The Segmentation of Printed Arabic Characters Based on Interest Point

Fitriyatul Qomariyah, Fitri Utaminingrum, Wayan Firdaus Mahmudy
Faculty of Computer Science, Brawijaya University, Indonesia
fitriyatulq1017@gmail.com

***Abstract*—Arabic characters are different compared to the other characters whether from their forms or the way they are read. Before conducting a recognition process, we should conduct segmentation or divide each character to identify each Arabic character of the word. The enormous problem of segmenting the connected Arabic characters is dividing each character with different positions, forms, and sizes for each character. Therefore, we suggested a method in segmentation process by using the interesting point, which successfully obtains the 86.5% average accuracy.**

***Index Terms*—Image Segmentation; Connected Arabic Characters Segmentation; Interest Point.**

I. INTRODUCTION

Arabic characters are different compared to the other characters whether from their form or the way they are read. Arabic characters, also called as hijayyah letters consist of 29 letters: 26 letters are the genuine consonants and 3 letters are the semi-vocal consonants, which are, "Alif", "Waw" and "Ya". Connected Arabic characters are divided into two types of character, which are the separate character and the connected characters in which, every connected character has three different forms at the beginning, middle, ending, and isolated place. It depends on the character's position in a word. For the recognition process of Arabic character, we need to separate each character. Hence, we conducted a research to separate each Arabic character in a word.

The image segmentation is a crucial part in the image processing to analyze the image [1]. The image segmentation is the process that divides a picture into a certain area for the purpose of extracting the necessary information and eliminating unnecessary information [2]. This segmentation process has a crucial role in recognizing each Arabic character. A good segmentation process helps the system to identify each Arabic character well [1]. Optical Character Recognition (OCR) is a process of modifying texts on the picture into the text code readable by the computer [3]. Although it is good to control the segmentation of separate characters using OCR, it has the disadvantage to control the segmentation of connected characters.

There are many algorithms to perform the segmentation of Arabic characters. Mohammad et al. separated the printed Arabic characters [1]. Wshah et al. separated the handwriting of Arabic characters using the contour and skeleton of Arabic characters [4]. Yi-Kai Chen combined the background and foreground analysis to the single and multi-touching handwritten numbers segmentation and used the Mixture Gaussian Function to decide which is the best o possible routes segmentation [5]. Wei suggested the Genetics Algorithm to Chinese connected character segmentation [6].

In image segmentation, an interest point can be used as the important parameter. Interest point is necessary for image processing and can be used as the descriptor to image matching [7]. The majority of the interest point extractions are the point gained based on intensity [8][9][10]. Interest point method is pushing aside the other important information. Some of the constructed interest point researches are based on the colors because the characteristic formed by the colors is superior [11]. Therefore, the interest point based on the colors is important for matching the image. However, the interest point method based on the colors is not suitable in some case of an image with the same color.

Based on the weakness identified in the previous research, this research suggested a segmentation method using the interesting point based on some rules to separate the connected Arabic character. On the segmentation process, the interest point is used as the coordinate reference to split each character of Arabic writing image data with a size of 769x149. The results from the experiment showed a fairly good accuracy. By using cameras with different resolutions, the accuracy gained is different because of the different camera quality [12]. The interest point of the suggested segmentation process can be seen in the design system below.

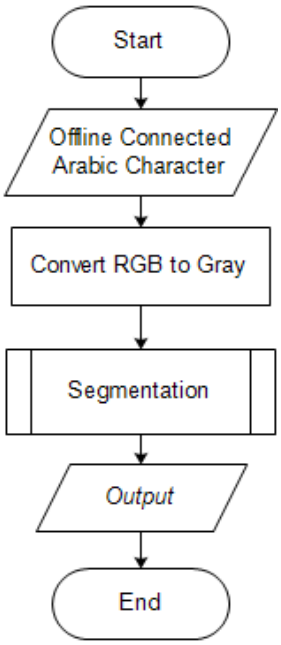


Figure 1: Design system

The detail of the segmentation process is explained in the second section. Whilst, the conclusion is explained in the subsequent section.

II. SEGMENTATION CONNECTED ARABIC CHARACTER

Arabic character segmentation is the process of separating each connected character in a word. The segmentation process in this study involved in the use of a few rules used as an interest point. Several steps have been taken before the process of segmentation, which are thresholding, complement, and thinning. These processes will be discussed below. The design system of segmentation process can be seen below.

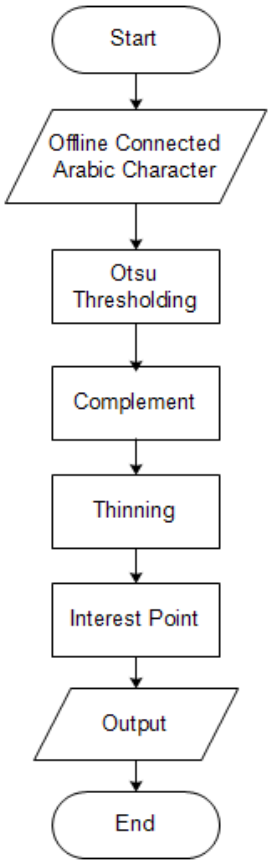


Figure 2: Design system of segmentation process

A. Otsu Thresholding

Otsu thresholding is a method for separating the background and object in the image based on a certain threshold value. However, due to bad lighting and non-uniform background, it is difficult to get a good threshold value. With the assumption that the original image has light and dark regions are denoted as C_1 and C_2, and the probability of each region, $P(s)$ is as following [13]:

$$P_1(s) = \sum_{i=0}^{s} P_i \quad for \quad C_1[s > 0] \tag{1}$$

$$P_2(s) = \sum_{i=s+1}^{L-1} P_i = 1 - P_1(s) \quad for \quad C_2[s+1, L-1] \tag{2}$$

Where $P_1(s)$ and $P_2(s)$ are the probability of light and dark regions (C_1 and C_2). P_i is the normalized histogram for the overall size of pictures. Besides, to get the threshold value k based on variations in the contrast of the image, there are some presumptions; bimodal histogram, statistics stationary, hence, adaptive and uniform illumination area can be modified. From these presumptions, Otsu thresholding is acquired by the following equation.

$$k = \frac{\sigma^2 B}{\sigma^2 G} \tag{3}$$

where k is threshold value, $\sigma^2 B$ is a global diversity of the whole picture, and $\sigma^2 G$ is between-class diversity.

B. Complement

Complement is a process to modify each pixel of the image (binary image) into the opposite value as the example showed in Figure 3.

1	1	1	0	0	0
0	0	1	1	1	0
0	1	1	1	0	0

Figure 3: Complement

C. Thinning

Thinning algorithm is largely based on looping layers of the pixels until no pixels can be removed. Usually, looping is performed twice to check the changes of pixels. If there is no change, the condition will be stopped [14]. ZS algorithm is the most popular and proven thinning algorithms, proposed by Zhang and Suen in 1984 [15]. The ZS algorithm is the thinning algorithm that works on 3x3 neighborhood parallel, shown in Figure 4.

P8	P2	P3
P7	P1	P4
P6	P1	P5

Figure 4: ZS 3x3 neighborhood

The ZS algorithm is an algorithm, which consists of two sub-iterations: The first iterations intends to remove the barrier pixel of South-East and the corner pixel of North-West, while the second sub-iteration aims to remove the barrier pixel of North-West and the corner pixel of South-East. These are the opposite orientation [14].

A method for extracting picture frame consists of removing all of the contour pixels in the image, except the contour appropriate with the framework. At the first sub-iteration (in the odd iteration), contour point in p_1 is removed from the pattern, if the following conditions exist:

$$2 \le B(p_1) \le 6 \tag{4}$$

$$A(p_1) = 1 \tag{5}$$

$$p_2 \times p_4 \times p_6 = 0 \tag{6}$$

$$p_4 \times p_6 \times p_8 = 0 \tag{7}$$

In the second sub-iteration (in the even iteration), contour point in p_1 is removed from the pattern based on the following conditions:

$$2 \le B(p_1) \le 6 \tag{8}$$

$$A(p_1) = 1 \tag{9}$$

$$p_2 \times p_4 \times p_6 = 0 \tag{10}$$

$$p_4 \times p_6 \times p_8 = 0 \tag{11}$$

where A (p_1) is the number pairs of 0-1 (white-black) in transverse clockwise from 8 neighborhood in (p_1) like (p_2),

(p_3), (p_4) ,..., (p_8), (p_9). B (p_1) is the number of which not zero in 8 neighbors, that is:

$$B(p_1) = \sum_{i=2}^{9} P_i \quad (12)$$

If there are no matching conditions, p1 is not removed from the foreground.

(a) (b)

Figure 5: (a) Image before thinning process and (b) image after thinning process

D. Segmentation

Each character in the Arabic letters has different form according to the position of these characters in a word (beginning, middle, end, and isolated position). This factor creates major challenges in the segmentation process. The segmentation process in this paper is to look for the important coordinate as a reference to separate each character.

Some studies proposed a method for determining interest points, such as Harris Point Detection method. However, Harris Point Detection method is less suitable when it is used to Arabic letters. Therefore, we made some new interest points to segment the connected Arabic characters. In this study, the important coordinates are based on interest point gained from some intersection rules using a number of conditions.

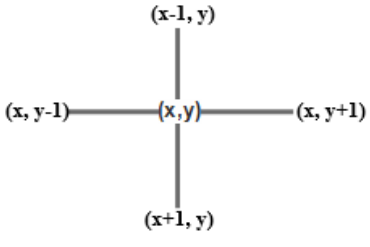


Figure 6: The intersection coordinates

(x,y) coordinates is 1. The conditions are including :

a First condition
(x-1, y) = 1
(x, y+1) = 1
(x+1, y) = 1
(x, y-1) = 1

b Second condition
(x-1, y) = 1
(x, y+1) = 1

c Third condition
(x+1, y) = 1
(x, y-1) = 1

d Fourth condition
(x-1, y) = 1
(x, y+1) = 1
(x+1, y) = 1

e Fifth condition
(x+1, y) = 1
(x, y-1) = 1

f Sixth condition
(x-1, y) = 1
(x+1, y) = 1
(x, y-1) = 1

g Seventh condition
(x-1, y) = 1
(x, y+1) = 1
(x, y-1) = 1

h Eighth condition
(x+1, y) = 1
(x, y+1) = 1

i Ninth condition
(x+1, y) = 1
(x, y+1) = 1
(x, y-1) = 1

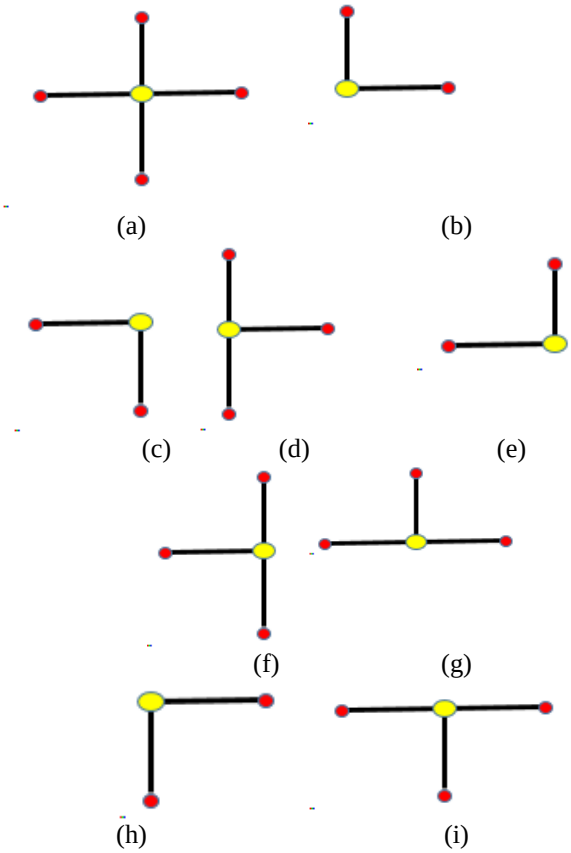


Figure 7: Image of first intersection (a), second intersection (b), third intersection (c), fourth intersection (d), fifth intersection (e), sixth intersection (f), seventh intersection (g), eighth intersection (h), end ninth intersection (i)

III. EXPERIMENT AND RESULT

A. Trial

In this study, we used Arabic writing image data with a size of 769x149. The data used in this study is shown in Figure 8.

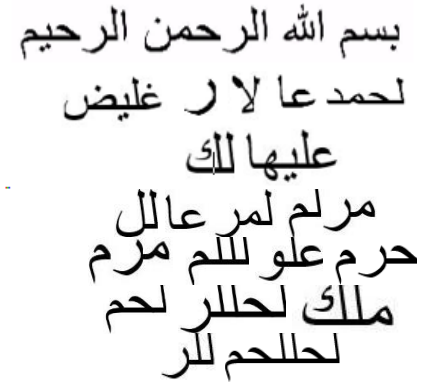


Figure 8: 16 Trials

In the next process, the sentences are separated based on their region, as shown in Figure 9.

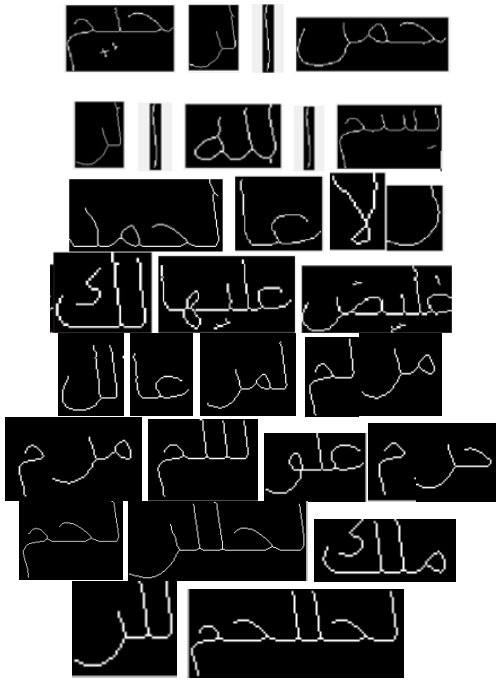


Figure 9: Word in each region

The next process is to acquire the interest point using predefined rules to separate each character, as shown in Figure 10.

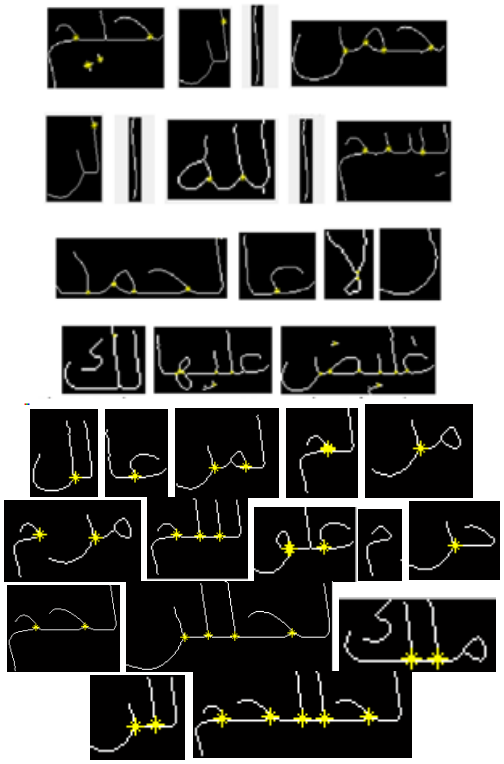


Figure 10: Word in each region with interesting point

From the interest point that we get in Figure 10, we split each character using interest point as a reference using the conditions:

1. Right Side
   a. Width_Column = width – y.
   b. Widht_Line = height.
   c. Crop = imcrop (figure, [y 0 Width_Column Width_Line])
2. Middle Side
   a. Width_Column = (y+1) – y
   b. Width_Line = height.
   c. Crop = imcrop (figure, [y 0 Width_Column Width_Line])
3. Left Side
   a. Width_Column = y
   b. Width_Line = height
   c. Crop = imcrop (figure, [0 0 Width_Column Width_Line])

B. Analysis of Trial Results

The test result that consists of regions is as shown in Figure 11.

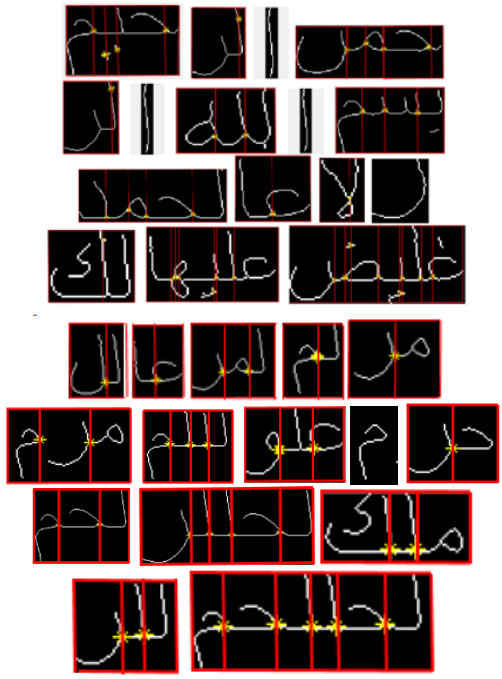
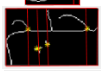
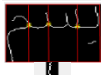
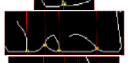
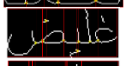
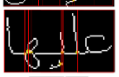
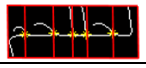


Figure 11: Bounding box result

From Figure 11 above, the table of testing results is shown in Table 1:

Table 1
Table of Testing Result

No.	Word	Number of Character	Segmented Character
1		3	5
2		1	1
3		2	2
4		3	5
5		3	4
6		1	1

7		3	3
8		1	1
9		2	2
10		1	1
11		2	2
12		2	2
13		4	5
14		4	9
15		5	7
16		2	2
17		2	2
18		2	2
19		3	3
20		2	2
21		2	2
22		3	3
22		4	4
23		3	3
24		1	1
25		2	2
26		3	3
27		5	5
28		3	3
29		3	3
30		6	6

The test results of the connected Arabic character segmentation using the calculation accuracy are presented in Table 2.

Table 2
Table Accuracy

ACTUAL		**Predicted**	
		Yes	**No**
	Yes	TP	FN
	No	FP	TN

Explanation of TP, TN, FP and FN are:

a. TP (True Positive), if the real value and the prediction is worth "Yes", that is if:
The real value is three letters, then the prediction decide three letters
b. TN (True Negative), if the real value and the prediction is worth "No", that is if:
The real value is not three letters, then the prediction decides three letters.
c. FP (False Positive), if the real value "No" but prediction decide "Yes," that is when:
Real value is nothing, but prediction decide the detected letters
d. FN (False Negative), if the real value "Yes" but prediction "No", that is if:
The real value stated there are letters, but prediction decides no letters from those detected.

System validation was assessed by counting the numbers of TP, TN, FP, and FN of Table I using the formula:

$$\text{System Performance} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$$

$$= \frac{74+0}{74+0+13+0} \times 100\%$$

$$= 86.5\ \%$$

IV. Conclusion

Based on the results of the experiment, it can be concluded that the Arabic character segmentation using interesting point is very good for some words with the good image condition. Based on the image size of 769x149, we got a fairly good accuracy value of 86,5 %. The good picture quality will increase the level of accuracy. For further research, the author will make the repairing process of the rule to improve the accuracy and continue this research to the process of Arabic character recognition.

References

[1] K. Mohammad, M. Ayyesh, A. Qaroush, and I. Tumar, "Printed Arabic optical character segmentation," *SPIE/IS&T Electron. Imaging*, vol. 9399, p. 939911, 2015.

[2] Tihao Chiang and Ya-Qin Zhang, "A new rate control scheme using quadratic rate distortion model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 1, pp. 246–250, 1997.

[3] S. Shastry, G. Gunasheela, T. Dutt, D. S. Vinay, and S. R. Rupanagudi, "'i' — A novel algorithm for optical character recognition (OCR)," *2013 Int. Mutli-Conference Autom. Comput. Commun. Control Compress. Sens.*, pp. 389–393, 2013.

[4] S. Wshah, Z. Shi, and V. Govindaraju, "Segmentation of Arabic handwriting based on both contour and skeleton segmentation," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 793–797, 2009.

[5] Y. K. Chen and J. F. Wang, "Segmentation of single- or multiple-touching handwritten numeral string using background and foreground analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1304–1317, 2000.

[6] W. Xianghui, M. Shaoping, and J. Yijiang, "Segmentation of connected Chinese characters based on genetic algorithm," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 2005, no. 60223004, pp. 645–649, 2005.

[7] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, pp. 264–271, 2003.

[8] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," *Procedings Alvey Vis. Conf. 1988*, pp. 147–151, 1988.

[9] T. Kadir and J. M. Brady, "Scale, Saliency and Image Description," *Int. J. Comput. Vis.*, vol. 45, no. 2, pp. 83–105, 2001.

[10] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point

detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.
[11] N. Sebe, T. Gevers, J. Van De Weijer, and S. Dijkstra, "Corner detectors for affine invariant salient regions: Is color important?," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4071 LNCS, no. Section 4, pp. 61–71, 2006.
[12] M. S. Karis *et al.*, "Fruit Sorting Based on Machine Vision Technique," vol. 8, no. 4, pp. 31–35, 1843.
[13] W. A. Mustafa and H. Yazid, "Background Correction using Average Filtering and Gradient Based Thresholding," vol. 8, no. 5, pp. 81–88, 2016.
[14] D. Kocharyan, "A Modified fingerprint image thinning algorithm," vol. 32, no. 1, pp. 13–18, 2013.
[15] R. M. Haralick, "A Fast Parallel Algorithm for Thinning Digital Patterns," vol. 27, no. 3, pp. 236–239, 1984.