**learning Arabic letter and sentences based on image processing.**

Ainatul Radhiah1, Carmadi Machbub*2*, Ary S. Prihatmanto3, Egi M.I. Hidayat4

Sekolah Teknik Elektro dan Informatika, Institusi Teknologi Bandung

Jl Ganesha 10, Bandung, Indonsia

[1ainawind27@gmail.com](mailto:1ainawind27@gmail.com), [2carmadi@stei.itb.ac.id](mailto:2carmadi@stei.itb.ac.id), [3asetijadi@lskk.ee.itb.ac.id](mailto:3asetijadi@lskk.ee.itb.ac.id), [4egi@lskk.ee.itb.ac.id](mailto:4egi@lskk.ee.itb.ac.id)

|  |
| --- |
| **Abstract**  Optical character recognition (OCR) functions to scan digital images into text that is in the image. in this research will be designed a system that can recognize isolated Arabic letters and Arabic letters in a sentence. System has five stage : pre-processing, thinning, segmentation, feature extraction and classification.. In the pre-processing stage is done by binarization, the image is converted to a binary that have valued at 0 and 1. In the thinning stage is done with a stentiford algorithm that has 4 templates, end point and number of cometivities to check whether an image can be deleted or not. In segmentation stage sub-word segmentation is done by connected pixel components, and letter segmentation is done by Zidouri algorithm. In the feature extraction is done by 3 features that extracted, the first is nuber of dots, the second is the position of dots and the third is normalized chain code. In the classification stage is done by Neural Network and Hidden Markov Model. the results show that the accuracy recognition of isolated Arabic letters with the Neural Network classification method reached 100% and the recognition result of the Arabic letters in the sentence reached 75%. While the recognition results of the isolated arabic letters with hidden markov model classification method reached 74% and the recognition result of the Arabic letters in the sentence reached 50%.  Keywords: OCR, thinning Stentiford, Chain code, Neural Network, Hiden Markov Model |

1. **INTRODUCTION**

Optical Character Recognition is very useful for human life, text images that are processed with OCR can be text that can be edited and manipulated.OCR is also useful for automatic reading license plate car, sorting mails and editing documents, bank inspection process, signature checking and product identification. research on OCR has been widely developed in Latin, Japanese and Chinese character. but little research on Arabic. Therefore in this research will be developed Arabic OCR system.

Arabic language is used by more than 1 billion people in the world. Arabic has 28 letters that written right to left and and written cursively both printed text or handwriting. Therefore the recognition of Arabic letters in sentences requires a segmentation process. Some Arabic letters have a similar shape and can be distinguished from the number of dots and the position of the dots. Each Arabic letter has a different shape, depending on it’s position in the sentence, that is isolated, at the beginning, in the middle and at the end.

1. **REASEARCH SYSTEM**

This research was conducted in 5 stages, the following is an explanation of each stage that is implemented.

**2.1. Binarization**

Binarization of the image is the process of converting the image into binary that have values 0 and 1. Grayscale image will be changed to black and white. A binarization process is required to perform the next steps on the introduction of Arabic letters and sentences. The way it does is by doing threshold on each color channel. The threshold used is 150. If the color channel is less than 150 it will be converted to black, and if the color more than 150 will be changed to white.

**2.2 Thinning**

One of the uses of thinning is in the pattern recognition application. The image used is the image that has been done binaryization first so that the image becomes binary image. This process erodes the pixels as much as possible without affecting the general shape. After thinnning process the pattern should still be recognized. The resulting image of the thinning algorithm is called the skeleton. In this research thinning algorithm used is Stentiford algorithm. This algorithm uses a set of four templates to scan the image, say T1, T2, T3, and T4 as shown in figure 1.



Figure 1. Templates of Stentiford Algorithm

Here are the steps to get the skeleton of an image with the Stentiford algorithm

Thinning stage of the Stentiford algorithm:

1. Initially locate the pixel (i, j) that matches the T1 template. With T1 template all pixels at the top of the image will be deleted. Matching this template moves from left to right and from top to bottom.
2. If the middle pixel is not an endpoint and has a number of connectivity = 1, then mark pixels for later deletion.

The endpoint is a pixel which is the end limit and is only connected 1 pixel only. That is, if the black pixel has only one black neighbor of the eight possible neighbors.

The number of connectivity is a measure of how many objects are connected to a certain pixel. Here is the formula to calculate the number of connectivity.



Where:

Nk is the value of the 8 neighbors around the pixels to be analyzed, and the value S = {1,3,5,7}

N0 is the value of the middle pixel.

N1 is the value of the pixel on the right of the central pixel and the rest are numbered sequentially in the opposite direction of the clock

1. Repeat steps 1 and 2 for all pixels that match the T1 template.
2. Similarly follow the above mentioned steps 1-3 for the templates: T2, T3, and T4.
3. Template T2, T3, and T4 matches the left, bottom, and right side of the image.
4. Pixels marked for deletion are set to white



Figure 2. Result of Stentiford Thinnning algorithm

* 1. **Segmentation**

There are two kinds of segmentation done in this research, namely subword segmentation and letter segmentation. Sub-word segmentation is done by using the connected pixel component, the method used in this segmentation is the method described by amin [13]. After getting connected pixel components, each component is grouped into one of three groups: the main body, the secondary object and the noise. The letter segmentation was done using the Zidouri algorithm [4]. There are no modifications in the implementation.



Figure 3. Results of Zidouri letter segmentation

* 1. **Feature Extraction**

In this research there are 3 stage feature extraction, that is number of dots, position of dots and chain code. Here is an explanation of each feature:

* + 1. **Chain code**

In pattern recognition, chaincode is a technique to describe a structure of an object. Chain code is obtained by tracing the pixels of the object boundary based on predetermined directions. The result of the chain code is the numbers that indicate the direction that represents the boundary of the object. Chain code can only be done on binary image.

Here is how to extract the chain code of an object in an image:

1. Find a black pixel that has only 1 neighbor by tracing the pixels in the image starting from the top left corner until it finds a black pixel that has 1 neighbor, if not found a black pixel that has only 1 neighbor then grab the first black pixels encountered.
2. 2. Do iteration on the image

1. Change the current pixel to 0

2. Follow the priority of directions 1 to 8

3. Move the pixel position

4. Append direction to the chain code

The length of the chain code of an object changes according to the shape of an object. In this research will be classified with Neural Network and Hidden Markov Model. The input of the neural network must be fixed, not changeable. Therefore in this research will be normalized chain code. That is, making the length of the chain code of a fixed image in number and not changing.

* + - 1. **Normalized Chain Code**

In this research the chain code of the object will be normalized to 10 for each object of the letter image. steps 1 and 2 follow the steps developed by Izakian, and steps 3 and 4 were developed in this research

Here are the steps of chain chain normalization:

1. Chain code is converted into 2 dimensional matrix. The first line is the value of the chain code. The second line is the frequency of occurrence of each number in the chain code.Like the following chain code: 7777311122222583353333, After the first stage of chain chain normalization will be 2 x 9 matrix:

7 3 1 2 5 8 3 5 3

4 1 3 5 1 1 2 1 4

1. Eliminate all values that have only 1 frequency.

7 3 1 2 5 8 3 5 3 7 1 2 3

4 1 3 5 1 1 2 1 4 4 3 5 6

1. Show chaincode according to frequency of occurrence: 777711122222333333
2. Perform chaincode mapping to 10chaincode :

N3 : 777711122222333333

N4[i] = n3[round(i/9 x N3length-1)

Normalized chaincode is :

7711222333

* + 1. **Number of Dots**

The feature of the number of dots is an important feature in Arabic letters, since some Arabic letters have the same shape but are only differentiated by the number of dots. Such as character ب, ث, and ت.

The number of dots obtained by iterating on the image of the letter from the top left corner to the right, then to the bottom, if found the first black point calculate the chain code. Letters that have a point will have more than 1 chaincode. Then it will be checked, if the chaincode found less than 7 then it will be calculated as chaincode point, and do summation of the number of points. If the chaincode has a length greater than 7 it will be counted as the chaincode of the letter body.

* + 1. **Position of Dots**

position of dots is an important feature of Arabic letters. Some Arabic letters have the same shape and number of dots, but are distinguished by the position of dots.

The point position is obtained by calculating the first black pixels present in the image called YPos, then the image will be divided into 5 parts. If YPos is in the position of less than 2/5 the image height then the position of the point is above which is represented with the number 0. If YPos is in position less than 3/5 the height of the image then the position of the point is in the middle which is represented with the number 1. If YPos is on position of more than 3/5 image height then position of point is below which represented with number 2

**2.5 Classification**

The classification stage is performed with Neural Network and Hidden Markov Model (HMM).

**2.5.1 Classification Using Neural Network**

Artificial neural networks are a computational system whose network structure mimics the human nervous system in order to produce responses and behaviors such as biological neural networks.

Information processing systems on artificial neural networks have similar features of bilogical nervous tissue.

How simple the Neural Network works compared to the biological neural network:

1. Processing of signals or information occurs in neurons
2. Signals are sent between the neurons through a link, the dendrites and axons
3. Liaison between neurons has a weight that will strengthen or weaken the signal
4. Each neuron has an activation function that serves to determine the output of a neuron, whether the signal will be forwarded to another neuron or not.

x1

w1

x2

w2

Y

w3

x3

w4

x4

Figure 4. Neuron Models

Neurons are the principal information processing units of artificial neural networks that act on the impulses they receive and are transmitted to other neurons.

Neurons consist of 3 main elements, namely:

1. Group of of units connected by path
2. The summing unit that sums the input signal already multiplied by its weight.
3. The activation function used to determine the output of a neuron, ie determining whether the signal from the neuron input will be forwarded to another neuron or not.

In this research type of network architecture of neural network that used is network plural layer with 1 hidden layer. The activation function used for the hidden layer is the sigmoid activation function, and the activation function used for the output layer is the softmax activation function. Input of neurons for each sample is 12, the first neuron is the number of points, the second neuron is the position of the point and the third neuron is the chaincode that has been normalized.

**2.5.2 Classification using Hidden Markov Model**

HMM is a stochastic process in which one process can not be observed (hidden). This unobservable process can only be observed through a process that can be observed. [10]

Basically HMM consists of 3 things:

1) Evaluation

Evaluation is the process of calculating the probability of the observation sequence on the HMM model. Evaluation using forward and backward algorithms [11].

* Forward Algorithm

If the forward variable α\_t (i), at time t and state i, then the equation

(1)

With O = the matrix index is observed Resolution with n circumstances and observations until Q is iterated

**Initialization :**

, (2)

Where π = initial state matrix and b\_i (O\_1) = the first observed matrix.

**Induction** :

… (3)

Where n\_i = the number of states and a\_ij = transition matrix.

**Termination :**

(4)

* Backward Algorithm

The state flows backward from the last observation at t. The backward probability equation β\_t (i) is as follows:

(5)

And analogous to forward procedure β\_t (i) with step:

**Initialization**

(6)

**Induction**

,

,

(7)

2) Decoding

Decoding is done to find the best state of observation sequence in HMM model with viterbi algorithm. The steps are as follows.

Initialization:

, (8)

**Rekursi :**

, (9)

**Terminasi :**

(10)

Where P = Probability

1. Parameter Estimation (Learning)

Baum - Welch algorithm performs learning to obtain parameters on the HMM model.

**Parameter A**

, Odds when state j at time t + 1 if at time t is in state I.

= Probability of transition from state I to state j

*P*  = Probability

*n* = The number of hidden states in the model

**Parameter B**

dan

= Probability of observation matrix distribution

*m* = The number of different observation symbols in each state

In continuos density HMM is often characterized by density function or mixed density function in each state [13,14]. Assuming the use of Gaussian Mixture, the state emission density 𝑗 is defined as:

(11)

Where K is the number of mixture and w\_jk is the mixing coefficient for the Gaussian k state of state j with the stochastic constraint:

(12)

where Shows Gaussian density function with mean A covarian matrix for *mixture*.

**Initial matrix at state i**

So HMM can be denoted by

.

**IV. RESEARCH METHODS**

The Arabic letter recognition system consists of pre-processing stages, line segmentation, word segmentation, letter segmentation, thinning, feature extraction and classification.

An overview of the system is given by the following figure.



Figure 7. The main block diagram of the Arabic letter recognition system

In this research is known data of training and test data. Here's an explanation of both types of data.

1. Training data

Train data is a collection of Arabic letter data according to their position in the sentence that is isolated, at the beginning, in the middle, and at the end of which will be trained to then stored as Arabic letter set data.

2. Test data

Test data is a collection of Arabic sentence data that will be entered in the testing phase to then be incorporated into the process of segmentation. Then the feature is taken and compared with the data set of the training data and determined the similarity of Arabic letter data.

The number of Arabic letters that will be trained as much as 28 Arabic letters, 1 Arabic letters have 4 different positions, isolated, at the beginning, in the middle and at the end.

The following techniques are used in training data.

1. Type 28 letters Arabic from ا to ى using Microsoft Word, Arabic letters are written according to their position on the sentence, isolated, at the beginning, in the middle and at the end

2. Cut each letter using the Snipping Tool

3. The result of the letter cut is stored as Arabic letter set data.

**A. System Design and Classifier Experiment**

1. Training and Testing Stage with Hidden Markov Model Method

In this study using Hidden Markov Model (HMM) for classification stage. In general, block diagrams of the Arabic letter recognition process use the classification method using the Hidden Markov Model in Figure 8.



Figure 8. Arabic letter recognition process using Hidden Markov Model

1. Training phase with Hidden Markov Model method

The training phase is performed to determine the estimation parameters, so formed a hidden markov form model .

Here is a training flow chart with hidden markov modeling methods.



Figure 9. Training flow chart with the Hidden Markov Model (HMM)

1. Testing phase with Hidden Markov Model method

Testing is done to find out how big value of opportunity of image test data against compatibility with train data. This testing or evaluation process uses viterbi algorithm because the only thing to look for is the probability value of the test data in each training result database. The highest probability value determines the match between the test data against the training database. The following viterbi algorithm is used in the testing phase by the HMM method.



Figure 10. Testing flow chart with Hidden Markov Model (HMM)

1. **CONCLUSION**

According to the experimental results, we observe that

1. After being tested with 100 sentences letter segmentation using zidouri agorithm achieve 89% accuracy. The zidouri algorithm is good for this research because it can be applied to all font types and sizes.
2. Classification using Hidden Markov Model (HMM) gives 50% accuracy. HMM has low accuracy due to less discriminatory characteristics of HMM.

**REFERENCE**

[1] Safwa Taha, Y. B. (2012). Optical character recognition of arabic printed text. *IEEE Student Conference on Research and Development*, 1-3.

[2] Ismail Bouazizi, F. B.-A. (2013). Arabic reading machine for visually impaired people using TTS and OCR. *4th International Conference on Intelligent Systems, Modelling and Simulation*, 1 - 3.

[3] Iping Supriana\*, A. N. (2013). Arabic character recognition system development. *4th International Conference on Electrical Engineering and Informatics (ICEEI 2013)*, 5.

[4] Zidouri, A. (2010). On multiple typeface arabic script recognition. *Research Journal of Applied Sciences Engineering and Technology 2(5): 428-435, 2010*, 1 - 3.

[5] *Pengertian OCR dan OMR (2)*. (2016, 11 15). Retrieved from information and communication technology solution: <http://www.skill.co.id/front/index.php/jakarta-tech-update/296-pengertian-ocr-dan-omr-2>

[6] Vidya Dwi Turisqi Wijaya, S. N. (2014). deteksi huruf arab menggunakan metode freeman chain code. 2.

[7] Alginahi, Y. M. (2013). A survey on arabic character segmentation. *International Journal on Document Analysis and Recognition*, 1.

[8] Anwar Ali Sanjrani, J. B. (2016). Handwritten Optical character recognition system for shindig numerals. *International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, 3.

[9] Ningsih, S. F. (2016, 11 16). *Penipisan Pola Citra (Thinning)*. Retrieved from Sri Fitri Ningsih: http://srifitri07.blogspot.co.id/2013/12/penipisan-pola-citra-thinning.html

[10] (n.d.). Retrieved from <http://a-research.upi.edu/operator/upload/s_mat_0706660_chapter3.pdf>

[11] Devi, H. (2016). Perbandingan metode kuantisasi vector dan model markov tersembunyi pada pengenalan pembicara berbahasa Indonesia.

[12] Kapogiannopoulos, George dan Kalouptsidis, Nicholas. A Fast High Precision Algorithm for the Estimation of Skew Angle Using Moments.  
In In: IASTED, *International Conference Signal Processing, Pattern Recognition and Application*, SPPRA (Crete, Grece); 2002. pp. 275-27

[13] Adnan, Amin (2000). Recognition of printed arabic text based on global features  
and decision tree learning techniques. *Journal of Pattern Recognition Society.*

[14] M. Albadr, LN. (2013) Pengembangan sistem pengenalan huruf Arab.

<https://www.researchgate.net/publication/4027380_Offline_Arabic_text_recognition_system/figures?lo=1>

[15]izakian