Implementation of Dynamic Time Warping Algorithm on an Android Based Application to Write and Pronounce Hijaiyah Letters

Mohamad Irfan[1], Imam Zainal Mutaqin[2], Rio Guntur Utomo[3]
Department of Informatics
Sunan Gunung Djati Bandung
State Islamic University
Jl. AH Nasution No. 105, Bandung, West Java, Indonesia
irfan.bahaf@uinsgd.ac.id, imamthegemini@gmail.com, rio.guntur@uinsgd.ac.id

Abstract— Understanding Hijaiyah letters is one of the important things for Muslims to be able to read the Qur'an. Dynamic Time Warping algorithm can be used to study similarity on writing and pronunciation for Android users. Dynamic Time Warping algorithm used to calculate the ratio of existing data and data input from the user to see the resemblance. In the calculation, Dynamic Time Warping Algorithm obtains image data from the extracted images and sound using two methods. To process the image data Principle Component Analysis is used and for the voice data processing Mel Frequency Cepstrum Coefficients method is used. From the calculation of the data, it can be seen the similarities on writing and pronunciation of users who will be raised in the form of value. For comparison accuracy Euclidean Distance method is used. From this research can be found Dynamic Time Warping algorithm can be implemented to determine the similarity of the writing and pronunciation of letters hijaiyah.

Keywords— Dynamic Time Warping; Euclidean Distance; Principle Component Analysis; Mel Frequency Cepstrum Coefficients; MFCC; Hijaiyah; Android.

I. INTRODUCTION

Al-Quran shall be read and studied by a Muslim. In addition to reading, a Muslim must learn how to write and recite the verses of the Qur'an correctly. Before able to do that, a Muslim needs to know the letters in the Qur'an. The letters in the Qur'an is Hijaiyah letter consisting of twenty nine letters.

On the field research it was founded that around 74.36% kindergarten Bani Hasan and ECD Lotus have not known how to write and pronounce Hijaiyah. Based on this problem, a learning media on how to learn, write, and pronounce Hijaiyah is needed.

In the android itself there is a gesture recognition feature that serves as an identifier movements by the human body. For the voice recognition there is a speech recognition feature. In the matching of data there are several methods that can be used. Among them is the Dynamic Time Warping algorithm and Euclidean Distance.

Dynamic Time Warping function compares between the input data and training data by calculating the optimal warping path between the two data, ie the smallest distance between two data matched resemblance. Euclidean Distance is a classification method to calculate the distance between the two pieces of data, recognition obtained by calculating the closest euclidean distance, ie the value of the smallest euclidean distance.

In this case Euclidean Distance is used as the comparison method to Dynamic Time Warping in calculating the accuracy of the suitability of the data.

II. THEORETICAL BASIS

A. Hijaiyah Letters

Hijaiyah letters is the letters that is used in the Qur'an, there are 28 basic letters in the Quran also known by the name Hijaan letters [1].

B. Gesture Touch

Gesture Touch is usually used for the finger movement or shift on the sensitive touch screen on Android, which is processed using a package android.gesture [2].

C. Speech Recognition

Based on the definition coined by Tan and Lindberg (2008) which refers to the Deller's opinion, voice recognition or commonly known as the speech recognition is the process of translating the voice signal into a sequence of words [3].

D. Principle Component Analysis (PCA)

PCA is a statistical technique that has been used widely both in terms of data processing, machine learning, as well as image processing or signal processing. This method was first made by statisticians and was discovered by Karl Pearson in 1901 and used in the field of biology.

This method is basically to reduce an image into a vector characteristics. Therefore, the computation that will

be done is become less. And this will relate to the time for the introduction, and make it faster [4]. Figure 1 is a block diagram for PCA.

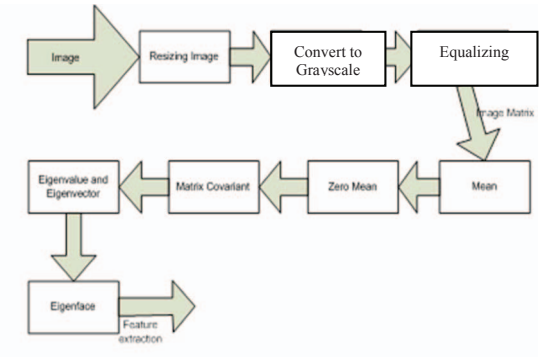


Figure 1. PCA Block Diagram

This process consists of the following stages:

1. Resizing (the process of enlargement or diminution of the dimension of the image)
2. Grayscaling (conversion of colors to grayscale phase)
3. Equalizing (correction image brightness level phase)
4. The formation of matrix data were taken from each image pixel.
5. Find the average (mean) results of the entire image using the following formula:

$$a = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3}{n}$$

6. The calculation of the average value of zero (Zero Mean) using the equation:

$$Y = \bar{x} - a$$

7. Finding the value of covariance matrix (C) by the equation:

$$C = Y^T * Y$$

8. Determine the eigenvalue matrix (d) and eigenvector matrices (v) using the equation:

$$[v, d] = eig(C)$$

9. Finding Eigenface (f) using the equation:

$$f = (Y * v)^t$$

E. Mel Frequency Cepstrum Coefficients (MFCC)

MFCC (Mel Frequency Cepstrum Coefficients) is a method that is widely used in the field of speech technology, both speaker recognition and speech recognition. This method is used to perform feature extraction, a process that converts the voice signal into multiple parameters [5]. Figure 2 is a block diagram of the MFCC.

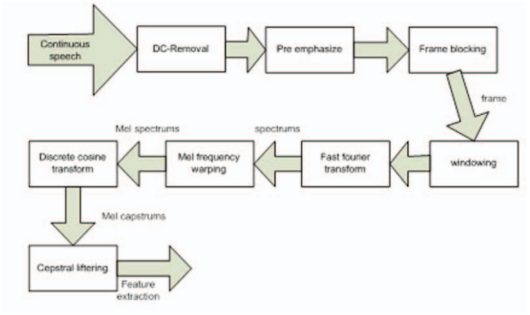


Figure 2. MFCC Block Diagram

This process consists of the following stages:

1. DC Removal aims to calculate the average of the data samples of the sound, and subtracting the value of each sampled sound with the average value.

$$y[n] = x[n] - \bar{\mathbf{x}} \ , 0 \leq n \leq N-1$$

Where:

y[n] = samples of DC removal process result signal

x[n] = samples of the original signal

$\bar{\mathbf{x}}$ = the average value of the original signal samples

N = signal length

Pre-Emphasize Filetering is a type of filter that is often used before a signal is processed further.

$$y[n] = s[n] - \alpha\, s[n-1]$$

Where:

y[n] = pre-emphasize filtering result signal

s[n] = signal before pre-emphasize filtering

In general, the value of α that most often used is between 0.9 and 1.0.

2. Frame Blocking is used because the voice signal continues to change as a result of the shift in the articulation of organ vocal production, the signal must be processed in short segments (short frame). Frames are used to calculate the amount of the formula:

Number of frames = ((I-N)/M)+1)

I = Sample rate

N = Sample point (Sample rate*framing time(s))

M = N/2

3. The next step is to windowing processing on each individual frame to minimize discontinuous signals at the beginning and the end of each frame.

$$x(n) = x_i(n)w(n) \quad n=0,1,\ldots,N-1$$

$x(n)$ = value of windowing result signal sample

$x_i(n)$ = sample value of signal frame i

$w(n)$ = window function

N = frame size, multiples of 2

Hamming window function is as follows:

$$0.54 - 0.46\cos\frac{2\pi n}{M-1}$$

Where:

n = 0,1,...,M-1

M = frame length

4. Fast Fourier Transform (FFT) acts to convert each frame of N samples from the time domain into the frequency domain. FFT is a fast algorithm for implementing the Discrete Fourier Transform (DFT) and is defined as follows:

$$S[k] = \sum_{n=0}^{N-1} s[n]e^{-j2\pi nk/N}, 0 \le k \le N-1$$

N = the number of samples to be processed (N ∈ N)

S(n) = value of signal sample

K = discrete frequency variable, which the value will be ($k = N/2, k \in N$)

5. Mel Frequency Wrapping generally done by using Filterbank. Filterbank is one form of a filter that is performed in order to determine the size of the energy of a certain frequency band in the voice signal. Here is the formula used in the Filterbank calculation:

$$Y[i] = \sum_{j=1}^{N} S[j]H_i[j]$$

N = the number of magnitude spectrum (N ∈ N)

S[j] = magnitude spectrum at the frequency j

$H_i[j]$ = coefficient filterbank at frequency j ($1 \le i \le M$)

M = the number of channels in the filterbank

6. Discrete Cosine Transform (DCT) is the last step of the main process of MFCC feature extraction. Here is the formula used to calculate the DCT:

$$C_n = \sum_{k=1}^{K} (\log S_k)\cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right]; \quad n = 1,2,\ldots,K$$

S_k= the output of the filterbank at index k

K= the expected number of coefficients

7. Results of the main MFCC feature extraction process has several drawbacks. Low order of cepstral coefficients are very sensitive to spectral slope, while the high orders are very sensitive to noise. Therefore, cepstral liftering become one of the standard techniques that can be applied to minimize the sensitivity. Cepstral liftering can be done by implementing the window function to the cepstral features

$$W[n] = \begin{cases} 1 + \frac{L}{2}\sin\left(\frac{n\pi}{L}\right) & n = 1,2,\ldots,L \\ 0 \end{cases}$$

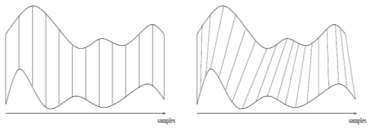
L= the number of cepstral coefficients

N= index of cepstral coefficients

F. Dynamic Time Warping

DTW (Dynamic Time Warping) is a method to calculate the distance between two times of series data. Advantages of the DTW method within the other is able to calculate the distance of two vectors of data with different lengths [6].

DTW distance between two vectors is calculated from the optimal bending lines (optimal warping path) of the two vectors. DTW matching method illustration shown in figure 3 below.



(a) (b)

Figure 3. Callibration of sequence (a) the original alignment from the two sequences (b) alignment with DTW

From some of the techniques used to calculate DWT, one of the most reliable is the method of dynamic programming. DTW distance can be calculated by the formula:

$$d(u,v) = (u-v)^2$$

$$\gamma(m,n) = d_{base}(u_i, v_j) + \min\begin{cases}\gamma(i-1,j)\\ \gamma(i-1,j-1)\\ \gamma(i,j-1)\end{cases}$$

$\gamma(0,0)=0, \gamma(0,\infty)=\infty, \gamma(0,0)=0, \gamma(\infty,0)=\infty$

$(i = 1,2,3...m;\ j = 1,2,3...n)$

Column with a value of $\gamma(i,j)(1 \le i \le m, 1 \le j \le n)$ is called cumulative distance matrix. Here is an example of the cumulative distance matrix:

	0	3	6	0	6	1
2	**4**	**5**	21	25	41	42
5	29	8	**6**	31	26	42
2	33	9	22	**10**	26	27
5	58	13	10	35	**11**	27
3	67	13	19	19	20	**15**

Figure 4. Cumulative distance matrix between 2 vectors. u={2,5,2,5,3}, v={0,3,6,0,6,1}

G. Euclidean Distance

Euclidean distance is the approach to look for cases by calculating the affinity between new cases with old cases, which is based on matching the weight of a number of features that are to be used in the process of image recognition. The following equation of euclidean distance method that works to get the value of eigen distance.

$$d_v = \sqrt{\sum_{j=1}^{m} (\text{Eigenface}_{\text{train ke j}} - \text{Eigenface}_{\text{uji}})^2}$$

III. METHODOLOGY

The methodology used to build this application is Luther method, because this method is suitable to the development of software-based multimedia, as shown in Figure 5.

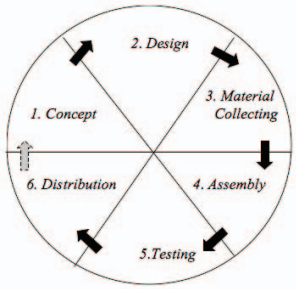


Figure 5. Multimedia Application Development Cycle According to Luther [7]

A. Concept

The purpose of this research is the creation of a learning media on how to know, write, and pronounce Hijaiyah with users' goal are kindergarten and early childhood children.

B. Design

In this phase of the software development Unified Modeling Language (UML) method is used because this application is object oriented.

C. Material Collecting

In this research, material collected by interview and literature study.

D. Assembly

The creation of this application based on the design stage that has been made in the previous stage.

E. Testing

This stage is also known as alpha testing phase and the testing performed by the researcher or the researcher's own environment. Aspect tested in this study is the algorithm of the application.

F. Distribution

At this stage the application that has been created is stored in the apk form which is the installation file for the android based applications.

IV. IMPLEMENTATION AND TESTING

A. User Interface

a. Main Menu

Main Menu display shown in Figure 6.

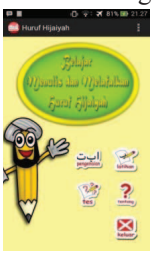


Figure 6. Main Menu display

b. Introduction

Introduction display shown in Figure 7.

Figure 7. Introduction display

c. Practice

Practice display shown in Figure 8.

Figure 8. Practice Display

d. Test

In the Figure 8 shown the Writing Test and Pronunciation Test display.

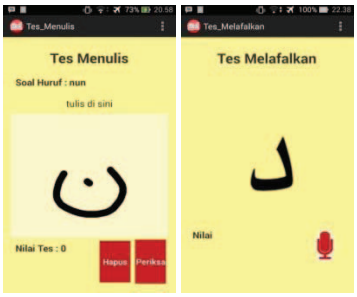


Figure 8. Writing Test and Pronunciation Test display.

B. Algorithm Testing

a. PCA and DTW Methods

In PCA and DTW methods testing, will be tested the accuracy of writing letters as much as five times by five individuals with different shape of letter on each letter. After the results of distance obtained, then compare it with the threshold value. The threshold value can be found using the formula:

$$T = \frac{Fmax - Fmin}{2}$$

T = Threshold value
Fmax = maximum distance value in the image
Fmin = minimum distance value in the image

Table I displays the test results PCA and DTW method by giving the Threshold value = $\frac{50.91-0.15}{2} = 25.53$

Distance Value < Threshold = Match
Distance Value > Threshold = Not Match

TABLE I
RESULTS OF TESTING THE ACCURACY
OF THE LETTER SHAPE WITH PCA AND DTW

No	Letter	Person Number	Distance Value	Result
1	Alif	1	5.94	Match
		2	4.32	Match
		3	5.33	Match
		4	6.92	Match
		5	5.77	Match
2	Ba	1	2.98	Match
		2	15.98	Match
		3	9.95	Match
		4	14.95	Match
		5	3.94	Match
3	Ta	1	1.36	Match
		2	13.13	Match
		3	17.45	Match
		4	29.81	Not Match
		5	27.54	Not Match

Of the 140 trials with 25.53 threshold value resulting in 122 match, 18 do not match. Therefore it can be seen the percentage of its suitability for $\frac{122}{140} \times 100\% = 87.14\%$.

b. PCA and Euclidean Distance Methods

In PCA and Euclidean Distance methods testing, will be tested the accuracy of writing letters as much as five times by five individuals with different shape of letter on each letter..

Table II shown the test results of PCA and Euclidean Distance methods by giving the Threshold value = $\frac{13.01-0.25}{2} = 6.63$

Distance Value < Threshold = Match
Distance Value > Threshold = Not Match

TABLE II
RESULTS OF TESTING THE ACCURACY
OF THE LETTER SHAPE WITH PCA AND EUCLIDEAN DISTANCE

No	Letter	Person Number	Distance Value	Result
1	Alif	1	0.42	Match
		2	1.37	Match
		3	0.48	Match
		4	1.21	Match
		5	0.77	Match
2	Ba	1	3.27	Match
		2	4.98	Match
		3	3.72	Match
		4	4.25	Match
		5	4.1	Match
3	Ta	1	4.52	Match
		2	4.82	Match
		3	5.85	Match
		4	3.81	Match
		5	5.44	Match

Of the 140 trials with 6.63 threshold value resulting in 130 match and 10 not match. Therefore it can be seen the percentage of its suitability is $\frac{130}{140} \times 100\% = 92.85\%$.

c. MFCC and DTW Methods

In the MFCC and DTW methods testing will be tested the pronounce precision of the letters five times by five people.

Table III shown the test results of MFCC and DTW methods by giving the Threshold value = $\frac{12.67-0.41}{2} = 6.54$

Distance Value < Threshold = Match
Distance Value > Threshold = Not Match

TABLE III
RESULTS OF TESTING THE ACCURACY
OF THE LETTER PRONUNCE WITH MFCC AND DTW

No	Letter	Person Number	Distance Value	Result
1	Alif	1	0.42	Match
		2	1.37	Match
		3	0.48	Match
		4	1.21	Match
		5	0.77	Match
2	Ba	1	3.27	Match
		2	4.98	Match
		3	3.72	Match
		4	4.25	Match
		5	4.1	Match
3	Ta	1	4.52	Match
		2	4.82	Match
		3	5.85	Match
		4	3.81	Match
		5	5.44	Match

Of the 140 experiment with 6.54 threshold value resulting in 100 match, 40 do not match. Therefore it can be seen the percentage of its suitability is $\frac{100}{140} \times 100\% = 71.42\%$.

Based on the results of the algorithm testing, it can be seen the results of accuracy tests performed by five people to test the 28 letters. From the test results using a different threshold value, it can be seen the accuracy of PCA and DTW comparing to PCA and Euclidean Distance methods in image matching. With the comparison shown in Table IV below.

TABLE IV
COMPARISON OF THE ACCURACY OF DTW ALGORITHM
AND EUCLIDEAN DISTANCE

Method	Percentage Value
Dynamic Time Warping	87.14%
Euclidean Distance	92.85%

Therefore it can be seen that Euclidean Distance is more accurate in matching the image data compared to DTW. As for the percentage of compatibility DTW with MFCC method on the pronunciation can be seen in Table V below.

TABLE V
MATCH PERCENTAGE OF MFCC AND DTW METHODS

Method	Percentage Value
Dynamic Time Warping	71.42%

V. CONCLUSIONS AND SUGGESTIONS

Dynamic Time Warping algorithm can be used in image and voice matching. To be able to calculate the data, image and voice input must go through the process of extraction. Principle Component Analysis (PCA) is used for the extraction of image and Mel Frequency Cepstrum Coefficients (MFCC) for the extraction of voice. In the image matching testing, the accuracy percentage of Dynamic Time Warping is 87.14% and 92.85% for Euclidean Distance. For the voice matching testing, Dynamic Time Warping accuracy percentage is 71.42%.

Euclidean Distance can be considered more accurate than Dynamic Time Warping in the image data matching. It is due to the result of distance value is smaller than the Euclidean Distance Dynamic Time Warping. And in the process of calculating matching, Euclidean Distance is shorter because the Dynamic Time Warping need to calculate the optimal warping path in advance, while the Euclidean Distance not.

Expected in the future the features of the app can be improved, such as writing letters continued Hijaiyah and additional punctuation.

And also expected to use other methods such as Edge Matching, Robert Cross, Linear Discriminant Analysis, and others to identify the similarity of image and voice with a higher degree of accuracy.

REFERENCES

[1] Gustadipura, Wana. 2014. Pembangunan Aplikasi Pelatihan Menulis Huruf Hijaiyah Berbasis Android. Universitas Komputer Indonesia. Bandung.

[2] Sakoe, H. and Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, 26(1) pp. 43– 49.

[3] Hyun Hoi James Kim, "Face Detection and Face Recognition", Survey Paper.

[4] Tan, Z. H., & Lindberg B. 2008. Automatic Speech Recognition on Mobile Devices and over Communication Networks. United Kingdom : Springer

[5] Manunggal, HS. 2005. Perancangan dan Pembuatan Perangkat Lunak Pengenalan Suara Pembicara dengan Menggunakan Analisa MFCC Feature Extraction. Surabaya : Universitas Kristen Petra.

[6] Resmawan, I Wayan Adi. 2010. Verifikasi Suara Menggunakan Metode MFCC dan DTW. Jimbaran: Universitas Udayana.

[7] Sutopo, Ariesto Hadi. 2003. Multimedia Interaktif dengan *Flash*. Graha Ilmu. Yogyakarta.