

Received: 21 Oct, 2017; Accept 8 Feb, 2018; Publish: 21 Feb, 2018

Framework of Page Segmentation for Mushaf Al-Quran Based on Multiphase Level Segmentation

Amirul Ramzani Radzid^{1*}, Mohd Sanusi Azmi², Intan Ermahani A. Jalil³, Azah Kamilah Muda⁴ and Laith Bany Melhem⁵, Nur Atikah Arbain⁶

^{1*, 2, 3, 4, 5, 6} Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia.
^{1*}amirulramzani@gmail.com, ²sanusi@utem.edu.my, ³ermahani@utem.edu.my, ⁴azah@utem.edu.my, ⁵lnm1989@gmail.com, ⁶nuratikah9.arbain@gmail.com

Abstract: This paper presents the framework of page segmentation for Mushaf Al-Quran based on Multiphase Level Segmentation (MLS). This study focuses to (a) extract multiform frame shape by using a novel technique Neighbouring Pixel Behaviors (NPB) and (b) segment text line by using a novel technique which is Hybrid Projection Based Neighbouring Properties (HPBNP). Since Mushaf Al-Quran pages are decorated with a different type of pattern and design of a decorative frame. Thus, the decoration frame must be properly to extract out from a page of Mushaf Al-Quran first before properly get only the text of Mushaf Al-Quran regardless of its decoration heterogeneity. Therefore, NPB technique was proposed to remove multiform frame shape from the page of Mushaf Al-Quran. While the text of Mushaf Al-Quran has a several of diacritical marks, hence it will block the process of segmenting text line. Therefore, HPBNP technique was proposed for segment overlapping text line that interfered by diacritical marks or the stroke of the Arabic word. Experimental results of the proposed technique is shown in this paper.

Keywords: Page Segmentation, Frame Extraction, Extraction Mushaf Al-Quran Decoration, Mushaf Al-Quran Text Segmentation, Line segmentation.

I. Introduction

Mushaf Al-Quran is the most preserved book in the mankind history [1]. It is decorated with various decorations that meant to embellish the presentation of the Holy Quran. However, this decoration will degrade the authentication process. Thus, page segmentation for Mushaf Al-Quran is an important task to extract the only text of Al-Quran from the pages without making any changes to the content of the Mushaf Al-Quran. Page segmentation is a preprocessing stage for document analysis. It is considered as an important initial step for document image analysis and understanding [2]. A document page contains several properties such as halftones, decoration, graphics, text or etc. which can be divided using columns or block [3][4]. Columns or block can be classified in document components such as texts, frames, lines, ornaments and etc that can be segmented. Thus, this page segmentation is the crucial step in order to understand the layout or content of the document. Page segmentation on Mushaf Al-Quran is

challenging due to many variations such as layout structure, decorations and etc. This paper proposed to establish a generic, flexible and multiform segmentation method to unrestricted of decoration frame and the overlapping component of the text line based on MLS.

Some page of Mushaf Al-Quran contains variety form and shape of decoration frame. It is unnecessary to form in order to prettify the page that surrounded the text. It is crucial to extract out decoration frame from the page due to analyses the text. In future work, by analyzing manuscript decoration frame illumination can discover the information of specific manuscript [5].

Page layout can be divided into two classes which are overlapping and nonoverlapping [6]. Overlapping can be found in text line or other component layouts. This paper is concerned with overlapping text line that causes by interfering of diacritical marks or stroke of the Arabic word. Punctuation and diacritic symbols, which are located between text lines make it more complicate deciphering the physical structure of text lines [7]. While nonoverlapping text line components are apparently clear separated by white space.

II. Related work

Document page analysis has two structure: Physical layout and logical structure [8]. The logical structure can be described as logical labels of document physical components where these labels derived from a set of rules. While, the physical layout can be described in various forms, independently of or jointly with document logical structure. These document structured analysis can be seen in studied by Tsujimoto and Asada [9]. In their study represent document the physical layout and the logical structure of trees. By using a set of generic transformation rules and a virtual field separator technique they modeled document understanding as the transformation of a physical tree into a logical one.

Document page image physical layout analysis algorithms can be categorized into three class: top-down approaches, bottom-up approaches and hybrid approaches [8] [10]. The top-down approach in page segmentation is segmenting large regions into smaller sub-regions. Deng Cai [11] in his study for

a vision-based page segmentation algorithm used an automatic top-down, tag-tree independent approach to detect web content structure. Sukhvir Kaur [12] in his study mentioned that the XY cut segmentation algorithm also stated as the recursive XY cuts (RXYC) algorithm and which is referred as tree-based top-down algorithm [13]. On the other hand, the bottom-up approach starts by grouping pixels of interest then merging into larger blocks or connected components. As studied by Akiyama and Hagita [14] perform bottom-up layout analysis that works both global and local text features along with generic properties of documents. It is in a similar to Fisher et al. [15] perform bottom-up segmentation in his studied [16]. While hybrid approaches is a combination of top-down approaches and bottom-up approaches. This approach can relate with Seyyed Yasser Hashemi [17] in his study indicated that hybrid method for segmenting the Persian/Arabic document images used to solve the complexity of layout document. In this study, this paper indicates top-down approach in order to segment page of Mushaf Al-Quran which is from a page into paragraph then paragraph into text line.

In 2016, Ha Dai-Ton et al. [18] in their study on adaptive over-split and merge algorithm for page segmentation. In their study, they had proposed an adaptive over-split and merge algorithm to reduce simultaneously over-segmentation and under-segmentation errors. While, in 2015, Kai Chen et al. [19] has studied on page segmentation of historical document images with convolutional autoencoders. On his paper proposed an unsupervised feature learning method for page segmentation available as color images in whereby applied convolutional autoencoders to learn features directly from pixel intensity values. On the other hand, in 2014, Kai Chen et al. [20] proposed another technique on page segmentation for historical handwritten document images using color and texture features. They proposed a physical structure detection method for the historical handwritten document. In 2016, Kai Chen et al. [2] proposed another technique on page segmentation for historical document images based on superpixel classification with unsupervised feature learning. Besides that, in 2017, Kai Chen et al. [21] proposed another technique which is convolutional neural networks for page segmentation of historical document images. In their paper presents a CNN based page segmentation method for handwritten historical document images. Based on these studies, those techniques unsuitable for Mushaf Al-Quran pages because Mushaf Al-Quran text contains overlapping cause by diacritics or stroke of the Arabic word and multiform frame shape.

In 2013, T. Abu-Ain et al. [22] was proposed text normalization in order for selection of the correct baseline region. This study complies with seven main stages that involved in order to straighten baseline and slant correction.

This research is continuity from past paleography research study. This study can relate to digital Jawi paleography field. Mohd Sanusi Azmi introduced features from triangle geometry for digit recognition on Jawi paleography field [23]. Moreover, this researcher applied his technique to Arabic or Jawi. Thus, it can be related to this research topic because of Mushaf Al-Quran were written in Arabic.

On the other hand, this research is also continuity

pre-processing stage from studied of removing Al-Quran illumination [24]. This studied focusing on removing illumination from the text. Past study also has been done for frame illumination removal and text line segmentation on Mushaf Al-Quran [25] [26] but the proposed techniques were ineffective to solve the problem.

Arabic language that is used is a sacred language of Mushaf Al-Quran [27]. On the other hand, studied has done on Arabic calligraphy classification [28]. This studied on Arabic calligraphy classification of the ancient manuscripts can give useful information to paleographers. Thus, this study can be applied on Mushaf Al-Quran for authentication purpose on future work.

III. Dataset

We experimented with six different type of Mushaf Al-Quran for multiform frame shape extraction. While for text line segmentation, we use four different type of text line in Mushaf Al-Quran that contain overlapping. Table 1 shows dataset of Mushaf Al-Quran pages for experimenting multiform frame shape extraction. While Table 2 shows dataset of Mushaf Al-Quran text lines for experimenting text line segmentation that contains overlapping.

Number	Source	Page
1	Image of Al-Quran Al-Karim from Mawarsoft Digital Furqan 1.0	2
2	Image of Al-Quran Al-Karim from Mushaf Al-Madinah Quran Majeed	1
3	Image of Al-Quran Al-Karim from KSU -Electronic Mosshaf	1
4	Image of Al-Quran Al-Karim from Mushaf Al-Madinah Quran Majeed	3
5	Image of Al-Quran Al-Karim from Mawarsoft Digital Furqan 1.0	4
6	Image of Al-Quran Al-Karim from Uthmani Script Mushaf	2

Table 1. Dataset of Mushaf Al-Quran pages.

Number	Source	Page	Row
1	Mushaf Al-Quran Rasm Uthmani publish by company S Abdul Majeed	6	11-13
2	Mushaf Al-Madinah Quran Majeed	3	3-5
3	Mushaf Al-Madinah Quran Majeed	3	6-8
4	Mushaf Al-Madinah Quran Majeed	3	8-10

Table 2. Dataset of Mushaf Al-Quran text lines.

IV. Proposed Method

A. Pre-processing

Before processing, dataset must be prepared. Page of Mushaf Al-Quran used in this experiment is the collection of text images from Mushaf Al-Quran that has been digitalized. Text image of Mushaf Al-Quran must contain any decoration, illumination, illustration in order to segment multiform of the frame. Conventional steps for instance noise removal and

filtering comprise text normalization for example baseline correction, slant normalization and skew correction must be applied. Those steps create the image to process more reliable and effective [29].

At this phase, image from the page of Mushaf Al-Quran performs preprocessing algorithm as data provision stage. This purpose is to improve and enhance the input image into the uniform format which is binary form. Colored input image will convert into grey-scale format then it will convert into binary format. The conversion process from grey-scale format into the binary format called binarization. This binarization format was refer studied by NB Venkateswarlu and RD Boyle [30] on their new segmentation techniques for document image analysis. The binary form will be labeled as “0” for the foreground while the background will be labeled as “1”.

Thresholding method one of the important technique for image preprocessing that converts a grey-scale image to create a binary image. Thresholding method used in this experiment was conducted by using Otsu’s method proposed by Scholar Otsu in 1979 [31]. The concept of thresholding is to select an optimal grey-level threshold value for separating objects of interest in an image from the background based on their grey-level distribution [31]. If $g(x, y)$ is a threshold version of $f(x, y)$ at some global threshold T , it can be defined as [32]

$$g(x, y) = 1 \text{ if } f(x, y) \geq T \\ = 0 \text{ otherwise} \quad (1)$$

Thresholding operation is defined as:

$$T = M [x, y, p(x, y), f(x, y)] \quad (2)$$

In the equation as stated above (1) and (2), T is stands for the threshold; while $f(x, y)$ is stand for the gray value of point (x, y) and $p(x, y)$ represents as some local property of the point such as the average gray value of the neighborhood centered on point (x, y) .

B. Operational framework page segmentation method

In this paper there is a three phase of segmentation method: a) input image and pre-processing image, b) frame extraction and text line segmentation, c) result output and d) feature extraction and result validation. Figure 1 shows an operational framework for page segmentation method. Input Image and Pre-Processing Image phase have been explained in Section A. On the other hand, Frame Extraction and Text Line Segmentation phase will be explained in Section C. While, Result Output phase will be described in IV. Experiment Result. The result of this experiment is in image form. This output image can extract its features to do validation and classification on proposed techniques. Classification of this experiment was conducted by using Unsupervised Machine Learning (UML) that are used minimum Euclidean distance and average accuracy mean [33].

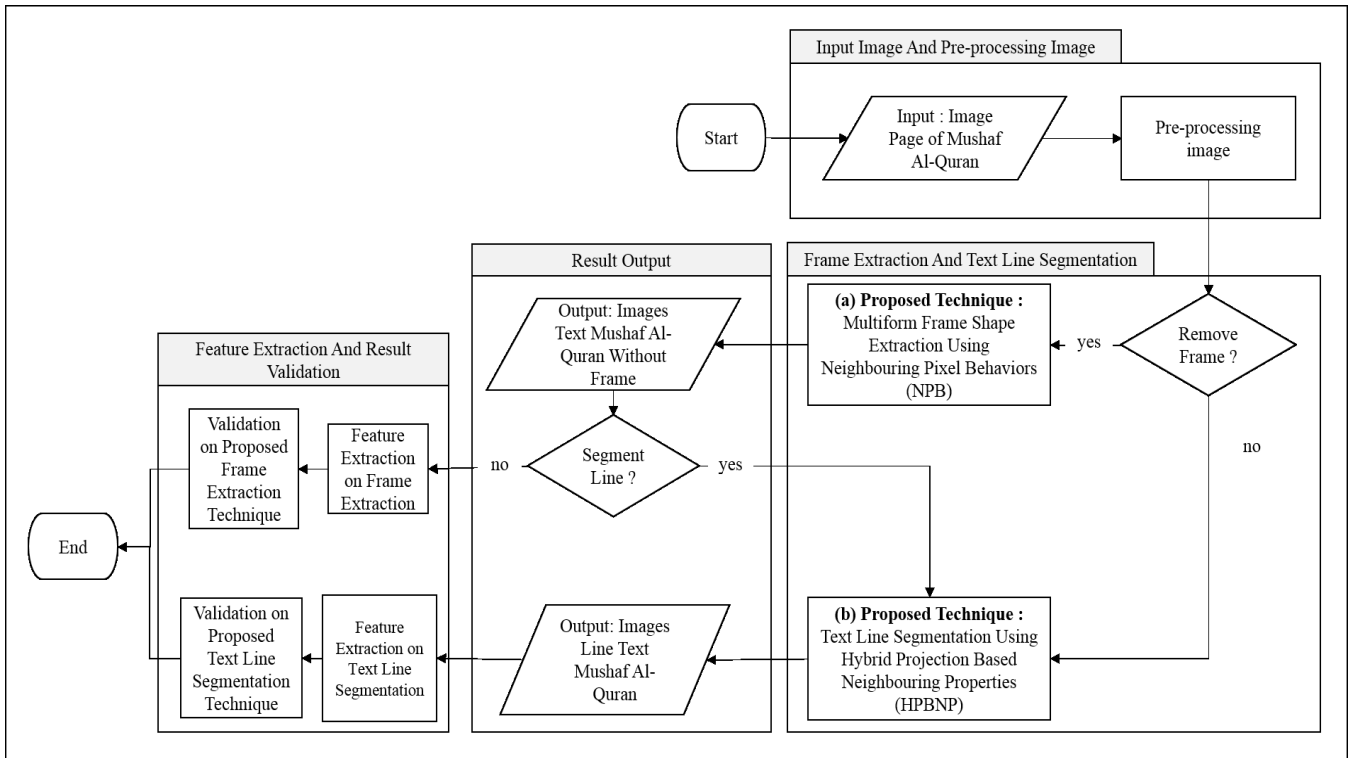
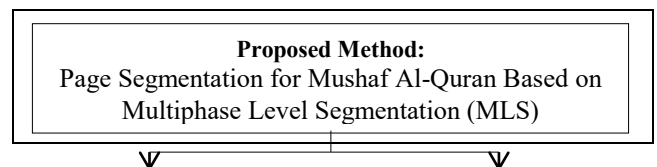


Figure 1. Operational framework page segmentation method

C. Page Segmentation Method

This paper present page segmentation for Mushaf Al-Quran based on Multiphase Level Segmentation (MLS). There are two proposed techniques on MLS indicated as a different level of segmentation method: 1) Neighbouring Pixel Behaviors (NPB) and 2) Hybrid Projection Based Neighbouring Properties (HPBNP). NPB is present to

solving multifom frame shape extraction while HPBNP is present to solving text line segmentation. Figure 2 shows a proposed page segmentation method in this paper.



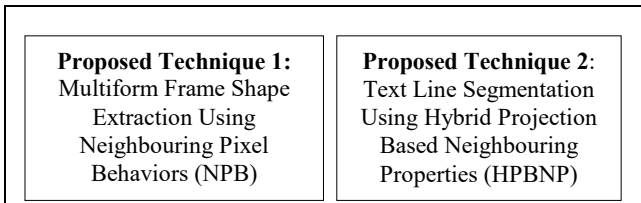


Figure 2. Proposed page segmentation method

1) *Multiform Frame Shape Extraction (MFSE)*

There are several challenges in extracting the significant information in existing Mushaf Al-Quran pages. One of the significant challenges is to extract text that contains different patterns and texture of decorations surround it. In order to extract text, decoration frame must be properly identified from a page of Mushaf Al-Quran. Therefore, multiform frame shape extraction was proposed in order to extract from a page of Mushaf Al-Quran.

This proposed method, multiform frame shape extraction using Neighbouring Pixel Behaviors (NPB) can solve one of the difficulties which are to extract frame decoration from a page. Without removing the decorations, the images can be mistakenly considered as part of Mushaf Al-Quran texts. Thus, this study aims to automatically extract the text of Al-Quran from the images without making any changes to the content of the Mushaf Al-Quran. This is to ensure the extracted images are only the Mushaf Al-Quran texts regardless of Mushaf frame decoration heterogeneity. Thus, this study proposed a novel Neighbouring Pixel Behaviors (NPB) technique to address this problem.

This technique will identify boundary regions. Gap or blank space regions between Arabic text (middle) and decoration (side) which is known as boundary regions. The algorithm computes a wide range of every pixel area to be analyzed which is 4% from the length of a page for vertical point and horizontal point that continually has the same properties of the pixel. Figure 3 shows an example of boundary regions on the Mushaf Al-Quran page.

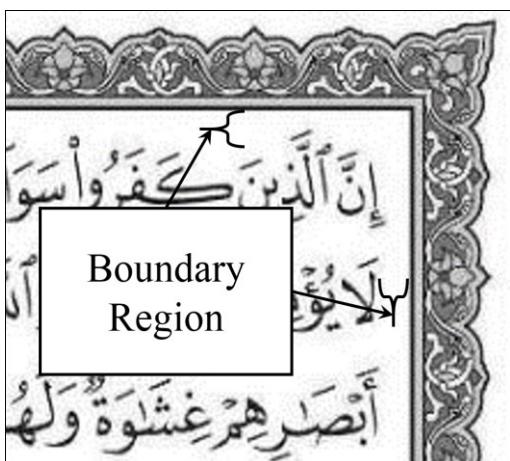


Figure 3. Example of boundary regions on the Mushaf Al-Quran page

The recognize boundary regions that locate outside text area (middle regions) will be passed to the next process of the point of region detection. Four different regions of interest are focused in this study, which are page region, decoration region, boundary region, and text region. With

this, the point of intersection between borders of every region will be identified. It can be applied to a different type of shapes and patterns of Mushaf Al-Quran decoration frame. For example, Figure 4 illustrates the point of detection on rectangle decoration frame, while Figure 5 illustrates the point of detection on oval decoration frame.

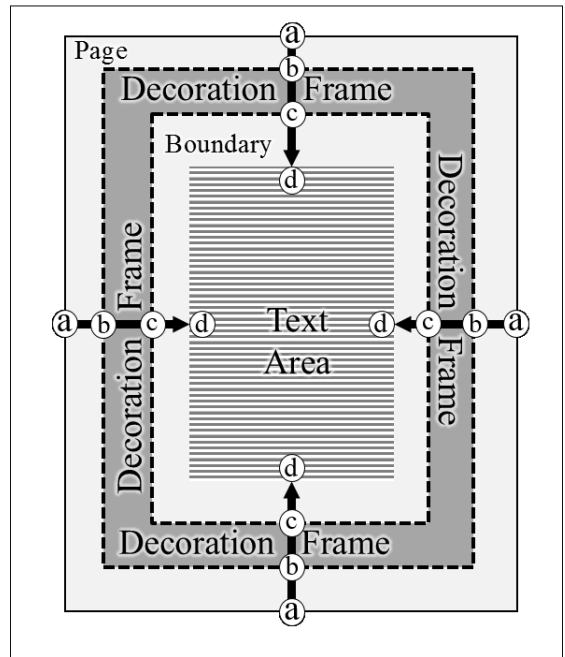


Figure 4. Example point of detection on rectangle decoration frame

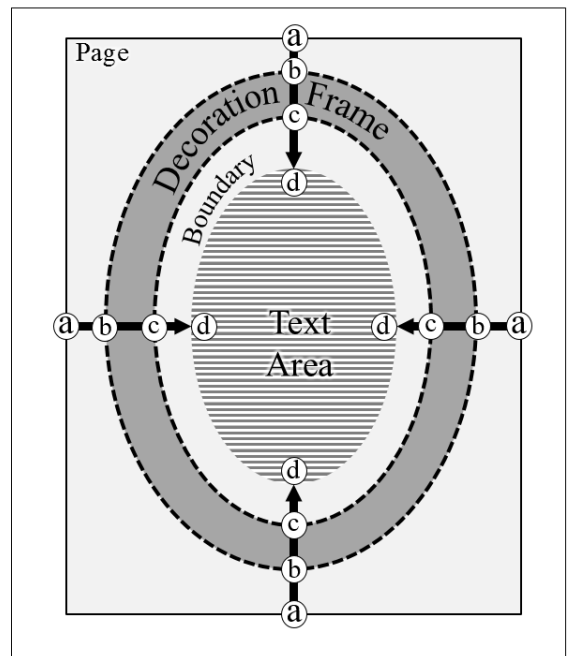


Figure 5. Example point of detection on oval decoration frame

Figure 4 and Figure 5 presented the information as below:

- (a) Point of detection on document page region.
- (b) Point of detection on decoration frame region.
- (c) Point of detection on boundary region.
- (d) Point of detection on text region.

After the point of detection on regions is applied to

recognize the decoration frame, the point of recognition pixels based on neighbouring pixels properties is taken the step. It will cluster pixels which have same properties. Figure 6 shows an example cluster to identify pixels point.

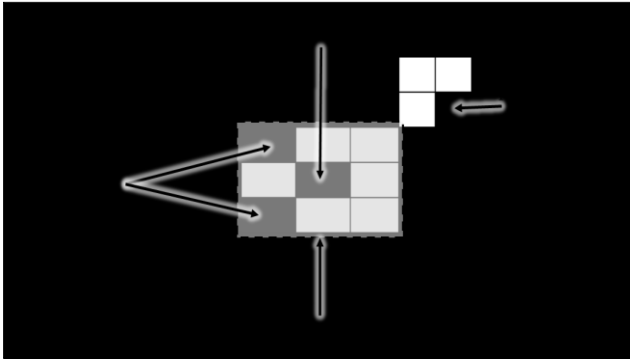


Figure 6. Cluster of pixels point identified by using neighbouring pixels properties

The process of the point of recognition pixels is used to identify balance regions of frame decoration. Balance regions of frame decoration are depicted as Figure 7.

This process can extract multiform of decoration frame from the page of Mushaf Al-Quran. The result as shown in section V. Experiment Result.

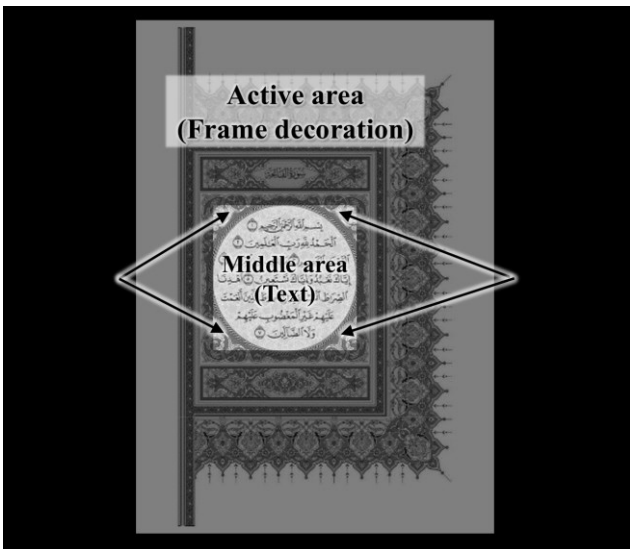


Figure 7. Balance regions of frame decoration

2) Text Segmentation

Text line segmentation is an important step in document image processing. Its part of the pre-processing stage to prepared the images before throughout either feature extraction or classification images. In this paper, we present a novel technique of text line segmentation for Mushaf Al-Quran text by using Hybrid Projection Based Neighbouring Properties (HPBNP). This is based on the pixel, object and histogram properties. This algorithm will identify overlaps between neighboring text lines and segment each line with precision. Overlap cause by interfering of diacritical marks or stroke of the Arabic word must be properly segmented without change the original meaning of the text. Figure 8 shows an example of the

diacritical mark that which cause overlapping of text line segmentation. This diacritical marks is an obstacle of during text line segmentation that causes overlapping as illustrated in Figure 8.

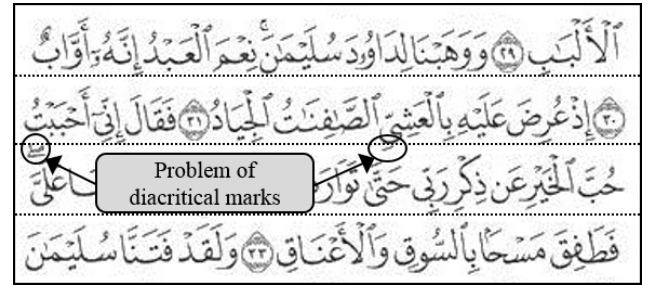


Figure 8. Example of diacritical marks that cause of overlapping

Fist step algorithm compute horizontal projection profile in order to calculate each pixel by row to project its graph as shown in Figure 9.

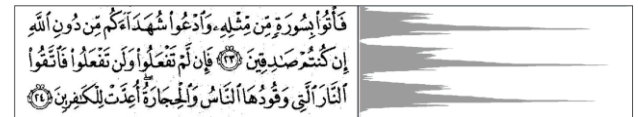


Figure 9. Result of horizontal projection histogram

Second step algorithm computes object ownership in order to calculate the distance of baseline or distance of the determined object.

In order to determine the object of diacritical marks ownership based on the distance of baseline, the algorithm will calculate the gap between diacritical marks or stroke of the Arabic word with upper text baseline and bottom text baseline. The nearest text baseline will be owned for the object. The distance of baseline and object are depicted as Figure 10.

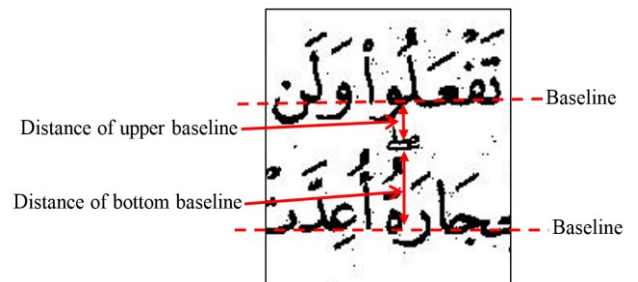


Figure 10. Illustration of the distance between the object (diacritical marks or stroke of the Arabic word) with baseline

The others process will be the distance of the determined object. In order to determine the object of diacritical marks ownership based on the distance of the determined object, the algorithm will calculate the gap between diacritical marks or stroke of the Arabic word with upper determined object and bottom determined object. The nearest text baseline will be owned for the object of diacritical marks. The distance of the object of diacritical marks objects and the determined object of are depicted as Figure 11.

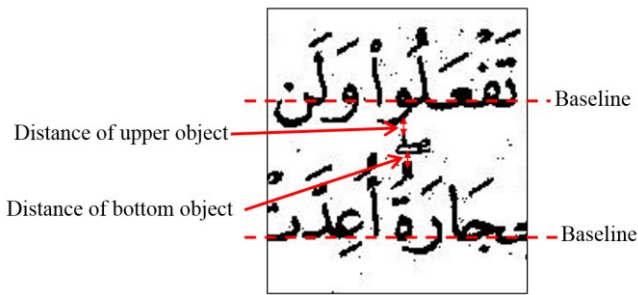


Figure 11. Illustration of the distance between the object (diacritical marks or stroke of the Arabic word) with the nearest object

Lastly, the algorithm will segment text line determined base on horizontal projection profile to detect its number of baseline. Then, it will consider the lower peak of contour as overlap. For overlap, it will determine object possession to determine its row number of the text line. The pseudocode is defining as shown in Figure 12.

```

1.0 Start
2.0 Read input image
3.0 Input image → pre-processing image
4.0 Detect baseline using horizontal projection profile
5.0 Fabricate object using neighbouring pixel properties
6.0 Determine object possession
    5.1 Define object possession using distance of baseline
    5.2 Define object possession using determined object
7.0 Output result image
8.0 End
    
```

Figure 12. Pseudocode Hybrid Projection Based Neighbouring Properties (HPBNP)

D. Feature extraction and result validation.

The result of this processing is in the form of images and binaries to facilitate the further process which is feature extraction. The resulting image results obtained will be extracted using the Geometric Triangle Using Background Foreground Image (STDIL) that has been suggested by N. A. Arbain et al. [34]. Experimental results are produced by comparing the results of the present techniques with the prior proposed techniques using unsupervised machine learning (UML). The UML used are minimum Euclidean distance and average accuracy mean (AAM). The result of this phase does not state in this paper will be stated in further research.

V. EXPERIMENT RESULT AND DISCUSSION

The experiment was implemented in Java and tested on the selected dataset of Mushaf Al-Quran as stated in section III Dataset. The result from the proposed method was compared with Binary Representation (BR) techniques that proposed by L. B. Melhem in 2015 [25] and 2017 [26].

A. Comparison Frame Extraction and Removal

In order to remove the multiform shape from Mushaf Al-Quran page, it must identify at first. Most research is focusing on removing illumination or ornament [2] [20] [21] [35] from the page. Past research has shown the object end

of the verse (Taskil) are misinterpreted as part of illumination or ornament, whereas in this study the object end of the verse is part of the text to guide as the end of the verse and the number of verse in Al-Quran. It also can be used later in further study to segmenting the verse of text Mushaf Al-Quran. Moreover, in this study will remove all text outside from decoration frame including the name of surah at the top of the page that does not effect on the ayah of Mushaf Al-Quran. This study also differs from past research that focuses on the different domain. *Table 3* shows the result of multiform frame extraction using the proposed method which is Neighbouring Pixel Behaviors (NPB). Our proposed method can identify or recognized the different shape of decoration on Mushaf Al-Quran page.

Source (Refer Table 1)	Input Image	Multiform Frame extraction (Binary Format)
1		
2		
3		
4		
5		
6		

Table 3. Result of Multiform Frame Identification.

Most related study for removing frame in domain Mushaf Al-Quran has been done by L. B. Melhem in 2015 [25] by using Binary Representation (BR). Unfortunately, BR unsuccessful to remove decoration frame from image source 1 (Image of Al-Quran Al-Karim from Mawarsoft

Digital Furqan 1.0 page 2), source 2 (Image of Al-Quran Al-Karim from Mushaf Al-Madinah Quran Majeed page 1) and source 3 (Image of Al-Quran Al-Karim from KSU - Electronic Mosshaf). This is because the method proposed by researcher only can be solved on the rectangle shape. While on this research solve multiform frame on page Mushaf Al-Quran as shown on Table 4. Table 4 shows a result comparison of multiphase for experimenting multiform frame shape extraction.

Source (Refer Table 1)	Input Image	Result of Binary Representation [25]	Result of Proposed Method (NPB)
1			cannot be processed
2			cannot be processed
3			cannot be processed
4			
5			
6			

Table 4. Result comparison of multiphase for page segmentation.

B. Comparison Text Segmentation

Comparison with BR techniques [26] is made because their research about text segmentation is in the same domain which is Mushaf Al-Quran. However, the proposed techniques were ineffective to solve the problem which Mushaf Al-Quran text. This is because Mushaf Al-Quran text contains diacritical marks and stroke of the Arabic word will cause overlapping. While this research proposed text segmentation to solve overlapping text on page Mushaf

Al-Quran as shown in Table 5 - Table 8. The result showed that proposed method HPBNP can solve overlapping problem.

Table 5 - Table 8 shows dataset of Mushaf Al-Quran text lines for experimenting text line segmentation that contains overlapping.

Input	
Result of Binary Representation [26] [25]	
Result of Proposed Method (HPBNP)	

Table 5. Result of text image of Mushaf Al-Quran Rasm Uthmani publish by company S Abdul Majeed page 6.

Input	
Result of Binary Representation [26] [25]	
Result of Proposed Method (HPBNP)	

يُخَالِدُونَ اللَّهَ وَالَّذِينَ آمَنُوا وَمَا يُخَالِفُونَ إِلَّا أَنفُسَهُمْ
Row 5

Table 6. Result of text image of Mushaf Al-Quran from Mushaf Al-Madinah Quran Majeed page 3.

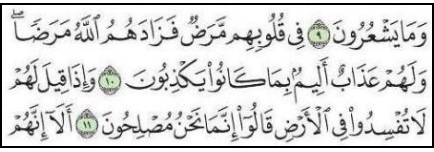
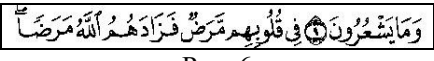

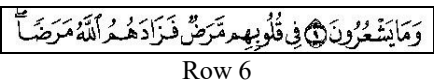
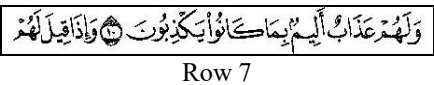
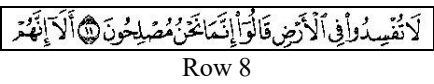
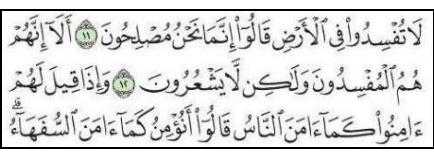
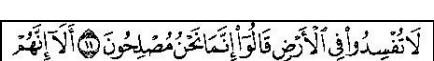
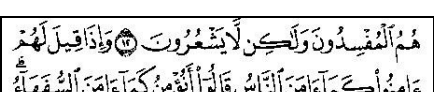
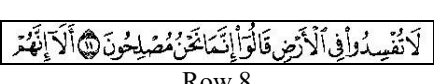

Input	 <p>Row 6-8</p>
Result of Binary Representation [26] [25]	 <p>Row 6</p>  <p>Row 7-8</p>
Result of Proposed Method (HPBNP)	 <p>Row 6</p>  <p>Row 7</p>  <p>Row 8</p>

Table 7. Result of text image of Mushaf Al-Quran from Mushaf Al-Madinah Quran Majeed page 3.

Input	 <p>Row 8-10</p>
Result of Binary Representation [26] [25]	 <p>Row 8</p>  <p>Row 9-10</p>
Result of Proposed Method (HPBNP)	 <p>Row 8</p>  <p>Row 9-10</p>

Row 10

Table 8. Result of text image of Mushaf Al-Quran from Mushaf Al-Madinah Quran Majeed page 3.

VI. CONCLUSION

In this paper, we present a framework for page segmentation for Mushaf Al-Quran based on Multiphase Level Segmentation (MLS). This study focusing to extract multiform frame shape by using the novel technique which is Neighbouring Pixel Behaviors (NPB) and segment text line by using the novel technique which is Hybrid Projection Based Neighbouring Properties (HPBNP). NPB technique will remove multiform frame shape from the page of Mushaf Al-Quran. While HPBNP technique will segmenting overlapping text line that caused of interfering with diacritical marks or stroke of the Arabic word.

The result is for multiform frame shape extraction are compared with Binary Representation technique that was proposed by L.B. Melhem [25] with the same dataset as shown in Table 4. Dataset that are being used for conducting this experiment are shown in Table 1. The result is shown that the proposed method named Neighbouring Pixel Behaviors (NPB) for multiform frame shape extraction is more efficient to solve the problem compare than prior research.

The result for text line segmentation are compared with L.B. Melhem [26] with the same dataset as shown in Table 5, Table 6, Table 7 and Table 8. The dataset that is being used for conducting this experiment is shown in Table 2. The result is shown that the proposed method named Hybrid Projection Based Neighbouring Properties (HPBNP) for text line segmentation are more efficient to solve the problem compare than prior research.

Feature work for this study will be verse segmentation. Object end of the verse (Taskil) will be guided to segment full sentence of the verse. This proposed method will be applied to conduct verse segmentation.

Acknowledgment

The authors would like to express their appreciation to the Universiti Teknikal Malaysia Melaka for the scholarship of Zamalah UTeM Scheme. Thank also to the Faculty of Information Technology and Communication for providing the excellent research faculties and facilities.

References

- [1] C. Paper, C. A. Language, and P. D. View, "Data Preparation and Handling for Written Quran Script Verification," no. October, 2016.
- [2] 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.
- [3] T. Pavlidis and J. Zhou, "Page segmentation and

- classification,” *CVGIP Graph. Model. Image Process.*, vol. 54, no. 6, pp. 484–496, Nov. 1992.
- [4] A. K. Jain and Y. Zhong, “Page segmentation using texture analysis,” *Pattern Recognit.*, vol. 29, no. 5, pp. 743–770, 1996.
- [5] M. S. Azmi and K. Omar, “Features Extraction of Arabic Calligraphy using extended Triangle Model for Digital Jawi Paleography Analysis,” *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 5, pp. 696–703, 2013.
- [6] K. Kise and A. Sato, “Page Segmentation Using the Area Voronoi Diagram,” *Tech. Rep. IEICE. PRMU*, vol. 96, no. 598, pp. 9–16, 1997.
- [7] R. Saabni, A. Asi, and J. El-Sana, “Text line extraction for historical document images,” *Pattern Recognit. Lett.*, vol. 35, no. 1, pp. 23–33, 2014.
- [8] S. Mao, A. Rosenfeld, and T. Kanungo, “Document Structure Analysis Algorithms: a Literature Survey,” *SPIE 5010, Doc. Recognit. Retr. X*, vol. 5010, no. 1, p. 197, 2003.
- [9] S. Tsujimoto and H. Asada, “Understanding multi-articled documents,” in *[1990] Proceedings. 10th International Conference on Pattern Recognition*, 1990, vol. i, no. 4, pp. 551–556.
- [10] Song Mao and T. Kanungo, “Empirical performance evaluation methodology and its application to page segmentation algorithms,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 242–256, 2001.
- [11] D. Cai, S. Yu, J. R. Wen, and W. Y. Ma, “VIPS: a visionbased page segmentation algorithm,” *Beijing Micosoft Res. Asia*, pp. 1–29, 2003.
- [12] S. Kaur, P. Mann, and S. Khurana, “Page Segmentation in OCR System-A Review,” *Ijcsit.Com*, vol. 4, no. 3, pp. 420–422, 2013.
- [13] G. Nagy, S. Seth, and M. Viswanathan, “A Prototype Document Image Analysis System for Technical Journals,” *Computer (Long. Beach. Calif.)*, vol. 25, no. 7, pp. 10–22, 1992.
- [14] T. Akiyama and N. Hagita, “Automated entry system for printed documents,” *Pattern Recognit.*, vol. 23, no. 11, pp. 1141–1154, Jan. 1990.
- [15] G. Nagy, S. Seth, and M. Viswanathan, “A prototype document image analysis system for technical journals,” *Computer (Long. Beach. Calif.)*, vol. 25, no. 7, pp. 10–22, Jul. 1992.
- [16] L. O’Gorman, “The document spectrum for page layout analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1162–1173, 1993.
- [17] S. Yasser Hashemi, “Persian/Arabic Document Segmentation Based on Hybrid Approach,” *Int. J. Comput. Sci. Appl.*, vol. 4, no. 1, pp. 23–34, 2014.
- [18] H. Dai-Ton, N. Duc-Dung, and L. Duc-Hieu, “An adaptive over-split and merge algorithm for page segmentation,” *Pattern Recognit. Lett.*, vol. 80, pp. 137–143, 2016.
- [19] K. Chen, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold, “Page segmentation of historical document images with convolutional autoencoders,” *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 2015–Novem, pp. 1011–1015, 2015.
- [20] K. Chen, H. Wei, J. Hennebert, R. Ingold, and M. Liwicki, “Page Segmentation for Historical Handwritten Document Images Using Color and Texture Features,” *2014 14th Int. Conf. Front. Handwrit. Recognit.*, pp. 488–493, 2014.
- [21] 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.
- [22] T. Abu-Ain, S. N. H. S. Abdullah, K. Omar, A. Abu-Ein, B. Bataineh, and W. Abu-Ain, “Text Normalization Method for Arabic Handwritten Script,” *J. ICT Res. Appl.*, vol. 7, no. 2, pp. 164–175, Nov. 2013.
- [23] M. S. Azmi, M. F. Nasrudin, K. Omar, C. W. S. B. C. W. Ahmad, and K. W. M. Ghazali, “Exploiting features from triangle geometry for digit recognition,” *2013 Int. Conf. Control. Decis. Inf. Technol. CoDIT 2013*, pp. 876–880, 2013.
- [24] A. R. Radzid, “Removing Al-Quran Illumination,” Thesis for Bachelor Degree, Universiti Teknikal Malaysia Melaka, 2016.
- [25] L. N. B. Melhem, “Illumination Removal And Text Segmentation For Al-Quran Using Binary Representation,” Thesis for Master, Universiti Teknikal Malaysia Melaka, 2015.
- [26] L. B. Melhem, M. S. Azmi, A. K. Muda, N. J. Bani-Melhim, and M. Alweshah, “Text Line Segmentation of Al-Quran Pages Using Binary Representation,” *Adv. Sci. Lett.*, vol. 23, pp. 11498–11502, 2017.
- [27] H. Ishkewy, H. Harb, and H. Farahat, “Azhar: An Arabic Lexical Ontology,” *Int. J. Web Semant. Technol.*, vol. 5, no. 4, pp. 71–82, 2014.
- [28] M. S. Azmi, K. Omar, M. F. Nasrudin, A. K. Muda, and A. Abdullah, “Arabic calligraphy classification using triangle model for Digital Jawi Paleography analysis,” *Proc. 2011 11th Int. Conf. Hybrid Intell. Syst. HIS 2011*, pp. 704–708, 2011.
- [29] F. Farooq, V. Govindaraju, and M. Perrone, “Pre-processing methods for handwritten Arabic documents,” in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2005*, vol. 2005, pp. 267–271.
- [30] N. Venkateswarlu and R. Boyle, “New segmentation techniques for document image analysis,” *Image Vis. Comput.*, vol. 13, no. 7, pp. 573–583, 1995.
- [31] M. H. J. Vala and A. Baxi, “A review on Otsu image segmentation algorithm,” *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 2, no. 2, pp. 387–389, 2013.
- [32] B. C. Rafael Gonzalez and R. E. Woods, *Digital Image Processing (2nd Edition)*. 2002.
- [33] 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.
- [34] N. Arbain, M. Azmi, L. Melhem, A. Muda, and H. Rashaideh, “Enhancement Of Triangle Coordinate

For Triangle Features For Better Classification,” *Jordanian J. Comput. Inf. Technol.*, vol. 2, no. 2, p. 107, 2016.

- [35] 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.

(UTeM). His research interests include Image processing and segmentation.



Nur Atikah Arbain was born in Melaka, Malaysia. She received her Bachelor of Computer Science in Database Management on 2015 and Master of Science in Information and Communication Technology on 2016 at Universiti Teknikal Malaysia Melaka. She is currently pursuing her PhD which is also at the same university. Her current research work is an offline subword handwriting and contributes in feature extraction domain.

Author Biographies



Amirul Ramzani Radzid received Bachelor in Computer Science of Software Development from University Teknikal Malaysia Melaka (UTeM) in 2016. Currently he is pursuing Master of Science in Information and Communication Technology at the same university which is Universiti Teknikal Malaysia Melaka (UTeM). His current research work is text image segmentation.



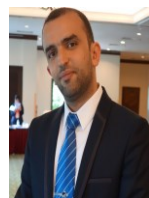
Mohd Sanusi Azmi received BSc., Msc and Ph.D from Universiti Kebangsaan Malaysia (UKM) in 2000, 2003 and 2013. He joined Department of Software Engineering, Universiti Teknikal Malaysia Melaka (UTeM) in 2003. Now, he is currently a senior lecturer at UTeM. He is the Malaysian pioneer researcher in identification and verification of digital images of Al-Quran Mushaf. He is also involved in Digital Jawi Paleography. He actively contributes in the feature extraction domain. He has proposed a novel technique based on geometry feature used in Digit and Arabic based handwritten documents.



Intan Ermahani A. Jalil received the BSc degree in Computer from Universiti Teknologi Malaysia (UTM), Malaysia, the MSc degree in Software Engineering from the University of Brighton, UK and Ph.D degree in Computer from Universiti Teknologi Malaysia (UTM), Malaysia, in 2003, 2004 and 2017 respectively. She is currently a lecturer in Faculty of Information and Communication Technology (FTMK) of Universiti Teknikal Malaysia Melaka (UTeM), Malaysia. Her research interests include the area of pattern recognition, handwriting identification, features ranking, software development, and software testing and software project management.



Azah Kamilah Muda is an Associate Professor at Faculty of ICT, UTeM. She has appointed as Deputy Dean of Post Graduate and Research since 2015. She received her PhD in 2010 from Universiti Teknologi Malaysia, specializing in image processing. Her research interest includes fundamental studies on data analytics using soft computing techniques, pattern analysis and recognition, image processing, machine learning, computational intelligence and hybrid systems. Her current research work is on pattern analysis of molecular computing for drug analysis, data analytic for various application and root cause analysis in manufacturing process.



Laith Bany Melhem received the BSc. in Computer Science from Jordan University of Science and Technology (JUST) in 2011, and Msc in Computer Science (Internetworking Technology) from Universiti Teknikal Malaysia Melaka (UTeM) in 2015. In 2015 he was awarded a Malaysia International Scholarship (MIS) to pursuing the Ph.D. Currently he is a Ph.D student at Universiti Teknikal Malaysia Melaka