنُ مُصْلِحُوْنَ ﴿١٢﴾ اَلَا اِنَّهُمْ</p> <p>Row 7-8</p>
Result of Proposed Method	<p>وَمَا يَشْعُرُوْنَ ﴿١٠﴾ فِيْ قُلُوْبِهِمْ مَّرَضٌ فَزَادَهُمُ اللّٰهُ مَرَضًا</p> <p>Row 6</p> <p>وَلَهُمْ عَذَابٌ اَلِيْمٌ يُمَاطِكُوْنَ اَنْۢوَابَهُمْ وَيَكُوْبُوْنَ ﴿١١﴾ وَاِذَا قِيْلَ لَهُمْ لَا تُفْسِدُوْا فِي الْاَرْضِ قَالُوْا اِنَّمَا نَحْنُ مُصْلِحُوْنَ ﴿١٢﴾ اَلَا اِنَّهُمْ</p> <p>Row 7</p> <p>وَمَا يَشْعُرُوْنَ ﴿١٠﴾ فِيْ قُلُوْبِهِمْ مَّرَضٌ فَزَادَهُمُ اللّٰهُ مَرَضًا</p> <p>Row 8</p>

Table 10: Result of Text Line Segmentation of Mushaf Al-Quran From Mushaf Al-Madinah Quran Majeed (Page 3)

Input	<p>اَبْصُرْ هِمْزٌ عَشْوَةٌ وَلَهُمْ عَذَابٌ عَظِيْمٌ ﴿١٠﴾ وَمِنَ النَّاسِ مَن يَقُوْلُ ءَاٰمَنَّا بِاللّٰهِ وَيَاْتُوْهُ الْاٰخِرُ وَمَا هُمْ بِمُؤْمِنِيْنَ ﴿١١﴾ يُخٰدِعُوْنَ اللّٰهَ وَالَّذِيْنَ ءَاٰمَنُوْا وَمَا يَخْدَعُوْنَ اِلَّا اَنْفُسَهُمْ</p> <p>Row 3-5</p>
Result of Binary Representation (Melhem, 2015) [2][28]	<p>اَبْصُرْ هِمْزٌ عَشْوَةٌ وَلَهُمْ عَذَابٌ عَظِيْمٌ ﴿١٠﴾ وَمِنَ النَّاسِ مَن يَقُوْلُ ءَاٰمَنَّا بِاللّٰهِ وَيَاْتُوْهُ الْاٰخِرُ وَمَا هُمْ بِمُؤْمِنِيْنَ ﴿١١﴾ يُخٰدِعُوْنَ اللّٰهَ وَالَّذِيْنَ ءَاٰمَنُوْا وَمَا يَخْدَعُوْنَ اِلَّا اَنْفُسَهُمْ</p> <p>Row 3</p> <p>اَبْصُرْ هِمْزٌ عَشْوَةٌ وَلَهُمْ عَذَابٌ عَظِيْمٌ ﴿١٠﴾ وَمِنَ النَّاسِ مَن يَقُوْلُ ءَاٰمَنَّا بِاللّٰهِ وَيَاْتُوْهُ الْاٰخِرُ وَمَا هُمْ بِمُؤْمِنِيْنَ ﴿١١﴾ يُخٰدِعُوْنَ اللّٰهَ وَالَّذِيْنَ ءَاٰمَنُوْا وَمَا يَخْدَعُوْنَ اِلَّا اَنْفُسَهُمْ</p> <p>Row 3-5</p>
Result of Proposed Method	<p>اَبْصُرْ هِمْزٌ عَشْوَةٌ وَلَهُمْ عَذَابٌ عَظِيْمٌ ﴿١٠﴾ وَمِنَ النَّاسِ مَن يَقُوْلُ ءَاٰمَنَّا بِاللّٰهِ وَيَاْتُوْهُ الْاٰخِرُ وَمَا هُمْ بِمُؤْمِنِيْنَ ﴿١١﴾ يُخٰدِعُوْنَ اللّٰهَ وَالَّذِيْنَ ءَاٰمَنُوْا وَمَا يَخْدَعُوْنَ اِلَّا اَنْفُسَهُمْ</p> <p>Row 3</p> <p>اَبْصُرْ هِمْزٌ عَشْوَةٌ وَلَهُمْ عَذَابٌ عَظِيْمٌ ﴿١٠﴾ وَمِنَ النَّاسِ مَن يَقُوْلُ ءَاٰمَنَّا بِاللّٰهِ وَيَاْتُوْهُ الْاٰخِرُ وَمَا هُمْ بِمُؤْمِنِيْنَ ﴿١١﴾ يُخٰدِعُوْنَ اللّٰهَ وَالَّذِيْنَ ءَاٰمَنُوْا وَمَا يَخْدَعُوْنَ اِلَّا اَنْفُسَهُمْ</p> <p>Row 4</p> <p>اَبْصُرْ هِمْزٌ عَشْوَةٌ وَلَهُمْ عَذَابٌ عَظِيْمٌ ﴿١٠﴾ وَمِنَ النَّاسِ مَن يَقُوْلُ ءَاٰمَنَّا بِاللّٰهِ وَيَاْتُوْهُ الْاٰخِرُ وَمَا هُمْ بِمُؤْمِنِيْنَ ﴿١١﴾ يُخٰدِعُوْنَ اللّٰهَ وَالَّذِيْنَ ءَاٰمَنُوْا وَمَا يَخْدَعُوْنَ اِلَّا اَنْفُسَهُمْ</p> <p>Row 5</p>

Table 11: Result of Text Line Segmentation of Mushaf Al-Quran From Mushaf Al-Madinah Quran Majeed (Page 3)

Input	
-------	--

Table 12: Result of Text Line Segmentation of Mushaf Al-Quran From Mushaf Al-Madinah Quran Majeed (Page 3)

Input	<p>لَا تُفْسِدُوْا فِي الْاَرْضِ قَالُوْا اِنَّمَا نَحْنُ مُصْلِحُوْنَ ﴿١٢﴾ اَلَا اِنَّهُمْ هُمُ الْمُفْسِدُوْنَ وَلٰكِن لَّا يَشْعُرُوْنَ ﴿١٣﴾ وَاِذَا قِيْلَ لَهُمْ ءَاٰمِنُوْا كَمَا ءَاٰمَنَ النَّاسُ قَالُوْا اَنْۢوَابُنَا كَمَا ءَاٰمَنَ السُّفَهَاءُ</p> <p>Row 8-10</p>
Result of Binary Representation (Melhem, 2015) [2][28]	<p>لَا تُفْسِدُوْا فِي الْاَرْضِ قَالُوْا اِنَّمَا نَحْنُ مُصْلِحُوْنَ ﴿١٢﴾ اَلَا اِنَّهُمْ هُمُ الْمُفْسِدُوْنَ وَلٰكِن لَّا يَشْعُرُوْنَ ﴿١٣﴾ وَاِذَا قِيْلَ لَهُمْ ءَاٰمِنُوْا كَمَا ءَاٰمَنَ النَّاسُ قَالُوْا اَنْۢوَابُنَا كَمَا ءَاٰمَنَ السُّفَهَاءُ</p> <p>Row 8</p> <p>لَا تُفْسِدُوْا فِي الْاَرْضِ قَالُوْا اِنَّمَا نَحْنُ مُصْلِحُوْنَ ﴿١٢﴾ اَلَا اِنَّهُمْ هُمُ الْمُفْسِدُوْنَ وَلٰكِن لَّا يَشْعُرُوْنَ ﴿١٣﴾ وَاِذَا قِيْلَ لَهُمْ ءَاٰمِنُوْا كَمَا ءَاٰمَنَ النَّاسُ قَالُوْا اَنْۢوَابُنَا كَمَا ءَاٰمَنَ السُّفَهَاءُ</p> <p>Row 9-10</p>
Result of Proposed Method	<p>لَا تُفْسِدُوْا فِي الْاَرْضِ قَالُوْا اِنَّمَا نَحْنُ مُصْلِحُوْنَ ﴿١٢﴾ اَلَا اِنَّهُمْ هُمُ الْمُفْسِدُوْنَ وَلٰكِن لَّا يَشْعُرُوْنَ ﴿١٣﴾ وَاِذَا قِيْلَ لَهُمْ ءَاٰمِنُوْا كَمَا ءَاٰمَنَ النَّاسُ قَالُوْا اَنْۢوَابُنَا كَمَا ءَاٰمَنَ السُّفَهَاءُ</p> <p>Row 8</p> <p>لَا تُفْسِدُوْا فِي الْاَرْضِ قَالُوْا اِنَّمَا نَحْنُ مُصْلِحُوْنَ ﴿١٢﴾ اَلَا اِنَّهُمْ هُمُ الْمُفْسِدُوْنَ وَلٰكِن لَّا يَشْعُرُوْنَ ﴿١٣﴾ وَاِذَا قِيْلَ لَهُمْ ءَاٰمِنُوْا كَمَا ءَاٰمَنَ النَّاسُ قَالُوْا اَنْۢوَابُنَا كَمَا ءَاٰمَنَ السُّفَهَاءُ</p> <p>Row 9</p> <p>لَا تُفْسِدُوْا فِي الْاَرْضِ قَالُوْا اِنَّمَا نَحْنُ مُصْلِحُوْنَ ﴿١٢﴾ اَلَا اِنَّهُمْ هُمُ الْمُفْسِدُوْنَ وَلٰكِن لَّا يَشْعُرُوْنَ ﴿١٣﴾ وَاِذَا قِيْلَ لَهُمْ ءَاٰمِنُوْا كَمَا ءَاٰمَنَ النَّاسُ قَالُوْا اَنْۢوَابُنَا كَمَا ءَاٰمَنَ السُّفَهَاءُ</p> <p>Row 10</p>

5. DISCUSSION

Here is the detail of page segmentation on Mushaf Al-Quran phase:

5.1. Decoration and text segmentation

This phase will identified multiform decoration frame on Mushaf Al-Quran pages and extract only text of Mushaf Al-Quran. In previous study [2], a method has been proposed for removing illumination on Mushaf Al-Quran by using Binary Representation. Then, improved has been made [27]. However, the problem with previous study was inaccurate segmenting for multiform decoration frame as shown in table 3. Time taken for previous method was slow. Hence, this research are focusing on segmenting for multiform decoration frame by using Dominant Foreground and conjointly with Fork and Join in order to speed up processing time for big data of Mushaf Al-Quran with high accuracy.

5.2. Text line segmentation

This phase will segmenting text line on Mushaf Al-Quran text. In previous study [28][2], a method for text line segmentation of Al-Quran pages using Binary Representation has been proposed. However, the problem with previous study is inaccurate segmenting for text line on Mushaf Al-Quran text. The result from previous study shows that the previous method can not solve overlapping that cause by interfering of diacritical marks or stroke of the Arabic word. Hence, this research focusing on segmenting for segmenting text line by using Dominant Foreground on Mushaf Al-Quran text to solve overlapping problem.

5.2. Verse segmentation

This phase will segmenting verse on Mushaf Al-Quran text. In order to segmenting verse, object end of verse (tashkil) must be identified correctly. Different Mushaf have different pattern and shape of object. Thus, this research focusing on identify multiform of object end of verse by using Dominant Foreground.

5.3. Sub-word segmentation

This phase will segmenting sub-word on Mushaf Al-Quran text. In order to segmenting sub-word, unoccupied space between words must be identified. The challenging task to identified unoccupied space for Mushaf Al-Quran text is existence of diacritical mark in arabic text of Mushaf Al-Quran. The existence of diacritical mark in arabic text of Mushaf Al-Quran can misleading or erroneous the result to identified unoccupied space in order to segmenting sub-word of Mushaf Al-Quran. Hence, this research focusing on segmenting sub-word on Mushaf Al-Quran text by using Dominant Foreground.

After the segmentation process completed, the feature extraction process will be performed to extract features using triangle geometry method. Triangle geometry features are proposed by M.S. Azmi [37] and improved by N.A. Arbain [38]. Study has been conducted that the triangle geometry features shows a better result for digit recognition as well as text recognition [36]. In this study, triangle geometry features apply dynamic multi-zoning divisions to solve the issues on big size of features in order to tackle big data that causes the time taken is longer when processing data.

Then, the proposed method will be evaluated in terms of accuracy and time in the Unsupervised Machine Learning (UML) environment. The classification process are UML

with ranking measures that are widely used in information retrieval. The UML algorithm used is the Euclidean Distance Method (EDM) whereas the information retrieval measures used for the classification are Majority Voting (MV) and Mean Average Precision (MAP). Study prove that the features from Triangle Model techniques with the UML and MAP techniques give better result compared to SML with Multi-layer Perceptrons and UML with MV [39].

6. CONCLUSION

In this paper, we present a segmentation method in Mushaf Al-Quran. There are 4 stage which is (A) Decoration and Text Segmentation on Mushaf Al-Quran; (B) Text Line Segmentation on Mushaf Al-Quran; (C) Verse Segmentation on Mushaf Al-Quran; and (D) Sub-word Segmentation on Mushaf Al-Quran. The result is for decoration segmentation are compared with Binary Representation technique that was proposed by L.B. Melhem [2] with the same dataset. Result show that proposed method more accurate than previous method.

The proposed method can be applied to document analysis on Mushaf Al-Quran. This could help researcher in order to investigate Mushaf Al-Quran authentication. This proposed method also can be used to study Mushaf Al-Quran layout and structure.

ACKNOWLEDGEMENT

The authors would like to express their appreciation to thank to the Ministry of Education for funding this study through the following grants: FRGS/1/2017/ICT02/FTMK-CACT/F00345. Gratitude is also due to Universiti Teknikal Malaysia Melaka and Faculty of Information Technology and Communication for providing excellent research facilities.

REFERENCES

1. C. L. Tan, "Text Line Segmentation for Handwritten Documents Using Constrained Seam Carving," *2014 14th Int. Conf. Front. Handwrit. Recognit. Text*, 2014.
2. L. N. B. Melhem, "Illumination Removal And Text Segmentation For Al-Quran Using Binary Representation," Thesis for Master, Universiti Teknikal Malaysia Melaka, 2015.
3. K. Chen, M. Seuret, H. Wei, M. Liwicki, J. Hennebert, and R. Ingold, "Ground truth model, tool, and dataset for layout analysis of historical documents," *IS&T/SPIE Electron. Imaging*, no. November, p. 940204, 2015.
4. O. Biller, A. Asi, K. Kedem, and I. Dinstein, "WebGT: An Interactive Web-Based System for Historical Document Ground Truth Generation," in *2013 12th International Conference on Document Analysis and Recognition*, 2013, pp. 305–308.
5. S. Pletschacher and A. Antonacopoulos, "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 257–260.

6. C. Ha Lee and T. Kanungo, “**The architecture of TrueViz: a groundTRUth/metadata editing and Visualizing ToolKit,**” *Pattern Recognit.*, vol. 36, no. 3, pp. 811–825, Mar. 2003.
7. A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, “**COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images,**” Jan. 2016.
8. X. C. Yin *et al.*, “**DeTEXT: A database for evaluating text extraction from biomedical literature figures,**” *PLoS One*, vol. 10, no. 5, 2015.
9. K. Chen, H. Wei, J. Hennebert, R. Ingold, and M. Liwicki, “**Page Segmentation for Historical Handwritten Document Images Using Color and Texture Features,**” in *2014 14th International Conference on Frontiers in Handwriting Recognition*, 2014, pp. 488–493.
10. K. Chen, C.-L. Liu, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold, “**Page Segmentation for Historical Document Images Based on Superpixel Classification with Unsupervised Feature Learning,**” in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 2016, pp. 299–304.
11. K. Chen, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold, “**Page segmentation of historical document images with convolutional autoencoders,**” in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, vol. 2015–Novem, no. August, pp. 1011–1015.
12. K. Chen, H. Wei, M. Liwicki, J. Hennebert, and R. Ingold, “**Robust text line segmentation for historical manuscript images using color and texture,**” *Proc. - Int. Conf. Pattern Recognit.*, pp. 2978–2983, 2014.
13. K. Chen, M. Seuret, M. Liwicki, J. Hennebert, C.-L. Liu, and R. Ingold, “**Page Segmentation for Historical Handwritten Document Images Using Conditional Random Fields,**” in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2016, pp. 90–95.
14. K. Chen, M. Seuret, J. Hennebert, and R. Ingold, “**Convolutional Neural Networks for Page Segmentation of Historical Document Images,**” in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 965–970.
15. H. Wei, K. Chen, R. Ingold, and M. Liwicki, “**Hybrid Feature Selection for Historical Document Layout Analysis,**” *Proc. Int. Conf. Front. Handwrit. Recognition, ICFHR*, vol. 2014–Decem, pp. 87–92, 2014.
16. Y. Yang, R. Pintus, E. Gobbetti, and H. Rushmeier, “**Automatic Single Page-Based Algorithms for Medieval Manuscript Analysis,**” *J. Comput. Cult. Herit.*, vol. 10, no. 2, pp. 1–22, Mar. 2017.
17. K. Chen, M. Seuret, H. Wei, M. Liwicki, J. Hennebert, and R. Ingold, “**Ground truth model, tool, and dataset for layout analysis of historical documents,**” in *SPIE 9402, Document Recognition and Retrieval XXII*, 2015, no. February, p. 940204.
18. M. Iwamura, T. Matsuda, N. Morimoto, H. Sato, Y. Ikeda, and K. Kise, “**Downtown Osaka Scene Text Dataset,**” vol. 9913, G. Hua and H. Jégou, Eds. Cham: Springer International Publishing, 2016, pp. 440–455.
19. R. Smith *et al.*, “**End-to-End Interpretation of the French Street Name Signs Dataset,**” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9913 LNCS, Springer, Cham, 2016, pp. 411–426.
20. J.-C. Burie *et al.*, “**Datasets - ICDAR2017 Competition on Multi-lingual scene text detection and script identification,**” *ICDAR 2017 Robust Reading competitions*, 2017. [Online]. Available: <http://rrc.cvc.uab.es/?ch=8&com=downloads>.
21. A. Fornes *et al.*, “**ICDAR2017 Competition on Information Extraction in Historical Handwritten Records,**” in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 1389–1394.
22. D. Karatzas, S. R. Mestre, J. Mas, F. Nourbakhsh, and P. P. Roy, “**ICDAR 2011 robust reading competition - Challenge 1: Reading text in born-digital images (web and email),**” *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 1485–1490, 2011.
23. D. Karatzas *et al.*, “**ICDAR 2013 Robust Reading Competition,**” in *2013 12th International Conference on Document Analysis and Recognition*, 2013, pp. 1484–1493. <https://doi.org/10.1109/ICDAR.2013.221>
24. A. Shahab, F. Shafait, and A. Dengel, “**ICDAR 2011 Robust Reading Competition Challenge 2: Reading Text in Scene Images,**” in *2011 International Conference on Document Analysis and Recognition*, 2011, pp. 1491–1496.
25. M. Iwamura, L. G. i Bigorda, D. Karatzas, and S. R. Mestre, “**Datasets - Text in Videos,**” *ICDAR 2017 Robust Reading competitions*, 2017. [Online]. Available: <http://rrc.cvc.uab.es/?ch=3&com=downloads>.
26. D. Karatzas *et al.*, “**Dataset - Incidental Scene Text,**” *ICDAR 2017 Robust Reading competitions*, 2017. [Online]. Available: <http://rrc.cvc.uab.es/?ch=4&com=downloads>.
27. A. R. Radzid, “**Removing AI-Quran Illumination,**” Thesis for Bachelor Degree, Universiti Teknikal Malaysia Melaka, 2016.
28. L. B. Melhem, M. S. Azmi, A. K. Muda, N. J. Bani-Melhim, and M. Alweshah, “**Text Line Segmentation of AI-Quran Pages Using Binary Representation,**” *Adv. Sci. Lett.*, vol. 23, pp. 11498–11502, 2017.
29. M. Camplani and L. Salgado, “**Background foreground segmentation with RGB-D Kinect data: An efficient combination of classifiers,**” *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 122–136, 2014.
30. R. Trabelsi, I. Jabri, F. Smach, and A. Bouallegue, “**Efficient and fast multi-modal foreground-background segmentation using RGBD data,**” *Pattern Recognit. Lett.*, vol. 97, pp. 13–20, 2017.
31. G. Moyà-Alcover, A. Elgammal, A. Jaume-i-Capó,

- and J. Varona, “**Modeling depth for nonparametric foreground segmentation using RGBD devices,**” *Pattern Recognit. Lett.*, vol. 96, pp. 76–85, 2017.
32. W. Yu, Z. Hou, P. Wang, X. Qin, L. Wang, and H. Li, “**Weakly Supervised Foreground Segmentation Based on Superpixel Grouping,**” *IEEE Access*, vol. 6, no. Cccv, pp. 12269–12279, 2018.
33. J. Ramdan, K. Omar, and M. Fyzul, “**Segmentation of Arabic words using area Voronoi diagrams and neighbours graph,**” *Int. J. Soft Comput.*, vol. 11, no. 5, pp. 282–288, 2016.
34. D. Lea, “**A Java fork/join framework,**” in *Proceedings of the ACM 2000 conference on Java Grande - JAVA '00*, 2000, pp. 36–43.
35. P. Smith, D. B. Reid, C. Environment, L. Palo, P. Alto, and P. L. Smith, “**A Threshold Selection Method from Gray-Level Histograms,**” *IEEE Trans. Syst. Man. Cybern.*, vol. 9, no. 1, pp. 62–66, 1979.
36. N. A. Arbain, M. S. Azmi, L. B. Melhem, A. K. Muda, and H. Rashaideh, “**Enhancement of Triangle Coordinates For Triangle Features For Better Classification,**” *Jordanian J. Comput. Inf. Technol.*, vol. 2, no. 2, pp. 108–119, 2016.
<https://doi.org/10.5455/jjcit.71-1448511760>
37. M. S. Azmi, “**Fitur Baharu Dari Kombinasi Geometri Segitiga Dan Pengezonan Untuk Paleografi Jawi Digital,**” Doctoral dissertation, Universiti Kebangsaan Malaysia, 2013.
38. N. A. Arbain, “**Improving Triangle Geometry Shape Features Through Triangle Points Selection In Digit Recognition,**” Thesis for Master, Universiti Teknikal Malaysia Melaka, 2016.
39. M. S. Azmi, M. F. Nasrudin, K. Omar, and K. W. M. Ghazali, “**Farsi/Arabic Digit Classification Using Triangle Based Model Features with Ranking Measures,**” *2012 Int. Conf. Image Inf. Process. (ICIIP 2012)*, vol. 46, no. Iciip, pp. 128–133, 2012.