ARABIC CHARACTER RECOGNITION USING LEARNING VECTOR QUANTIZATION

Keumala Anggraini[1], Lestari Handayani[2]
Informatics Engineering Department
Faculty of Science and Technology
Universitas Islam Negeri Sultan Syarif Kasim Riau
Pekanbaru - Indonesia
Email :[1]*keumalaanggraini93@gmail.com*, [2] *lestari.handayani@uin-suska.ac.id*

***ABSTRAK* – Arabic character recognition should be research again. Arabic character have 28 characters with 4 different positions in sentence, then Arabic character has 28-100 characters. The method is used for Arabic character recognition is learning vector quantization neural network. It is because, the learning vector quantization could classify input in category defined on training network. The objective of this study is to testing LVQ method in Arabic characters recognition. The experiment conducted using all types' position of character in sentence, there are isolated, begin, middle, and end. The testing data of Arabic character passed preprocessing phase to get vector number that was the size of matrix is used as input for learning vector quantization. The size of matrix was 8x10 for isolated, middle; end and 7x12 for begin. The success accuracy rate for isolated was 76, 43%, begin was 65, 45%, middle was 62, 73%, and end was 80%. The success accuracy percentage for all Arabic character was 72%.**

Keywords: Arabic Character, Character Recognition, Learning Vector Quantization, Size of Matrix, Success Accuracy.

I. INTRODUCTION

Pattern recognition is a technique used to classify an object by main feature or main properties [1]. Pattern recognition done by identifying pattern in an object on one of group. Pattern recognition can be done on characters of letters.

A letter is smallest unit or smallest information from a sentence that needs to be defined so that information on a sentence could understand. A letter have different form between one another, to distinguish it adapted characteristic one of the letters [2]. The letters have differences pronunciation, although the letters have same characteristic. One of the letters has difference characteristic and pronunciation is Arabic character.

Arabic characters have 28 difference characters between one another. Arabic characters have difference characters in accordance with position on a sentence. Parts of Arabic characters have 4 form differences position on sentences, another part of Arabic characters have 2 form differences position on sentences. Therefore, Arabic characters have 28-100 characters, with addition of changes in the form characters. Difference characters will change if the characters on isolated, begin, middle and end [3].

Learning vector quantization (LVQ) was introduced by Tuevo Kohonen, who also introduced Kohonen method. LVQ is one of artificial neural network and supervised network. LVQ classify input in category that has been defined through supervised network (Putra, 2010). On research which title Difference between Kohonen Neural Network and Learning Vector Quantization on The Real Time Handwritten System have purpose that recognition with LVQ better than Kohonen in terms of accuracy.

Another research entitle Analysis and Implementation of the Kohonen Neural Network for Arabic Character Recognition [5], featuring error result of experiment kohonen neural network. On that research, the error result was 43, 64% on begin of Arabic characters. In this study, used LVQ method decrease the error of Arabic character recognition. Based on [5] and this study same in Arabic characters, there were begin, middle, end, and isolated. But different in site of matrix image of Arabic characters. In this study found best practice for matrix image have various size. It was 8x10 for isolated, middle, and end, 7x12 for begin of Arabic characters.

II. CHARACTER RECOGNITION

Almost all type of character recognition use the same phases to recognition the characters. Generally, procedure to character recognition is used by many researches [5].

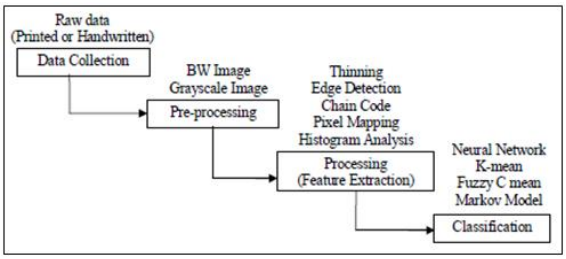

Fig 1. Character Recognition Phases [5]

On this research, Arabic character recognition using learning vector quantization phases in figure 2:

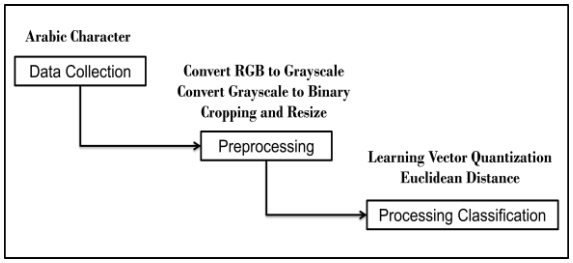

Fig 2. Arabic Character Recognition using Learning Vector Quantization Phases

The necessary data collection is handwritten Arabic character on canvas area on system and the character stored in image. Next, the processing phase, the step is cropping, conversion image to gray scale type, conversion into binary and resize. Next, the processing classification phase is process used learning vector quantization and Euclidean distance.

II. ARABIC CHARACTER

Arabic character's features [6]:

1. Arabic character have 28 characters and written right to left.
2. Arabic character didn't use different between uppercase and lowercase.
3. The width of each other is different.
4. Arabic characters have 4 difference forms according to isolated, begin, middle and end.

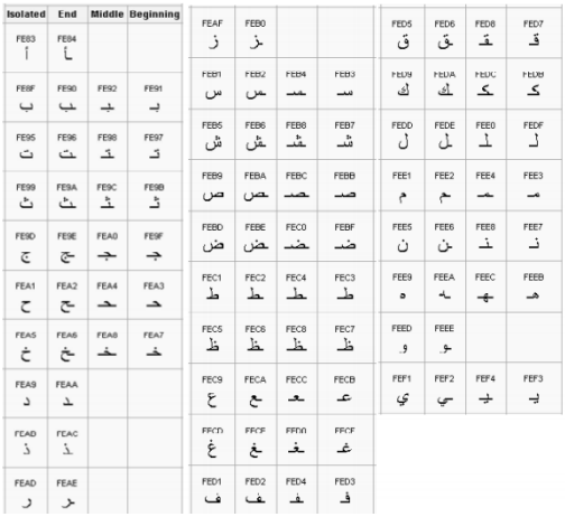

Fig 3. Arabic Character [5]

III. ANALYSIS

The analysis is the process analysis. The process analysis is analysis steps and process recognition using learning vector quantization, there are data collection, pre-processing and processing classification.

3.1 Data Collection

Data collection is the input which will be incorporated to do recognition. Input stored in image. The input is Arabic character.

3.2 Pre-processing

Pre-processing is the process change colors in image and separation between object and character. Pre-processing phase is conversion to gray scale type, conversion into binary, cropping and resize.

Conversion to Grayscale

Calculation to obtain the value for a pixel gray scale is as:

$$\text{Grayscale} = (0{,}2989 * \text{R}) + (0{,}5870 * \text{G}) + (0.1141 * \text{B}) \quad \text{..... (1)}$$

Conversion into Binary

Calculation to obtain the value for a pixel gray scale is as:

$$g(x,y) = \begin{cases} 1, & \text{jika } f(x,y) \geq T \\ 0, & \text{jika } f(x,y) < T \end{cases} \quad \text{..... (2)}$$

With g(x,y) is the binary from grayscale type and T is middle value.

Cropping and Resize

Cropping step is separation process object and background in image. Resize step is conversion process size of image matrix into m*n size of matrix. On resize step, the used matrix is 8x10 for isolated; middle and end, 7x12 for begin.

Fig 4. Character Sheen (ش)

Sheen (ش) has passed cropping step, conversion to grayscale and conversion into binary, sheen (ش) must pass resize step. Size of matrix sheen (ش) is 8x10, it's because, and sheen (ش) is isolated.

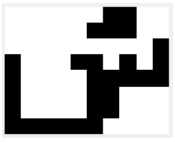

Fig 5. 8x10 of Sheen (ش)

Binary representation from resize for sheen is:

Table 1. Binary of 8x10 Sheen (ش)

1 [1,1]	1 [1,2]	1 [1,3]	1 [1,4]	1 [1,5]	1 [1,6]	**0 [1,7]**	**0 [1,8]**	1 [1,9]	1 [1,10]
1 [2,1]	1 [2,2]	1 [2,3]	1 [2,4]	1 [2,5]	**0 [2,6]**	**0 [2,7]**	**0 [2,8]**	1 [2,9]	1 [2,10]
1 [3,1]	1 [3,2]	1 [3,3]	1 [3,4]	1 [3,5]	1 [3,6]	1 [3,7]	1 [3,8]	1 [3,9]	**0 [3,10]**
0 [4,1]	1 [4,2]	1 [4,3]	1 [4,4]	**0 [4,5]**	**0 [4,6]**	1 [4,7]	**0 [4,8]**	1 [4,9]	**0 [4,10]**
0 [5,1]	1 [5,2]	1 [5,3]	1 [5,4]	1 [5,5]	**0 [5,6]**	**0 [5,7]**	**0 [5,8]**	**0 [5,9]**	**0 [5,10]**
0 [6,1]	1 [6,2]	1 [6,3]	1 [6,4]	1 [6,5]	**0 [6,6]**	**0 [6,7]**	1 [6,8]	1 [6,9]	1 [6,10]
0 [7,1]	1 [7,2]	1 [7,3]	1 [7,4]	1 [7,5]	**0 [7,6]**	**0 [7,7]**	1 [7,8]	1 [7,9]	1 [7,10]
0 [8,1]	**0 [8,2]**	**0 [8,3]**	**0 [8,4]**	**0 [8,5]**	**0 [8,6]**	1 [8,7]	1 [8,8]	1 [8,9]	1 [8,10]

3.3 Processing Classification

Processing classification is learning vector quantization and Euclidean distance.

Learning Vector Quantization

Learning Vector Quantization (LVQ) is supervised network which training of competitive layer. Competitive layer will learn automatically to classify input [7].

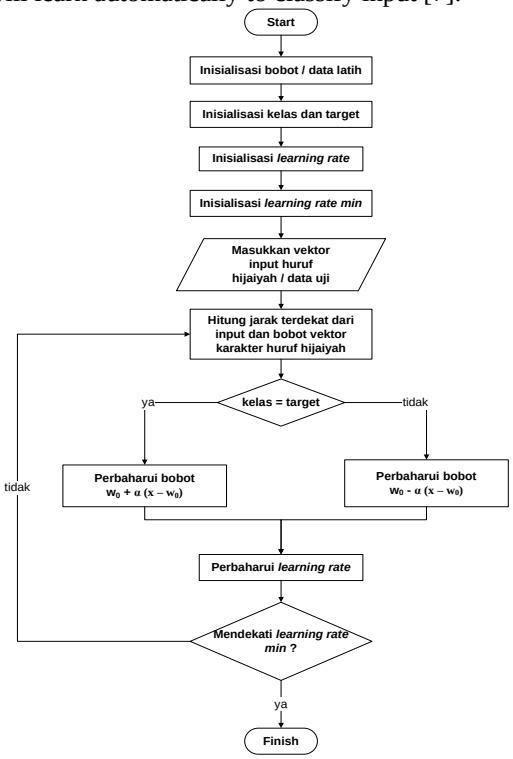

Fig 6. Flowchart of Learning Vector Quantization for Arabic Character Recognition

LVQ algorithm [1]:

LVQ Algorithm LVQ is [1] :

1. The first is determined each output, width and learning rate.
2. Compare each input and each width b measuring the distance between width (w0) and input (xp).
3. Distance measurement is calculated using Euclidean distance.
4. Minimum value from comparison result will determine class from vector input and change in weight from that class vector input.
5. Changes in new weight calculated.

Calculation for input and width has same class.

$$w_0' = w_0 + \alpha(x - w_0) \quad \text{..... (3)}$$

Calculation for input and width has same class.

$$w_0' = w_0 - \alpha(x - w_0) \quad \text{..... (4)}$$

Euclidean Distance

Euclidean distance is done to get smallest distance from new width used LVQ. Smallest distance from result will get pattern resemble input. Calculation Euclidean distance is:

$$j(v_1, v_2) = \sqrt{\sum_{k=1}^{N}(v_1(k) - v_2(k))^2} \quad \text{..... (5)}$$

IV. Implementation and Result

Learning Vector Quantization for Arabic character recognition implemented by processor Intel core i3 2, 4 GHz, RAM 2GB, hard disk 320 GB, operation system Windows 7, and programming Matlab 2010a, which used with not network.

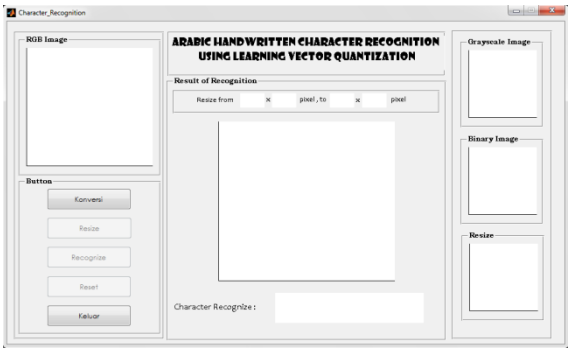

Fig 6. Character Recognition Interface

Character will be drawing on canvas interface. Character which drawing on canvas is Sheen (ش). Sheen (ش) will pass pre-processing and processing classification.

Fig 7. Interface Character Recognition Sheen ((ش)

Each Arabic character will be tested. The tested character is isolated, begin, middle and end. The experiment is Arabic character will be tested by 5 characters each Arabic character.

Table 2. The Experiment and Result Arabic Character Recognition

Character Type	Character	Tested Character	Result	Percentage (%)
Isolated	ا	أ	ا	100
		ا	ا	
		أ	ا	
		ا	ا	
		ا	ا	
	. . .	. . .	. . .	. . .
	ي	ي	ي	60
		ي	ي	
		ي	ي	
		ي	ز	
		ي	س	
Begin	بـ	بـ	بـ	100
		بـ	بـ	
		بـ	بـ	
	. . .	. . .	. . .	. . .
	يـ	يـ	يـ	80
		يـ	كـ	
		يـ	يـ	
		يـ	يـ	
		يـ	يـ	
Middle	بـ	ـبـ	بـ	100
		ـبـ	بـ	
		ـبـ	بـ	
		ـبـ	بـ	
		ـبـ	بـ	
	. . .	. . .	. . .	. . .
	يـ	ـيـ	بـ	20
		ـيـ	بـ	
		ـيـ	بـ	
		ـيـ	بـ	
		ـيـ	يـ	
End	ا	ـا	ا	100
		ـا	ا	
		ـا	ا	

Character Type	Character	Tested Character	Result	Percentage (%)
		ﺎٔ	ا	
		ﺎٕ	ا	
	. . .	. . .	. . .	. . .
	ي	ﻳ	ي	40
		ﻳ	ج	
		ﻳ	ي	
		ﻳ	ف	
		ﻳ	ف	

The experiment used size of matrix tested, and then percentage calculation of success and error from the experiment is calculated. The calculation for success percentage:

$$\frac{\text{Number of Success}}{\text{Number of Tested}} \text{x} 100\% \quad \text{..... (5)}$$

And the calculation for error percentage:

$$\text{Error} = 100\ \% - \text{Success Percentage} \quad \text{..... (6)}$$

Table 3. Percentage of Result Arabic Character Recognition

Character Type	Number of Tested Character	Number of recognize Character	Percentage (%)	Error (%)
Isolated	140	107	76,43	23,57
Begin	110	72	65,45	34,55
Middle	110	69	62,73	37,27
End	140	112	80	20
Total	500	360		
Average			72	28

V. CONCLUSION

1. Size of matrix is 8x10 for isolated, middle, end and 7x12 for begin.
2. Learning Vector Quantization can recognize with percentage rate of success is 76, 43 for isolated, 65, 45 for begin, 62, 73 for middle, 80% for end. An average success is 72% for 500 tested Arabic characters.
3. LVQ method for classification in Arabic characters more less error than [5]. Shown from table 3 wuth average error 28%.
4. The high percentage from every tested percentage is end character. It's because, end character have clear differences each other.

REFERENCES

[1] Putra, Darma. *Pengolahan Citra Digital.* Yogyakarta: Penerbit Andi. 2010.

[2] P, Tacbir Hendro, Agus Komaruddin, dan Dila Fadhilah. "Pengenalan Pola Huruf Arab menggunakan Jaringan Saraf Tiruan dengan Metode Backpropagation.", *Seminar Nasional Teknologi Informasi dan Komunikasi (SNASTIKOM 2012).* 2012.

[3] Batawi, Yusof A, dan Osama A Abulnaja. "Accuracy Evaluation of Arabic Optical Character Recognition Voting Technique : Experimental Study ." *International Journal of Electrical & Computer Sciences IJECS-IJENS Vol: 12 No: 01*. 2012.

[4] Asworo. "Perbandingan antara Metode Kohonen Neural Network dan Learning Vector Quantization pada Sistem Pengenalan Tulisan Tangan secara Real Time.". 2010.

[5] Handayani, Lestari, Iwan Iskandar, dan Weli Andrian.. "Analysis and Implementation of the Kohonen Neural Network for Arabic Character Recognition ." *ICoSTechS | http://www.icostechs.org*. 2014.

[6] AbdelRaouf, Ashraf, Colin A Higgins, dan Mahmoud Khalil. "A Database for Arabic Printed Character Recognition." *A. Campilho and M. Kamel (Eds.): ICIAR 2008*. 2008.

[7] Kusumadewi, Sri. *Membangun Jaringan Syaraf Tiruan Menggunakan Matlab & Excel Link.* Yogyakarta: Graha Ilmu. 2004