p-ISSN: 2301-5373
e-ISSN: 2654-5101

Hijaiyah Letter Segmentation Using Connected Component Labeling Method

Audini Nifira Putri[a1], I Putu Gede Hendra Suputra[a2]

[a]Informatics Department, Udayana University
Bali, Indonesia
[1]audininifirap@email.com
[2] hendra.suputra@unud.ac.id

Abstract

Arabic letters or Hijaiyah letters recognition is a challenge in itself because one letter consists of more than one character, namely the main character, companion character such as dots and lines, and punctuation called harakat. The image segmentation process is the most important in a character recognition system because it affects the separation of objects in an image. In this research, Hijaiyah letter segmentation aims to separate the letters according to the character of each letter using the Connected Component Labeling (CCL) method. Merging labels on each character will be done by looking for the Euclidean distance value from adjacent centroids. The experiment succeeded in segmenting each Hijaiyah character with an accuracy value of 86%.

Keywords: *Hijiaiyah letters, Segmentation, Connected Component Labeling, Euclidean distance*

1. Introduction

Hijaiyah letters are the constituent letters in the Quran. It is the Arabic alphabet which has a function to compose words or sentences and consists of 28 letters from '*alif'* to '*ya'*. Arabic is a language used in religious rituals in Islam. There are about 1.57 billion Muslims who use Arabic in their religious tradition [1]. There are several points of view related to the purpose of people to learn Arabic, including to learn the Quran, Islamic science, and communication [2]. These three points of view are related to the use of the Hijaiyah letters as a compiler of Arabic in writing texts, books, and the Quran.

The development of technology in the digital world today forces a lot of research related to digitizing text documents to help humans work. Pattern recognition is a branch of computer science that can provide the advantage of being able to translate text documents into digital data to be stored as images and processed as needed, such as recognition systems, translators, text-to-speech, and text mining. [3]. In the object of research, Hijaiyah letters also use pattern recognition for basics in building letter recognition systems, letter translation systems as well as other media for digitizing Arabic texts as needed.

One branch of pattern recognition is Optical Character Recognition (OCR), which is the process of converting an image of a printed or handwritten text into text encoded in machine-recognized code [4]. There have been many studies related to Optical Character Recognition (OCR) in Arabic letters, even though they are still behind when compared to research on other letter types such as Latin letters. This lag is due to several factors [1], including the recency of the related journal, the lack of books that discuss AOTR (Arabic Optical Text Recognition), and others.

In general, the character recognition system through the process of segmentation, feature extraction, and classification [5]. The image segmentation process is the most important in a character recognition system because it will divide the image into several parts according to their respective objects. Segmentation also means separating the background from objects in image data. Segmentation will be a challenge in itself in research related to Hijaiyah letters because one letter consists of more than one character, namely the main character, companion character such as dots and lines, and punctuation called harakat. Segmentation is important in

recognition, translator, and text-to-speech systems, especially for data in the form of text documents. This data usually has more than one character or object which are related to form a word or sentence and can be recognized and translated through processing each character or object of the letter. The importance of this segmentation process requires research to use right and precise methods to get maximum results.

One of the segmentation methods that can use in separating objects is Connected Component Labeling (CCL), which is a method used to classify objects in a digital image. This method applies pixel connectivity theory where all pixels on the character or object are connected or called connected if they obey adjacency rules or pixel "*proximity*" [6]. Connected Component Labeling (CCL) is an algorithm with a concept in which related components will be uniquely labeled based on a given heuristic value to cut the image into several single images by separating the background from the object to be examined [7]. This research is to segment images that have more than one Hijaiyah character so that they can separate images based on letters according to the constituent components of each letter character.

Other researchers have conducted similar research using the Connected Component Labeling (CCL) method, namely the image of the Javanese script character [3] and the Carakan Madura script image [8]. Research [3] succeeded in solving the problem of cutting a letter in the image by separating each character in a text document that has more than one main character component. The segmentation process by calculating centroids on adjacent labels using Euclidean distance results in an accuracy value of 93.26% using Intersection over Union (IoU) accuracy. Whereas in the research [8], there were 119 words from a total of 150 words that could be segmented using the Connected Component Labeling (CCL) method with an accuracy value of 79%.

2. Research Methods

This research is divided into several steps of the process, including data collection, preprocessing, segmentation, and testing. Data collection is the step of collecting primary data in the form of scanned images of Hijaiyah letters consisting of more than one abjad in Arabic language learning books. The next step is preprocessing, which is a stage that has two sub-stages, namely Grayscaling and Binarization, to produce a binary image or black and white image. After getting the binary image, the Connected Component Labeling (CCL) method will segment the images to separate the characters in one line.

The Connected Component Labeling (CCL) method will separate each object in the image, which means it will divide it into several parts. It is because one Hijaiyah letter consists of more than one character, namely the main character, companion character such as dots and lines, and punctuation called harakat. So, to get Hijaiyah letter segmentation, a rule will apply based on the centroid position label and the Euclidean distance calculation from each object that already has a label to find other components that accompany to the main Hijaiyah character. The final step in the research is testing to calculate the correct results in the research and get an accuracy value. The following is a research design on Hijaiyah character segmentation:

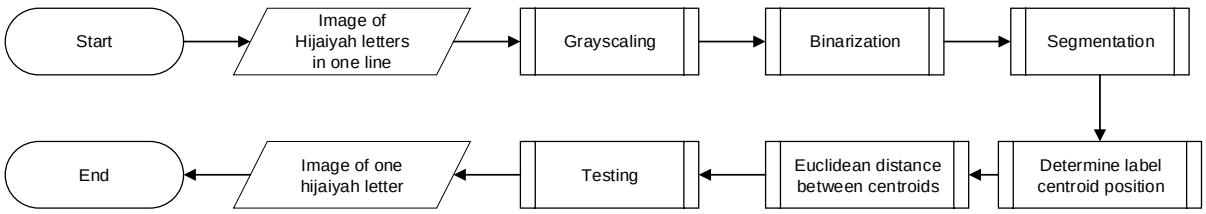


Figure 1. Research Design Flowchart

2.1. *Data Collection*

This research will use data in the form of scanned images of Hijaiyah letters taken from the Hijaiyah letter learning book, namely Iqro Volume 1 by K. H. As'ad Humam. There are 50

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Hijaiyah letters divided into 10 images. The total number of objects is 135 objects taken through the image scanning process.

2.2. Grayscaling

Grayscaling is the process of converting an RGB image or color image into a gray. The RGB image is an image where each pixel has 3 color components, namely red, green, and blue. The system will receive an RGB image for the first time and then convert it into a gray. The resulting Grayscaling image has a gray level between 0 and 255, where 0 represents the black value, and 255 represents the white value.

This research will use a weighted method in the grayscaling process, namely a grayscale method that uses the concept of the human eye's sensitivity to color [9]. This method works by reducing the value of the red and blue elements for each pixel and contributing more to the green. The calculation of this method is in equation (1) [9]:

$$Grayscale = R * 0.299 + G * 0.587 + B * 0.114 \quad (1)$$

Information:

R = red color intensity

G = green color intensity

B = blue color intensity

2.3. Binarization

Binarization is the process of converting a gray image resulting from the Grayscaling process into a black and white or a binary image with pixel values of only 0 and 1. It means the pixel value in the matrix of the image is between 0 (black) and 1 (white).

The binarization method will change the gray image into a binary image using the Thresholding method. This method will determine a value called the threshold value, which is the value limit in specifying the color of the image, whether it is above the threshold value or below. The thresholding process [10] to produce a binary image can be seen in equation (2):

$$g(x,y) = \begin{Bmatrix} 1 \; if \; f(x,y) \geq T \\ 0 \; if \; f(x,y) < T \end{Bmatrix} \quad (2)$$

Information:

x = row of pixels

y = column of pixels

g(x,y) = binary image

f(x,y) = gray image

T = threshold value

2.4. Segmentation

Segmentation is the process of separating each object in an image. Segmentation is also a technique to separate the background from the objects in the image. There are many methods used in the image segmentation process, one of which is in research that will use the Connected Component Labeling (CCL) methods.

Connected Component Labeling (CCL) method is a method that applies pixel connectivity theory in which all pixels on an object are connected or called connected if they obey adjacency rules or pixel “closeness” [6]. There are two types of connectivity rules used in two-dimensional imagery in this method, namely 4-Connected Neighborhood and 8-Connected Neighborhood. The difference is that in the 8-Connected Neighborhood pixels, the proximity will be checked not only vertically and horizontally, but also diagonally. The algorithm stages in the method include the following [11]:

a. Search for each pixel starting from the row matrix then continuing with the column to find the different pixel values to called (p).
b. In pixels (p), check each neighbor pixel p by using 4-Connected Neighborhood or 8-Connected Neighborhood on the top, left, bottom, and right diagonals.
c. If the fourth or eight neighboring pixels are 0, put a new mark on pixel (p).
d. If only one of the pixels has a value of 1 then, mark the neighboring pixel in pixel (p).
e. If two or more pixels are worth 1, put one mark on the pixel (p), then marks on the neighboring pixel (p) that are worth 1 are equivalent.

2.5. Determine Label Centroid Position

The centroid is the location of the midpoint of each component label detected [3]. The result of segmentation using Connected Component Labeling (CCL) is a separate label on each object. It is because one letter consists of more than one character, namely the main character, companion character such as dots and lines, and punctuation called harakat, so by using this CCL one letter will have several different labels. The companion character and punctuation (harakat) in one letter are not related to the main character, so they don't meet the neighborliness rules in CCL and are considered as different objects. For this reason, this research will determine the centroid on each label to find the closest distance value from each centroid. With the following equation, we will get the centroid [12] (3):

$$x_c = \frac{x_1.p_1 + x_2.p_2 + \ldots + x_n.p_n}{p_1 + p_2 + \ldots + p_n}$$
$$y_c = \frac{y_1.p_1 + y_2.p_2 + \ldots + y_n.p_n}{p_1 + p_2 + \ldots + p_n} \quad (3)$$

Information:

x_c = x coordinate of centroid

y_c = y coordinate of centroid

x_n = x coordinate of -n pixel

y_n = y coordinate of -n pixel

2.6. Euclidean Distance Between Centroid

The function of knowing the coordinates of the centroid on each label is to calculate the Euclidean distance. If the Euclidean distance is less than or equal to the threshold value, then we will combine as one label. Meanwhile, if the Euclidean distance is more than the threshold value, the opposite will apply. Calculating the Euclidean distance can be found using equation (4) [3]:

$$dist\big((x,y),(a,b)\big) = \sqrt{(x-a)^2 + (y-b)^2} \quad (4)$$

Information:

(x, y) and (a,b) : adjacent centroid

2.7. Testing and Evaluation

To get the accuracy value, we will carry out tests that reflect how good the system we have created. We will use all image data as a reference in calculating the accuracy value. In research, the calculation of the accuracy value [12] can use equation (5):

$$P(N) = \frac{I_N}{N}\ x\ 100\% \quad (5)$$

Information:

P (N) = Accuracy value

p-ISSN: 2301-5373
e-ISSN: 2654-5101

IN = Amount of data successfully segmented

N = Total data

3. Result and Discussion

In the research on Hijaiyah character text image segmentation using the Connected Component Labeling algorithm, we will implement it using the Python programming language. This study will use data in the form of scanned images of Hijaiyah letters from the Hijaiyah letter learning book, namely Book Iqro Volume 1 by K. H. As'ad Humam.

We will use images with a maximum size of 500 x 150 pixels, which contains 5 Hijaiyah letters in 1 line. An example of the scan looks like the image below:

Figure 2. Original Image

The system will receive a scanned letter image to enter the Grayscaling process with a weighted method which, will produce a grayscale image and binarization with Thresholding which, will produce a binary image. We use a threshold value in the Thresholding process, which is a pixel middle value of 128. Below is each output image from the process:

Figure 3. Grayscale Image

Figure 4. Binary Image

Figure 5. Binary Image Inversions

The difference between Figure 3 and Figure 4 is in determining the binary value. In Figure 3, pixels that have a value of 1 will be black, whereas in Figure 4, the opposite. Furthermore, the binary image will be included in the segmentation process using the Connected Component Labeling method with the 8-Connected Neighborhood rule. We will represent the image of the results of labeling each object in the binary image using Connected Component Labeling through color differences as in the image below:

Figure 6. CCL Image Segmentation

The Connected Component Labeling method will divide one letter into several objects with different labels. As seen in Figure 5, in this image, the Connected Component Labeling method divides the image into 12 labeled objects. This label does not represent letters, so images cannot be cut to separate them. It is not following the objectives of the research, so we will enforce the rule by finding the Euclidean distance value from the nearest centroid to divide the image according to the letter. The following is the centroid coordinate value of each labeled object:

Table 1. The Labeled Object's Centroid Value

Label	Centroid	
	width (x)	height (y)
1	21,700	461,646
2	21,762	358,706
3	27,072	154,320
4	29,971	257,051
5	36,739	47,130
6	76,708	351,805
7	68,831	451,584
8	54,017	151,254
9	58,052	452,910
10	79,213	250,431
11	83,569	55,661
12	90,083	155,675

The data in Figure 1 has 12 objects that already have labels. Thus, we can find the centroid of each label object according to Table 1. After getting the centroid, we will look for the Euclidean distance value for each label pair based on the centroid to get the labels to be combined. The proximity of the centroids between the object labels represents that these labels are components of the same Hijaiyah letter. If the Euclidean distance value is less than the threshold value, we will concatenate the labels to cut the object based on the updated labels. The threshold value in this study is 100. It means that we will combine label pairs whose Euclidean distance value is less than 100. The threshold value is the result of dividing the width of the image by the number of letters in it. The following are the Euclidean distance values for each label pair:

Table 2. The Euclidean Distance Value of Each Label Pair

Label	1	2	3	4	5	6	7	8	9	10	11	12
1												
2	102,9											
3	307,4	204,5										
4	204,8	102,0	102,8									
5	414,8	311,9	107,6	210,0								
6	122,8	55,4	203,6	105,7	307,3							
7	48,2	104,1	300,2	198,4	405,7	100,1						
8	312,1	209,9	27,1	108,5	105,5	201,8	300,7					
9	37,4	101,0	300,2	197,9	406,3	102,8	10,9	301,7				
10	218,9	122,6	109,3	49,7	207,7	101,4	201,4	102,3	203,6			

p-ISSN: 2301-5373
e-ISSN: 2654-5101

11	410,7	309,3	113,7	208,4	47,6	296,2	398,1	100,1	378,1	194,8	
12	313,5	214,2	63,0	117,9	120,9	196,6	299,0	36,3	279,1	100,3	100,3

In the table, several values meet the requirements for the threshold value of less than 100. Based on the table above, we can see which label pairs are close to one another so that we can combine them. Label 1 is close to 7 and 9, Label 2 with 6, Label 3 with 8 and 12, Label 4 with 10, Label 5 with 11. Each label represents the number of objects in each Hijaiyah letter. It means there will be 3 Hijaiyah letters with two objects and 2 Hijaiyah letters with three objects in the image data.

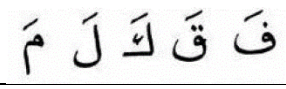
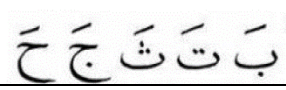
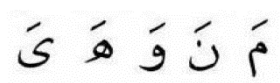
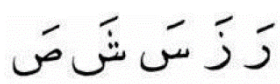
After getting the newest label, we will cut the image according to the label that we previously merged. The results of cropping images with their respective objects are as follows:

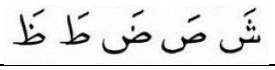
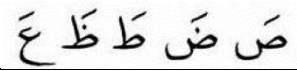
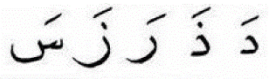
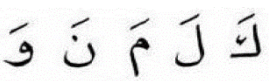
Table 3. The Results of Letter Segmentation

Letter 1	Letter 2	Letter 3	Letter 4	Letter 5
				

In Table 3, we succeeded in segmenting the letter image from 1 line into 5 different letters according to the Euclidean distance value we got with the threshold value, not 12 disparate objects. In this research, we tested 50 letters with a total of 135 objects in the image. The data that we have implemented using the Connected Component Labeling method to label objects in each image are as shown in the table below:

Table 4. The Results of The CCL Implementation

No.	Original Image	The Result		
		Color Representation of The Detected Object	Number of objects	Information
1.			14	True
2.			17	True
3.			11	True
4.			13	True
5.			14	True

6.			15	True
7.			12	True
8.			12	True
9.			15	True
10.			12	True
		TOTAL	135	

However, through testing that we have done, it turns out that there are still images that are not perfectly segmented. It is because of an error in determining the cut area of each labeled letter character. The following is an example of an imperfectly cropped image:

Figure 7. An Imperfectly Cropped Image

From a total of 50 letters consisting of 135 objects that we have tested, the number of letter characters that have successfully been labeled according to their respective letter-forming components and are properly segmented is 43 letters. The process of segmenting Hijaiyah characters by combining their letter labels results in an accuracy rate of 86%.

4. Conclusion

Based on the results of the tests carried out, the conclusion that we can draw is the process of segmenting the Hijaiyah character text image using the Connected Component Labeling method has reached an accuracy value of 86% of the 50 letters that have a total of 135 objects in it. The error in determining the cropping area for each letter character is the cause of the image failing to crop completely. The proximity of the letter components to other components is the reason for their failure.

References

[1] Anwar, K., & Nugroho, H. (2015, December). A segmentation scheme of arabic words with harakat. In 2015 IEEE International Conference on Communication, Networks and Satellite (COMNESTAT) (pp. 111-114). IEEE.

[2] Iswanto, R. (2017). Pembelajaran Bahasa Arab Dengan Pemanfaatan Teknologi. *Arabiyatuna: Jurnal Bahasa Arab*, *1*(2), 139-152.

[3] Sugianela, Y., & Suciati, N. (2019). CHARACTER IMAGE SEGMENTATION OF JAVANESE SCRIPT USING CONNECTED COMPONENT METHOD. Jurnal Ilmu Komputer dan Informasi, 12(2), 67-74.

[4] Kholimi, A. S. (2016, November). Segmentasi Citra Teks Pada Dokumen Berbahasa Arab Dengan Menggunakan Average Longest Path. In Prosiding SENTRA (Seminar Teknologi dan Rekayasa) (No. 2, pp. 202-206).

[5] Chaudhuri, A., Mandaviya, K., Badelia, P., & Ghosh, S. K. (2017). Optical character recognition systems. In Optical Character Recognition Systems for Different Languages with Soft Computing (pp. 9-41). Springer, Cham.

[6] Yudhistiro, K. (2017, September). Menghitung Obyek 2D Menggunakan Connected Component Labeling. In Seminar Nasional Sistem Informasi (SENASIF) (Vol. 1, No. 1).

[7] O. Salem, "*Connected Component Labeling* Algorithm," Code Project, no. (Segmentation)., 2014.

[8] Farid, M., Santoso, J., & Setyati, E. (2020). Handwritten Image Segmentation Carakan Madura Based Projection And Connected Component Labeling. JOINCS (Journal of Informatics, Network, and Computer Science), 4.

[9] D. Salomon, "The Computer Graphics Manual," Springer-Verlag, vol. 42, 2011.

[10] Yanti, C. P., Aristamy, I. G. A. A. M., & Pascima, I. B. N. (2020). PELABELAN HURUF PADA PRASASTI TEMBAGA MENGGUNAKAN THINNING STENTIFORD DAN CONNECTED COMPONENT LABELLING. Jurnal Pendidikan Teknologi dan Kejuruan, 17(2), 220-230.

[11] R. C. W. R. E. &. E. S. L. Gonzalez, "Digital Image Processing Using Matlab - Gonzalez Woods & Eddins.pdf. Education.," https://doi.org/10.1117/1.3115362, 2004.

[12] R. K. F. Provost, "Glossary of Terms," Journal of Machine Learning, pp. 271-274, 1998.

This page is intentionally left blank