Indonesian sign language recognition using kinect and dynamic time warping

Wijayanti Nurul Khotimah, Tiara Anggita, and Nanik Suciati
Informatics Department, Institut Teknologi Sepuluh Nopember Surabaya, Indonesia

Article Info

Article history:

Received Oct 1, 2018
Revised Dec 10, 2018
Accepted Jan 25, 2019

Keywords:

Dynamic Time Wrapping
Indonesian sign language recognition
User independent system

ABSTRACT

Sign Language Recognition System (SLRS) is a system to recognise sign language and then translate them into text. This system can be developed by using a sensor-based technique. Some studies have implemented various feature extraction and classification methods to recognise sign language in the different country. However, their systems were user dependent (the accuracy was high when the trained and the tested user were the same people, but it was getting worse when the tested user was different to the trained user). Therefore, in this study, we proposed a feature extraction method which is invariant to a user. We used the distance between two users' skeleton instead of using the users' skeleton positions because the skeleton distance is independent to the user posture. Finally, forty-five features were extracted in this proposed method. Further, we classified the features by using a classification method that is suitable with sign language gestures characteristic (time-dependent sequence data). The classification method is Dynamic Time Wrapping. For the experiment, we used twenty Indonesian sign languages from different semantic groups (greetings, questions, pronouns, places, family and others) and different gesture characteristic (static gesture and dynamic gesture). Then the system was tested by a different user with the user who did the training. The result was promising, this proposed method produced high accuracy, reach 91% which shows that this proposed method is user independent.

Corresponding Author:

Wijayanti Nurul Khotimah,
Informatics Department,
Institut Teknologi Sepuluh Nopember Surabaya,
Sukolilo 60111, Surabaya, Indonesia.
Email: wijayanti@if.its.ac.id

1. INTRODUCTION

Sign Language Recognition System (SLRS) is a system which captures sign language used by people with hearing disability, extracts the features, classifies the features and then shows the result into text. The system can identify both static sign and dynamic sign. The SLRS can be developed by using either sensor-based technique or image-based technique [1]. The sensor-based technique employs a variety of sensor devices to recognise sign (human movement), e.g. Smart glove, Kinect, and Leap motion controller [2-7]. While in the image-based technique, a sequence of images was captured by using a camera. Then, those images were processed by using an image-processing method [8-10].

Many features extraction methods have been implemented in the recognition by using the sensor-based technique. First was the study conducted by Ibrahim et al. which used 28 variables from glove sensor to recognise Arabic sign language [11]. Then, Pei yin extracted 17 features of hand position. These features were thumb outside, thumb top (on a thumbnail), index finger, middle finger, ring finger, little finger, wrist perpendicular to bones, wrist parallel to fingers, wrist perpendicular to palm, shoulder elevation (forward), shoulder (outward) [12]. In 2016, Tamas et al. used atomic gestures of each skeleton to recognise gesture in

Arabic sign language while Teerawat et al. used skeleton trajectory [13-14]. In the same year, Lee et al. extracted hand positions, hand signing direction, and hand shapes to recognise Taiwanese sign language [15].

Further, the sign language recognition accuracy is influenced not only by the features but also by the classification method. During these years, some researchers had implemented different classification methods to solve this problem. Nearest Neighbour algorithm had been used by Ibrahim, Tama, and Teerawat. Their study reported that the accuracy of their system was roughly 95%. However, the studies only recognised a small number of sign languages [12, 14-15]. Another algorithm was SVM. Agarwal and Sun C had used SVM algorithm and wrote that their system accuracy was around 90%. It was lower than the previous algorithm, but the number of the recognised sign language was higher than the previous algorithm [16-18]. Hidden Markov Models was also popular. It had been used by Quadri and Ulrich. The accuracy was varying, but it could be used to recognise several numbers of sign languages [19-21]. Even though those researchers had used different classification methods, they had similarities; their system was user dependent (the accuracy was high when the trained user and the tested user were the same people, but it was getting worse when the tested user different to the trained user).

Therefore, we need to find feature extraction method which represented the sign language and invariant to the user characteristic. In the classification method, we require a classification method that suitable with sign language characteristic since sign language is a time-dependent sequence. Thus, in this research, we used skeleton distance instead of skeleton position for the feature, and we used Dynamic Time Wrapping method which suitable to classify time-dependent sequence data.

The organization of this paper is as follows. After a general introduction of the previous research in the sign language recognition, the detail about Indonesian Sign Language and Dynamic Time Wrapping were discussed in section 2. In section 3, we discussed the recording process, the feature extraction and DTW implementation. In section 4, dataset, experiment detail and analysis have been presented. And the study has been concluded in section 5.

2. REFERENCES

2.1. Indonesian Sign Language

Some different sign languages exist in the world. Each sign language in different country might has different gesture, but most of them used the hand gestures. For instance, the word ‘person’ is signed by using Indonesian sign language that is shown in Figure 1(a), but it is signed by using Brazilian sign language that is shown in Figure 1(b) [22].

Based on the gesture’s characteristics, some sign languages have a static gesture, and others have dynamic gestures. The static gesture is a gesture that does not involve a movement in its delivery as seen in Figure 2. While the dynamic gesture is a gesture that has movement in its delivery as seen in Figure 3.

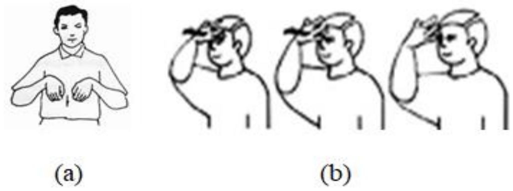


Figure 1. Sign Language of word 'person' in (a) Indonesian Sign Language (b) Brazilian Sign language

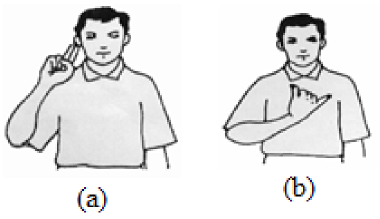


Figure 2. Sign Language with static gestures, the word of (a)’mother’ and (b)’I’ in Indonesian Sign Language

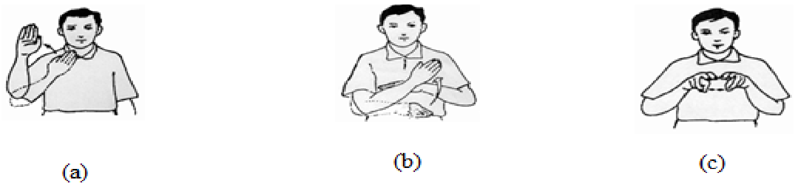


Figure 3. Sign Language with Dynamic gestures, the word of (a) ‘Hello’, (b) ‘morning’, and (c) ‘meet’ in Indonesian Sign Languae

2.2. Dynamic Time Wraping

Dynamic Time Warping (DTW) is a well-known technique for finding optimal alignment between two time-dependent sequences under a certain limit [23]. DTW is also called as a non-linear sequence alignment [24]. DTW is more realistic to use in measuring pattern matching between two time-dependent sequences than just using linear measurement algorithms such as Euclidean Distance, Manhattan, Canberra, and Mexican Hat; and widely used for speech recognition, handwriting recognition and signatures, data mining, clustering, gesture processing, and music. Further, sign language gesture is included into a time-dependent sequence. For example, the gesture of word 'hello' might be signed with slow speed by a person a; and it might be signed with higher speed by a person b. As a result, the sequence of the word 'hello' by a person a and a person b might be slightly different. Thus, an alignment between those two gestures (sequences) is required to recognize the meaning of the gestures. The example of alignment between those two sequences can be seen in Figure 4.

To compute the alignment, assume we have two time-dependent numeric sequences a (a1, a2, ..., an) and b (b1, b2, ..., bm). First, we calculate the local distance between elements of two sequences; they are calculated by finding the absolute value of the difference between the two series of data as formulated in the (1).

$$c_{ij} = |a_i - b_j|, i = 1, \dots, n \; and \; j = 1, \dots, m \tag{1}$$

Second, we find the warping path. Warping path is a path through a matrix that contains a minimal distance from an element D11 (an element in the DTW matrix) to an element Dnm. To compute the first-row element and the first-column element in the DTW matrix, we used (2) and (3), respectively. Then, to compute the other elements of DTW matrix, (4) is used. For example, the DTW matrix of training data sequence A={2,5,2,5,3,4} and testing data sequence B={0,3,6,0,6,1}was shown in Figure 5.

$$D_{(1,j)} = D_{(1,j-1)} + c_{(1,j)} \; where \; D_{(1,0)} = 0 \tag{2}$$

$$D_{(i,1)} = D_{(i-1,1)} + c_{(i,1)} \; where \; D_{(0,1)} = 0 \tag{3}$$

$$D_{ij} = c_{ij} + \min\left(D_{i-1,j-1}, D_{i-1,j}, D_{i,j-1}\right) \tag{4}$$

Since the sequence A= {2,5,2,5,3,4} and the sequence B={0,3,6,0,6,1} have the same element number, six, so the DTW matrix size is 6 × 6. Either (2), (3), or (4) is used to fill the elements in the matrix. For example, to fill an element D(1,4) we used (2).

$$\begin{aligned} D_{(1,4)} &= D_{(1,3)} + c_{(1,4)} \\ D_{(1,4)} &= D_{(1,3)} + |a_1 - b_4| \\ D_{(1,4)} &= 7 + |2 - 0| \\ D_{(1,4)} &= 7 + 2 \\ D_{(1,4)} &= 9 \end{aligned}$$

To fill the element in the first-column, we used (3). For example, an element D(2,1) was computed as follows:

$$\begin{aligned} D_{(2,1)} &= D_{(2-1,1)} + c_{(2,1)} \\ D_{(2,1)} &= D_{(1,1)} + |a_2 - b_1| \\ D_{(2,1)} &= 2 + |5 - 0| \\ D_{(2,1)} &= 2 + 5 \\ D_{(2,1)} &= 7 \end{aligned}$$

And the other element was filled by using (4). For example, an element D(4,4) was computed as follows:

$$\begin{aligned} D_{(4,4)} &= c_{4,4} + \min\left(D_{4-1,4-1}, D_{4-1,4}, D_{4,4-1}\right) \\ D_{(4,4)} &= |a_4 - b_4| + \min\left(D_{3,3}, D_{3,4}, D_{4,3}\right) \\ D_{(4,4)} &= |5 - 0| + \min(8, 6, 6) \\ D_{(4,4)} &= 5 + 6 \\ D_{(4,4)} &= 11 \end{aligned}$$

Finally, after DTW matrix element had been filled, the optimal warping cost was computed by computing the minimum total cost from D(m,n) to D(1,1). The example of warping path is shown by the red element in Figure 5. Start from the element D(m,n), we called it as the current wrapping element, the next wrapping element is min $\left(D_{m-1,n-1}, D_{m-1,n}, D_{m,n-1}\right)$. This process is done until reach element D(1,1). The optimal wrapping cost is the sum of the wrapping element.

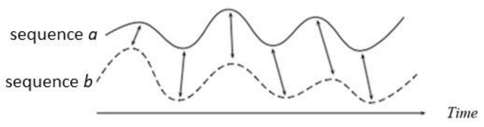
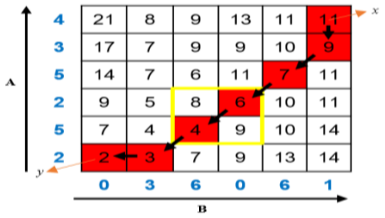


Figure 4. Example of Alignment between Sequence a and Sequence b



4	21	8	9	13	11	11
3	17	7	9	9	10	9
5	14	7	6	11	7	11
2	9	5	8	6	10	11
5	7	4	4	9	10	14
2	2	3	7	9	13	14
	0	3	6	0	6	1

Figure 5. The DTW matrix of sequence A={2,3,2,5,3,4} and sequence B={0,3,6,0,6,1}

3. RESEARCH METHOD

The proposed Indonesian Sign Language Recognition System (ISLRS) was developed by using Microsoft Visual Studio 12.0 and Kinect 2.0. The system has two stages: training and testing. The training stage has two main processes (gesture recording and feature extraction), and the testing stage has three main processes (gesture recording, feature extraction, and recognition by using DTW). The gesture recording and feature extraction process in the training and testing are the same. The difference between the training and testing are in training the extracted features are saved in the dataset together with its label, while in testing the extracted features are processed by using DTW to predict the class label. The process diagram of testing is shown in Figure 6. The detail of the process is as follows:

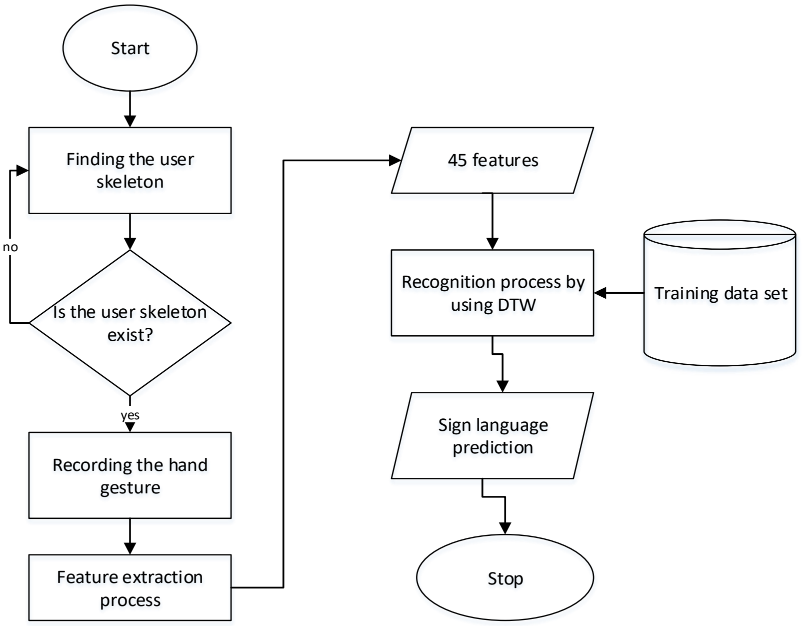

Start
Finding the user skeleton
no
Is the user skeleton exist?
yes
Recording the hand gesture
Feature extraction process
45 features
Recognition process by using DTW
Training data set
Sign language prediction
Stop

Figure 5. The Process Dyagram

3.1. Recording Process

In this system, we ask a user to stand up between 1-1.5 meters in front of the Kinect. When the system captured the user skeleton, it started to record the user skeleton position. This system would record 12 skeleton position on the axis X, Y, and Z in each frame the skeletons are shown in Figure 7, those skeletons were: Head (H), Middle Shoulder (MS), Left Shoulder (LS), Right Shoulder (RS), Left Elbow (LE), Right Elbow (RE), Left Wrist (LW), Right Wrist (RW), Left Arm (LA), Right Arm (RA), Spine (S), and Hip (Hi). In this system, we assume that one sign language gesture was expressed for 39 frames long, so we recorded the skeletons for 39 frames long. Finally, we would get 39x12x3=1.404 skeletons data for one sign language gesture see at Figure 8.

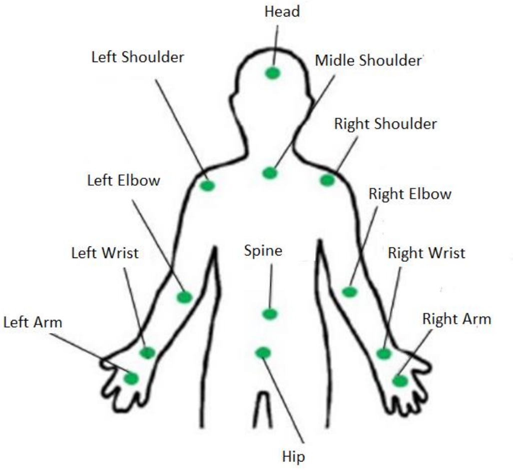
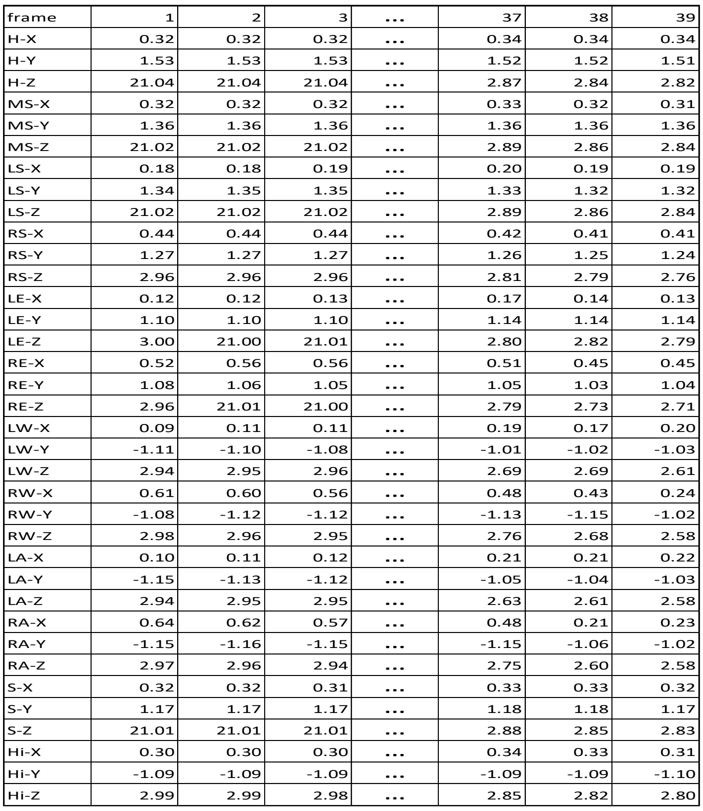


Figure 7. The Skeletons Captured by Kinect



frame	1	2	3	…	37	38	39
H-X	0.32	0.32	0.32	…	0.34	0.34	0.34
H-Y	1.53	1.53	1.53	…	1.52	1.52	1.51
H-Z	21.04	21.04	21.04	…	2.87	2.84	2.82
MS-X	0.32	0.32	0.32	…	0.33	0.32	0.31
MS-Y	1.36	1.36	1.36	…	1.36	1.36	1.36
MS-Z	21.02	21.02	21.02	…	2.89	2.86	2.84
LS-X	0.18	0.18	0.19	…	0.20	0.19	0.19
LS-Y	1.34	1.35	1.35	…	1.33	1.32	1.32
LS-Z	21.02	21.02	21.02	…	2.89	2.86	2.84
RS-X	0.44	0.44	0.44	…	0.42	0.41	0.41
RS-Y	1.27	1.27	1.27	…	1.26	1.25	1.24
RS-Z	2.96	2.96	2.96	…	2.81	2.79	2.76
LE-X	0.12	0.12	0.13	…	0.17	0.14	0.13
LE-Y	1.10	1.10	1.10	…	1.14	1.14	1.14
LE-Z	3.00	21.00	21.01	…	2.80	2.82	2.79
RE-X	0.52	0.56	0.56	…	0.51	0.45	0.45
RE-Y	1.08	1.06	1.05	…	1.05	1.03	1.04
RE-Z	2.96	21.01	21.00	…	2.79	2.73	2.71
LW-X	0.09	0.11	0.11	…	0.19	0.17	0.20
LW-Y	-1.11	-1.10	-1.08	…	-1.01	-1.02	-1.03
LW-Z	2.94	2.95	2.96	…	2.69	2.69	2.61
RW-X	0.61	0.60	0.56	…	0.48	0.43	0.24
RW-Y	-1.08	-1.12	-1.12	…	-1.13	-1.15	-1.02
RW-Z	2.98	2.96	2.95	…	2.76	2.68	2.58
LA-X	0.10	0.11	0.12	…	0.21	0.21	0.22
LA-Y	-1.15	-1.13	-1.12	…	-1.05	-1.04	-1.03
LA-Z	2.94	2.95	2.95	…	2.63	2.61	2.58
RA-X	0.64	0.62	0.57	…	0.48	0.21	0.23
RA-Y	-1.15	-1.16	-1.15	…	-1.15	-1.06	-1.02
RA-Z	2.97	2.96	2.94	…	2.75	2.60	2.58
S-X	0.32	0.32	0.31	…	0.33	0.33	0.32
S-Y	1.17	1.17	1.17	…	1.18	1.18	1.17
S-Z	21.01	21.01	21.01	…	2.88	2.85	2.83
Hi-X	0.30	0.30	0.30	…	0.34	0.33	0.31
Hi-Y	-1.09	-1.09	-1.09	…	-1.09	-1.09	-1.10
Hi-Z	2.99	2.99	2.98	…	2.85	2.82	2.80

Figure 6. Example of the data from recording process for one sign gesture

3.2. Feature Extraction Process

Characteristics that represent best the data are required to improve the accuracy of the sign language recognition. The characteristics are obtained through feature extraction process. In this study, we extracted forty-five features data from the 1.404 skeletons data received from the recording process.

Further, we extracted 45 features which were divided into four groups with the following details:

a. Group 1 (four features), the features that represent the position of the palms in the vertical plane. Here, the vertical distances between the right palm (RA-Y) and the head (H-Y) are calculated in each frame. We used Euclidian distance to compute the distance. Since the number of frames is 39, this computation produces 39 values; and the minimum of those values is used as the feature. The same process was implemented to the vertical distance between the right palm (RA-Y) and the hip (Hi-Y), the left palm (LA-Y) and the head (H-Y), the left palm (LA-Y) and the hip (Hi-Y).
b. Group 2, the features that represent the distance of the palms against other joints except for hip and head. For example, the position between the right palm (RA-X, RA-Y, RA-Z) and the middle shoulder (MS-X, MS-Y, MS-Z), the position between the right palm (RA-X, RA-Y, RA-Z) and the left elbow (LE-X, LE-Y, LE-Z), and the position between the left palm (LA-X, LA-Y, LA-Z) and the middle shoulder (MS-X, MS-Y, MS-Z). Thus, seventeen features candidate exist in each frame. The process to find the features in 39 consecutive frames are the same as the previous group by finding the minimum value. In the end, seventeen features were extracted from this group.
c. Group 3, the features that represent the statistical value of each candidate features in Group 1. In Group one, each candidate features has 39 values. Thus, the statistical value (average, median, modus, variant, and standard deviation) of each candidate feature is calculated. The number of features in this group is 20 features.
d. Group 4, the features (four features) that represent the depth position of the palm against the head and the hip. Here, the depth distance between the right palm (RA-Z) and the head (H-Z); the right palm (RA-Z) and the hip (Hi-Z), the left palm (LA-Z) and the head (H-Z), the left palm (LA-Z) and the hip (Hi-Z) are computed in each frame. Then, the process to find the features in 39 frames is the same as Group 1.

In training stage, those 45 features which represent one sign language gesture was saved into a training dataset with the label simultaneously. While in the testing stage, those 45 features were processed by using DTW.

3.3. Recognition by Using DTW

The Words and the Accuracy of Each Word on Each Scenario as shown in Table 1. The recognition by using DTW was implemented in the testing stage. Here, a sign language gesture was processed by these two previous processes producing 45 features. Thus, the optimum warping path cost between the testing gesture feature and 300 training data is computed. Since each data has 45 features, the DTW matrix size is 45x45. Then, (2), (3), or (4) is used to fill the DTW matrix. Finally, the label of the testing gesture is defined based on the label of the training data that produces the minimum warping path cost.

Table 1. The Words and the Accuracy of Each Word on Each Scenario

Semantic Category	Words	Abbreviation	Used Hand	Dynamic/ Static	Scenario		
					A(%)	B(%)	C(%)
Greetings	Hello (*Halo*)	HLO	Right	Dynamic	100	100	100
	Greeting for moslem (*Salam*)	GTG	Left, right	Static	80	80	100
	Morning (*Pagi*)	MNG	Left, right	Dynamic	100	100	100
	Meet (*Jumpa*)	MET	Left, right	Dynamic	100	100	100
Questions	What (*Apa*)	WHT	right	Dynamic	100	80	100
	How (*Bagaimana*)	HOW	Left, right	Dynamic	80	60	60
Family	Family (*Keluarga*)	FMY	Left, right	Dynamic	80	60	100
	Mother (*Ibu*)	MTR	right	Static	80	80	100
	Child (*Anak*)	CHD	Right	Dynamic	100	100	100
	Father (*Ayah*)	FTR	Right	Dynamic	100	100	80
	Brother *(Saudara*)	BTH	Left, Right	Dynamic	100	100	100
Pronouns	I (*Saya*)	I	Right	Dynamic	100	100	80
	You (*Kamu*)	YOU	Right	Dynamic	100	100	100
Places	House (*Rumah*)	HUE	Left, right	Dynamic	80	100	100
	School (*Sekolah*)	SCL	Left, right	Dynamic	100	100	100
Others	Good (*Baik*)	GOD	Right	Static	100	100	100
	Evil (*Jahat*)	EVL	Left, Right	Static	100	100	100
	Problem (*Masalah*)	PBM	Left, Right	Dynamic	100	80	60
	Work (*Kerja*)	WRK	Left, Right	Dynamic	80	100	100
	Mad (*Marah*)	MAD	Left, Right	Dynamic	100	80	100
Average (%)					**94**	**91**	**94**

4. EXPERIMENT AND ANALYSIS

4.1. Data Set

In this research, we tried to recognize 20 sign languages from various semantic category (greetings, questions, pronouns, places, family and others). The sign languages consist of 4 static sign languages and 16 dynamic sign languages. Some of them use one hand, and others use two hands. Those sign languages are Hello (‘Halo’), Greeting for muslem (‘Salam’), Morning (‘Pagi’), Meet (‘Jumpa’), What (‘Apa’), How (‘Bagaimana’), Family (‘Keluarga’), Mother (‘Ibu’), Child (‘Anak’), Father (‘Ayah’), Brother (‘Saudara’), Me (‘Saya’), You (‘Kamu’), House (‘Rumah’), School (‘Sekolah’), Fine (‘Baik’), Evil (Jahat), Problem (Masalah), Work (‘Kerja’), and Mad (‘Marah’). The gesture of those sign languages was shown in Figure 9.

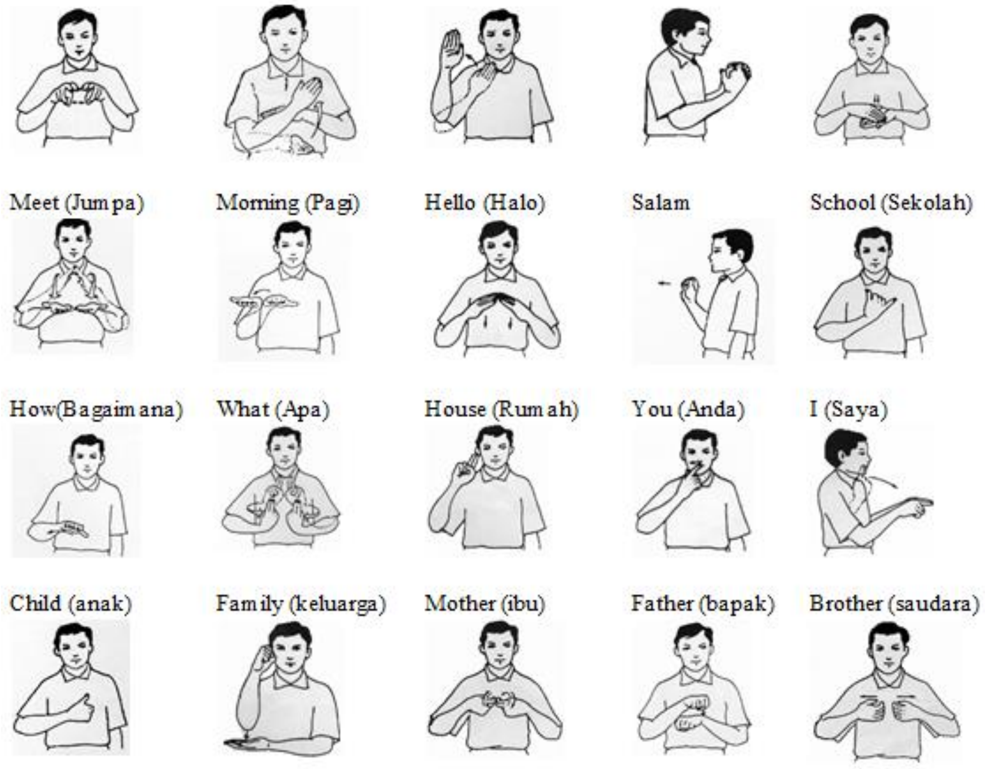


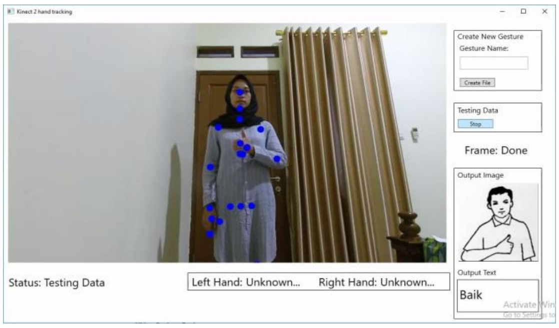
Figure 7. The sign language gesture example

4.2. Experiment

In training, we collected 15 gestures for each word. Totally, we had collected 300 gestures from a user. The testing was done in real-time. Our system recorded the gesture when the user was giving sign language and extracted the features. Then, the system computed the warping cost between the testing data and the training data. The label of the testing data was based on the label of the training data which produced the minimum warping cost. Then the predicted label and the sign language picture of the label were shown in the user interface (UI). The example of the UI is shown in Figure 10.

Also, we divided the testing into three scenarios to test whether the proposed method user-dependent or no. Those scenarios are as follows:

a. Scenario A, the sign language testing was performed by a user who had performed the training. The height of the user is 163 cm.
b. Scenario B, the testing and the training were performed by a different user. The height of the user who performed the training is 163 cm, while the height of the user who performed the testing is 173 cm. This scenario was conducted with the aim of identifying whether significant differences in height between the training user and the testing user effect on the accuracy.
c. Scenario C, the testing and the training were performed by two users. Those users have different height.
d. The total accuracy of this proposed method on every word can be seen in Table 1. While the detail of the result for five times testing on each word in scenario A, B, and C is shown in Figure 11, 12, and 13 respectively.



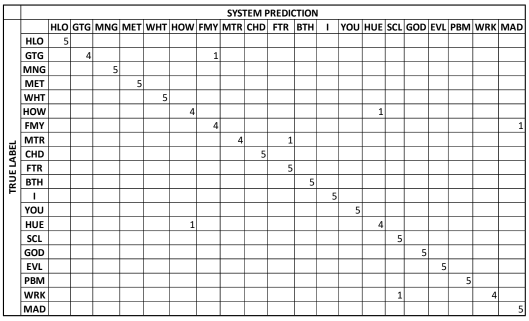
Frame: Done

Baik

Status: Testing Data

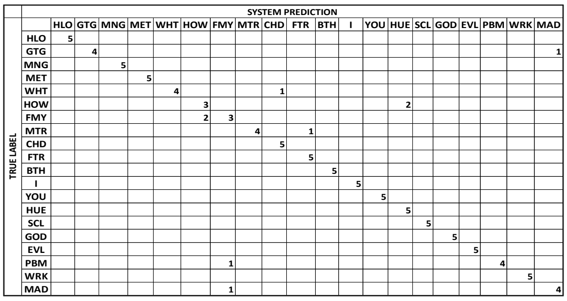
Left Hand: Unknown... Right Hand: Unknown...

Figure 8. The User Interface of the System



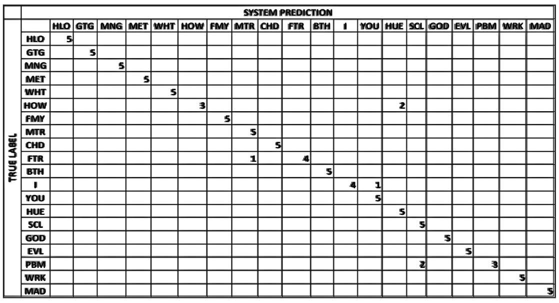
		SYSTEM PREDICTION																			
		HLO	GTG	MNG	MET	WHT	HOW	FMY	MTR	CHD	FTR	BTH	I	YOU	HUE	SCL	GOD	EVL	PBM	WRK	MAD
TRUE LABEL	HLO	5																			
	GTG		4					1													
	MNG			5																	
	MET				5																
	WHT					5															
	HOW						4								1						
	FMY							4													1
	MTR								4		1										
	CHD									5											
	FTR										5										
	BTH											5									
	I												5								
	YOU													5							
	HUE						1								4						
	SCL															5					
	GOD																5				
	EVL																	5			
	PBM																		5		
	WRK															1				4	
	MAD																				5

Figure 9. The Confusion Matrix of Scenario A



		SYSTEM PREDICTION																			
		HLO	GTG	MNG	MET	WHT	HOW	FMY	MTR	CHD	FTR	BTH	I	YOU	HUE	SCL	GOD	EVL	PBM	WRK	MAD
TRUE LABEL	HLO	5																			
	GTG		4																		1
	MNG			5																	
	MET				5																
	WHT					4				1											
	HOW						3								2						
	FMY						2	3													
	MTR								4		1										
	CHD									5											
	FTR										5										
	BTH											5									
	I												5								
	YOU													5							
	HUE														5						
	SCL															5					
	GOD																5				
	EVL																	5			
	PBM							1											4		
	WRK																			5	
	MAD							1													4

Figure 10. The Confusion Matrix of Scenario B



		SYSTEM PREDICTION																			
		HLO	GTG	MNG	MET	WHT	HOW	FMY	MTR	CHD	FTR	BTH	I	YOU	HUE	SCL	GOD	EVL	PBM	WRK	MAD
TRUE LABEL	HLO	5																			
	GTG		5																		
	MNG			5																	
	MET				5																
	WHT					5															
	HOW						3								2						
	FMY							5													
	MTR								5												
	CHD									5											
	FTR								1		4										
	BTH											5									
	I												4	1							
	YOU													5							
	HUE														5						
	SCL															5					
	GOD																5				
	EVL																	5			
	PBM															2			3		
	WRK																			5	
	MAD																				5

Figure 11. The Confusion Matrix of Scenario C

4.3. Analysis

Based on the experiments, we can see that:

a. The proposed method produced accuracy more than 91%. It is higher than previous research using K-NN which only produce 88% accuracy [25].

b. The results obtained from scenario A and scenario C are quite good, i.e. 94%. While the result obtained from scenario B is slightly lower, but it is good enough, that is 91%. This proofs that the proposed method is user independent. Even though the user who did the training different to the user who did the testing, the accuracy was not affected a lot.

c. Some sign languages have similar gesture which increased the error rate. In scenario B and scenario C, the gesture of the word 'Family' was misclassified into the gesture of the word 'House'; and the gesture of 'Problem' was misclassified into the gesture of 'School'. This problem occurred because their gestures are similar. Both gestures used two hands, dynamics (moves from top to bottom), and position of the hand is in front of the body.

5. CONCLUSION

This purpose of this study was to recognize Indonesian Sign Language. In this study, we proposed to use the distance among user skeletons for the features instead of the position of the user skeletons because the distance of the skeletons is invariant to the user posture. Therefore, the system will independent to a user. Further, we proposed the used of DTW to compute the distance between the features because the speed of the user when giving sign language is various. We hoped the DTW could improve the accuracy of the recognition.

Based on our experiment, either when the user is the same or when the user is different to the user who does the training, the result of our proposed method is promising. Even though when the user who tests the data is new (her data did not exist in training data), in scenario B, the accuracy decreased, but the decrease was not much, only from 94% to 91%. This result shows that the proposed method is invariant to a user (user independent).

However, when two sign languages have similar gestures, their misclassification rates are increasing. Also, the sign languages used in the experiment are limited to 20 sign languages. Therefore, further study about feature extraction which can distinguish similar gestures is required. And it would be interesting to assess the proposed method in more sign languages.

ACKNOWLEDGE

This research was supported by a research grant for “Penelitian Pemula” in 2018 from Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.

REFERENCES

[1] [N. B. Ibrahim, M. M. Selim, and H. H. Zayed, “An Automatic Arabic Sign Language Recognition System (ArSLRS),” *J. King Saud Univ. - Comput. Inf. Sci.*, Oct. 2017.

[2] M. I. Sadek, M. N. Mikhael, and H. A. Mansour, “A New Approach for Designing A Smart Glove for Arabic Sign Language Recognition System Based on the Statistical Analysis of the Sign Language,” in *2017 34th National Radio Science Conference (NRSC)*, 2017, pp. 380–388.

[3] W. N. Khotimah, Y. A. Susanto, and N. Suciati, “Combining Decision Tree and Back Propagation Genetic Algorithm Neural Network for Recognizing Word Gestures in Indonesian Sign Language using Kinect,” *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 2, pp. 292–298, Jan. 2017.

[4] C. H. Chuan, E. Regina, and C. Guardino, “American Sign Language Recognition Using Leap Motion Sensor,” in *2014 13th International Conference on Machine Learning and Applications*, 2014, pp. 541–544.

[5] W. N. Khotimah, R. A. Saputra, N. Suciati, and R. R. Hariadi, “Alphabet Sign Language Recognition Using Leap Motion Technology and Rule Based Backpropagation-Genetic Algorithm Neural Network (RBBPGANN),” *JUTI J. Ilm. Teknol. Inf.*, vol. 15, no. 1, pp. 95–103, Jan. 2017.

[6] L. E. Potter, J. Araullo, and L. Carter, “The Leap Motion Controller: A View on Sign Language,” in *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, New York, NY, USA, 2013, pp. 175–178.

[7] T. C. Hou, W. N. W. Zakaria, T. S. Jing, R. Tomari, T. K. Sek, and A. A. M. Suberi, “Vision Based Human Decoy System for Spot Cooling,” *TELKOMNIKA Telecommun. Comput. Electron. Control*, vol. 15, no. 4, pp. 1512–1519, Dec. 2017.

[8] C. Chansri and J. Srinonchat, “Hand Gesture Recognition for Thai Sign Language in Complex Background Using Fusion of Depth and Color Video,” *Procedia Comput. Sci.*, vol. 86, pp. 257–260, Jan. 2016.

[9] R. Elakkiya, K. Selvamani, and A. Kannan, “An Intelligent Framework for Recognizing Sign Language from Continuous Video Sequence Using Boosted Subunits,” in *IET Conference Proceedings; Stevenage*, Stevenage, United Kingdom, Stevenage, 2013.

[10] P. Doliotis, “Viewpoint Invariant Gesture Recognition and 3D Hand Pose Estimation Using RGB-D,” Ph.D., The University of Texas at Arlington, United States -- Texas, 2013.

[11] I. M. Al-saihati, "Real Time Arabic Sign Language Recognition". Dahran, Saudi Arabia: King Fahd University of Petrolium and Minerals, 2006.

[12] P. Yin, "Segmental Discriminative Analysis for American Sign Language Recognition and verification". Georgia Institute of Technology, 2010.

[13] T. Aujeszky and M. Eid, “A Gesture Recogintion Architecture for Arabic Sign Language Communication System,” *Multimed. Tools Appl.*, vol. 75, no. 14, pp. 8493–8511, 2016.

[14] T. Kamnardsiri, L. Hongsit, and N. Wongta, “Designing A Sign Language Intelligent Game-Based Learning Framework with Kinect,” in *ICEL2016-Proceedings of the 11th International Conference on e-Learning: ICEl2016*, 2016, p. 64.

[15] G. C. Lee, F.-H. Yeh, and Y.-H. Hsiao, “Kinect-Based Taiwanese Sign-Language Recognition System,” *Multimed. Tools Appl.*, vol. 75, no. 1, pp. 261–279, 2016.

[16] A. Agarwal and M. K. Thakur, “Sign Language Recognition Using Microsoft Kinect,” in *Contemporary Computing (IC3), 2013 Sixth International Conference on*, 2013, pp. 181–185.

[17] C. Sun, T. Zhang, B.-K. Bao, and C. Xu, “Latent Support Vector Machine for Sign Language Recognition with Kinect,” in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, 2013, pp. 4190–4194.

[18] M. M. Saad, N. Jamil, and R. Hamzah, “Evaluation of Support Vector Machine and Decision Tree for Emotion Recognition of Malay Folklores,” *Bull. Electr. Eng. Inform.*, vol. 7, no. 3, pp. 479–486, 2018.

[19] S. I. M. Quadri, "An Integrated System for Arabic Sign Language Recognition". U.M.I., 2007.

[20] U. Von Agris, J. Zieren, U. Canzler, B. Bauer, and K.-F. Kraiss, “Recent Developments In Visual Sign Language Recognition,” *Univers. Access Inf. Soc.*, vol. 6, no. 4, pp. 323–362, 2008.

[21] I. N. Yulita, “Feature Extraction Analysis for Hidden Markov Models in Sundanese Speech Recognition,” *TELKOMNIKA Telecommun. Comput. Electron. Control*, vol. 16, no. 5, 2018.

[22] S. G. Moreira Almeida, F. G. Guimarães, and J. Arturo Ramírez, “Feature Extraction in Brazilian Sign Language Recognition Based on Phonological Structure and Using RGB-D Sensors,” *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7259–7271, Nov. 2014.

[23] M. Müller, “Dynamic Time Warping,” *Inf. Retr. Music Motion*, pp. 69–84, 2007.

[24] Sutarman, M. A. Majid, and J. M. Zain, “A Review on the Development of Indonesian Sign Language Recognition System,” *J. Comput. Sci.*, vol. 9, no. 11, pp. 1496–1505, Nov. 2013.

[25] W. N. Khotimah, N. Suciati, Y. E. Nugyasa, and R. Wijaya, “Dynamic Indonesian Sign Language Recognition by Using Weighted K-Nearest Neighbor,” in *Information & Communication Technology and System (ICTS), 2017 11th International Conference on*, 2017, pp. 269–274.