PUP Journal of Science and Technology (PUPJST)
Vol 16 (2023), (pp. 15 to 29)
https://doi.org/10.70922/7txkfp49

Original article

Application of Earth Mover's Distance Algorithm for Gesture Recognition of American Sign Language Hand Gesture

Lorgio Paul N. Gomez, Nehemias P. Locading, Kate N. Manuel Samson A. Mendoza, Zeus V. Misa, and Roselito E. Tolentino

Polytechnic University of the Philippines, Sta. Rosa, Laguna 4026, Philippines

Corresponding author: kenmetara@yahoo.com

Received: 07 March 2024; Accepted: 13 March 2024; Available online: 03 March 2026

Abstract: This study utilizes computer vision in the interpretation of static and dynamic human gestures for the American Sign Language. This is another way of communicating by people who understands and do not understand American Sign Language. They propose the application of Earth Mover's Distance, which is define as distance between two feature descriptors by the minimal amount of work needed to transform one into the other. It is use for the recognition of static and dynamic gesture for American Sign Language of users with different hand shape and orientation.

Keywords: *American Sign Language, Dynamic gesture, Earth Mover's Distance*

1. INTRODUCTION

American Sign Language (ASL) is one of the most popular sign languages which, instead of acoustically conveyed sound patterns, uses visually transmitted sign patterns to convey meaning. It involves a simultaneous combination of the hand shapes, movement and orientation of the hand, arms or body, and also the facial expressions to express the speaker's thoughts. ASL has two distinct kinds of signs that are described as follows: Static signs (gestures) are signs without movement and are determined by a certain configuration of the hand. Examples of these are the letters in the ASL alphabet, with exception of J and Z. Dynamic sign is a moving gesture determined by a sequence of hand movements and configurations. Sometimes, they are also accompanied by body and facial expressions that can convey as much meaning as the hand posture, these signs are shown in Figures 1 and 2.

Early studies in this field involve sensors attached to certain parts of the user's body. Abualola (2016) proposed glove-based system for hand-gesture recognition. The system tracks fine-grain hand movements using inertial and attitude measurements. Gestures are recognized in real-time by feeding the sensor readings to a machine- learning algorithm. In addition, gestures are communicated wirelessly to external devices for display and control purposes. The machine learning algorithm is based on Linear Discriminant Analysis (LDA), which allows for accurate and low complexity classification by projecting into a space with improved clustering and reduced dimensionality. The feature vector comprises the angles between each finger relative to the hand palm. A real-time algorithm is developed to ensure features are captured when the gestures are at a steady state as opposed to gesture transitions. Abhishek et al. proposed gesture recognition glove based on charge-transfer touch sensors for the translation of the American Sign Language. The device is portable and can be implemented with low-cost hardware. The prototype recognizes gestures for the numbers 0 to 9 and the 26 English alphabets, A to Z. Based on 1080 trials, the glove experimentally achieved an overall detection accuracy of over 92%, which is comparable with the current high-end computer parts.

American Sign Language (ASL) alphabet recognition, using marker-less vision sensors, is a challenging task due to the complexity of ASL alphabet signs, self-occlusion of the hand, and limited resolution of the sensors. Many research and studies have been done and developed regarding Sign Language recognition. Funasaki (2017) proposed a sign language recognition using leap motion controller to give an alternative method of voice input for deaf people. The recognition target is finger spelling of 24 letters, exclude two letters that require finger movement. They use 16 kinds of conditions that focus on the characteristics of hand and finger. By changing the order of the conditional branches, a different decision tree is generated with a different average recognition rate for the finger spelling of 24 letters. But the problem of their system is the hand orientation, user's hand must be properly oriented on the leap sensor in order to recognize the employed sign language based on the condition that they have been set on each letter. Sign language recognition using depth sensor has become more widespread. However, it is difficult to detect meaningful signs from a contiguous hand-motion stream because the signs vary in both motion and shape in 3D space. Yang (2015) proposed a sign language recognition using Kinect. In this research, they use three-dimensional depth information from hand motions, generated from Microsoft's Kinect sensor and apply a hierarchical conditional random field (CRF) that recognizes hand signs from the hand motions. The proposed method uses a hierarchical CRF to detect candidate segments of signs using hand motions, and then a BoostMap embedding method to verify the hand shapes of the segmented signs. The BoostMap embedding method decreases the insertion and substitution errors by verifying the hand shape; however, it reduces the correct detection rate because of its own classification errors.

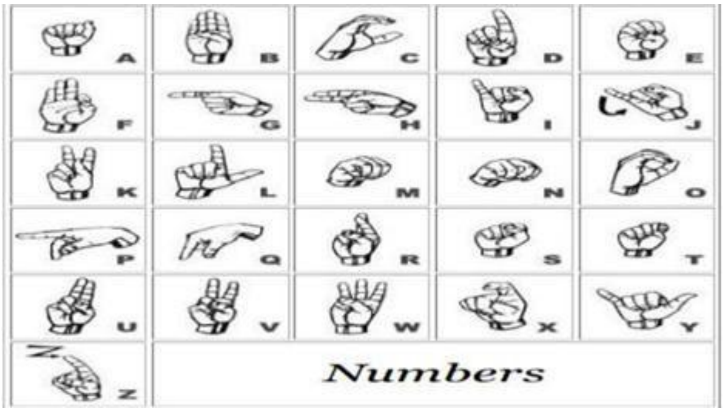

Figure 1. ASL Static signs

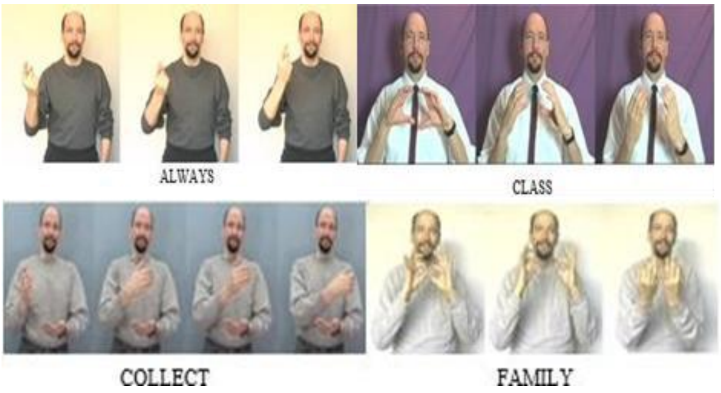

Figure 2. ASL Dynamic signs

The problem of this research is the different hand shape and orientation in recognizing static and dynamic gestures for American Sign Language. To solve this problem, the use of Earth Mover's Distance was proposed to be able to recognize different hand shapes and orientation of various users. This study was essential and beneficial in helping deaf and mute people to communicate in any person in social media through hand gestures. The focus of the study was to enable different users, trained or untrained, to be able to use a system without significantly affecting the recognition rate.

The study was created in order to improve the ASL letter and words recognition. The output of this research can be used as a concrete tool for communication between deaf and non-deaf people. To future researchers, this proposed study can serve as a guide educator a way to teach and train ASL to students who want to learn ASL. Furthermore, this may also be a means of creating the other ways of improving the accuracy of ASL recognition. Thus, it can also be used as a reference for other related study.

2. METHODOLOGY

A. Obtaining the Earth Mover's Distance to be used in the recognition of static and dynamic gestures for American Sign Language

2.1 *Image Acquisition and Conversion to Grayscale Image*

The first step of the whole process of the system is started with the image acquisition for generating the input frames, done by a camera. The researchers used an A1Tech 720p HD Camera. A grayscale image is one in which the only colors are shades of gray. The reason for this is that it removes all color information, thus less information needs to be provided for each pixel. In fact, gray color is one in which the red, green and blue components all have equal intensity in RGB space. In this process, these three values would be combined into a single value after removing color from an image. After this process, the image frame would be in grayscale and have only one pixel value.

2.2 *Image Scaling*

Image scaling is the resizing of a digital image. When scaling a graphic image, the pixels that make up the image can be scaled using geometric transformations, with no tangible loss of information. This puts every frame acquired on a comparable scale.

2.3 *Binary Image and Image Complement*

Binary images are images whose pixels have only two possible intensity values that are normally displayed as black and white; numerically, the two values are often 1 for black, and 0 for white. In this process, the grayscale image in the last step is converted to a binary image to make the source image free from noise and to separate an object in the image from the background.

The image complement computes the complement of a binary or intensity image. In the complement of a binary image, zeros become ones and ones become zeros; black and white are reversed. The hand will be separated from the background.

2.4 *Hand area Extraction*

Feature extraction is a type of dimensionality reduction that efficiently represents interesting parts of an image as a compact feature vector. This approach is useful when image sizes are large, and a reduced feature representation is required to quickly complete tasks such as image matching and retrieval.

After the image has been converted into a binary image and has successfully separated the hand from the background, there would only be two values of 1 for white and 0 for black, the 1's represent the region where the white pixels are present, which is the area of the hand. These represent the bins of the feature and are stored to be used for the computation of EMD. The same is done for both the input and the template.

2.5 *Recognition using Earth Mover's Distance*

After all the image pre-processing, the Earth Mover's Distance is computed that will lead to the recognition of the sign. The smallest measure of dissimilarity will mean that it had the least cost transportation — or the movement that required the least amount of work. Using the value of the feature extracted from the previous process, and the value of the feature of all templates in the database, they will all be computed in input-template pairs. To get the EMD, the researchers first solved optimal flow by using formula below which is formalized as the following linear programming problem.

$$\text{WORK(P, Q, F)}= \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij} \qquad (1)$$

P – first signature; Q – second signature; D = [dij] the ground distance matrix between clusters P and Q, where Euclidean Distance was used. F=[fij] flow between pi and qj. Once the transportation problem is solved, and we have found the optimal flow F, the earth mover's distance was defined as the resulting work by the total flow:

$$\text{EMD(P,Q)} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}} \qquad (2)$$

When used to compare distributions that have the same overall mass, the EMD is a true metric. To allow for partial matching or of application to signatures of different sizes, there must be a normalization factor done by

Formula 2. The smallest computed EMD value means it has the smallest minimum dissimilarity distance between the input hand gesture and each template. Each of the input-template pair EMD value will all be compared to one another, then each subtracted 1, and the largest value will represent having the highest confidence value or the best match, and the system will choose it and that will lead to the recognition of the sign. The same process is done for both the static and dynamic gestures. The only difference is that for dynamic gestures, the first and last frame of the dynamic gesture must be recognized in sequence shown in Figure 3.

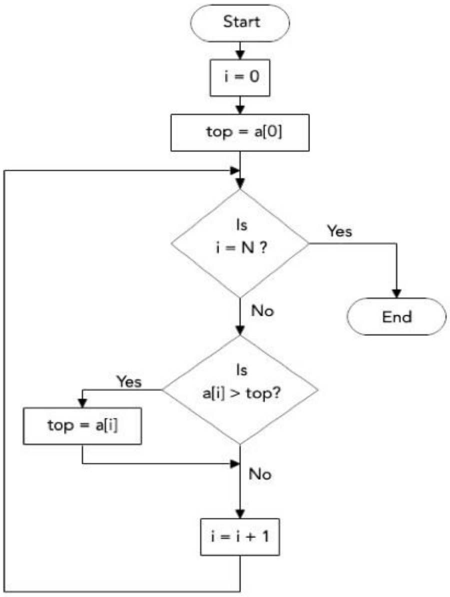

Figure 3. Process flow of choosing top recognition

B. *Obtaining the Reliability of the System in Recognizing Static and Dynamic Gestures for American Sign Language*

In order to obtain the reliability of the system in recognizing static and dynamic gestures for American Sign Language, the researchers will conduct sets of 15 trials for every letter and signed words of ASL chosen by the proponents. Each static and dynamic gesture is performed 15 times by three (3) different users, including the person who trained the system. The number of trials and number of users was based on the previous study for ideal comparison.

The total number of successful recognitions and the total number of trials performed will both be recorded. The formula to obtain the percentage of the reliability of recognition is presented in formula 3.

$$Recognition\,(\%) = \frac{total\,number\,of\,successful\,recognitions}{total\,number\,of\,trials} x100 \qquad (3)$$

3. RESULTS AND DISCUSSION

Acquired Reliability of the System in Recognizing Static and Dynamic Gestures for American Sign Language

The total number of successful recognitions and the total number of trials performed were both recorded. The results of the trials are shown in Table 1 and Table 2 below for gestures of ASL Fingerspelling and for ASL Words respectively. These tables indicate the average system recognition reliability for every gesture obtained from the test recognition performed by all three (3) users.

Table 1. Average Recognition Rate for ASL Fingerspelling

ASL Finger Spelling	**Gathered Data**			**Reference Data**		
	Number of Successful Recognition	Number of Unsuccessful Recognition	Average Recognition Rate	Number of Successful Recognition	Number of Unsuccessful Recognition	Average Recognition Rate
A	45	0	100.00%	45	0	100.00%
B	42	3	93.33%	41	4	91.11%
C	41	4	91.11%	45	0	100.00%
D	43	2	95.56%	42	3	93.33%
E	42	3	93.33%	41	4	91.11%
F	45	0	100.00%	35	10	77.78%
G	44	1	97.78%	42	3	93.33%
H	45	0	100.00%	42	3	93.33%
I	37	8	82.22%	38	7	84.44%

ASL Finger Spelling	Gathered Data			Reference Data		
	Number of Successful Recognition	Number of Unsuccessful Recognition	Average Recognition Rate	Number of Successful Recognition	Number of Unsuccessful Recognition	Average Recognition Rate
J	38	7	84.44%	29	16	64.44%
K	44	1	97.78%	37	8	82.22%
L	45	0	100.00%	45	0	100.00%
M	40	5	88.89%	31	14	68.69%
N	41	4	91.11%	30	15	66.67%
O	40	5	88.89%	37	8	82.22%
P	40	1	97.78%	39	6	86.67%
Q	44	3	93.33%	45	0	100.00%
R	42	6	86.67%	43	2	95.56%
S	39	2	95.56%	42	3	93.33%
T	43	7	84.44%	39	6	86.67%
U	44	1	97.78%	38	7	84.44%
V	45	0	100.00%	41	4	91.11%
W	45	0	100.00%	41	4	91.11%
X	45	0	100.00%	38	7	84.44%
Y	45	0	100.00%	41	4	91.11%
Z	42	3	93.33%	32	13	71.11%
Average	94.36%			88.55%		

Table 2. Average Recognition Rate for ASL Words

ASL Words	Gathered Data			Reference Data		
	Number of Successful Recognition	Number of Unsuccessful Recognition	Average Recognition Rate	Number of Successful Recognition	Number of Unsuccessful Recognition	Average Recognition Rate
Always	38	7	88.44%	42	3	93.33%
Class	45	0	100.00%	40	5	88.89%
Collect	43	2	95.56%	41	4	91.11%
Corner	45	0	100.00%	45	0	100.00%
Doctor	33	12	73.33%	38	7	84.44%
Family	45	0	100.00%	34	11	75.56%
Go	41	4	91.11%	45	0	100.00%
Group	44	1	97.78%	42	3	93.33%
Help	40	5	88.89%	45	0	100.00%
House	43	2	95.56%	39	6	86.67%
Impossible	45	0	100.00%	31	14	68.69%
Last	34	11	75.56%	34	11	75.56%
Make	42	3	93.33%	40	5	88.89%
Meet	42	3	93.33%	45	0	100.00%
Moment	43	2	95.56%	32	13	71.11%
Name	41	4	91.11%	42	3	93.33%
Need	45	0	100.00%	40	5	88.89%
Proud	39	6	86.67%	36	9	80.00%
Seems	39	6	86.67%	41	4	91.11%
There	36	9	80.00%	45	0	100.00%

ASL Words	Gathered Data			Reference Data		
	Number of Successful Recognition	Number of Unsuccessful Recognition	Average Recognition Rate	Number of Successful Recognition	Number of Unsuccessful Recognition	Average Recognition Rate
Thing	45	0	100.00%	40	5	88.89%
Turn	39	6	86.67%	40	5	88.89%
Until	45	0	100.00%	34	11	75.56%
With	45	0	100.00%	44	1	97.78%
You	33	12	73.33%	38	7	84.44%
Your	43	2	95.56%	43	2	95.56%
Average	91.71%			88.55%		

For letter I that has a recognition rate of 82.22%, it is often mistakenly recognized as letter D due to similarity in shape as shown below in Figure 4. The reason for this is because of the confusion in the system due to the only difference between these two signs is the position of a finger.

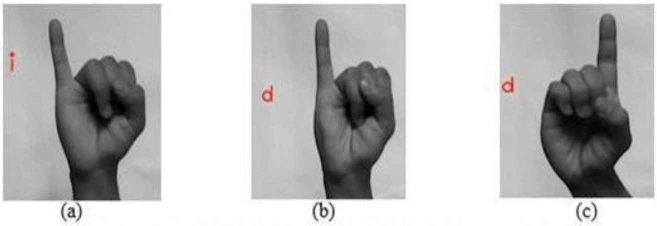

Figure 4: (a) letter I; (b) letter I misidentified as D and (c) letter D

For letter R that has a recognition rate of 86.67%, it is mistakenly recognized as letter D due to similarity shown in Figure 5. The reason for this is because of the overlapping fingers of letter R, which might be taken by the system as one finger. It it also very similar in the respect that the two letters have the same position of fingers.

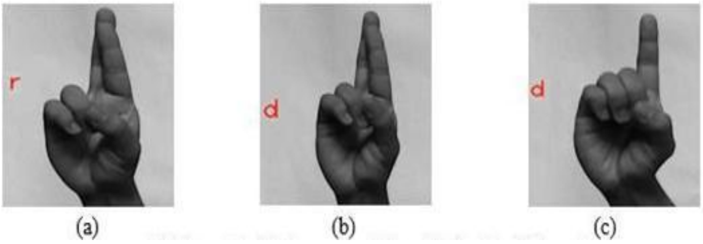

Figure 5: (a) letter R; (b) letter R misidentified as D and (c) letter D

For letter T that has a recognition rate of 84.44%, it is often confusing with the letter N. The system is making an error in recognition because the overlapping fingers are taken as one, similar to the case with letter R, as shown in Figure 6.

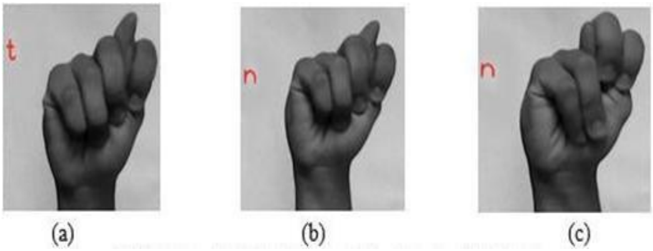

Figure 6: (a) letter T; (b) letter T misidentified as N and (c) letter N

For the gesture ‘Doctor’ that has a recognition rate of 73.33%, it is often mistakenly recognized as ‘Impossible’ due to similarity in signs as shown below in Figure 7.

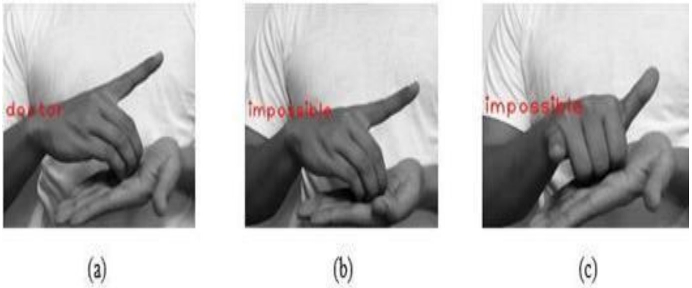

Figure 7: (a) sign for “Doctor; (b) “Doctor” misidentified as “Impossible” and (c) sign for “Impossible”

For the gesture ‘Last’ that has a recognition rate of 75.56%, it is often mistakenly recognized as ‘Name’ due to similarity in sign as shown below in Figure 8.

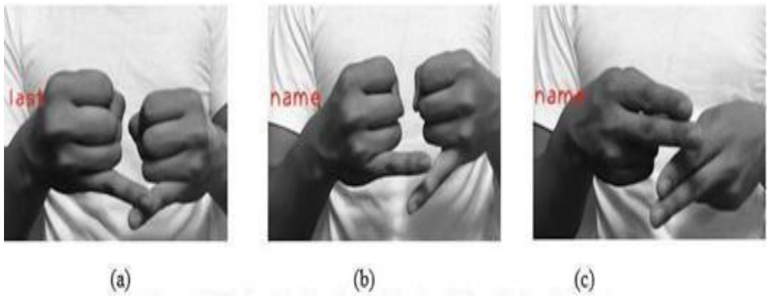

Figure 8: (a) sign for “Last; (b) “Last” misidentified as “Name” and (c) sign for “Name”

For the gesture ‘You’ with a recognition of 73.33% and There with 80.00%, they are often mistakenly recognized as one another due to similarity in sign, the only difference is the direction it points towards, and the different hand used when signing as shown below in Figure 9.

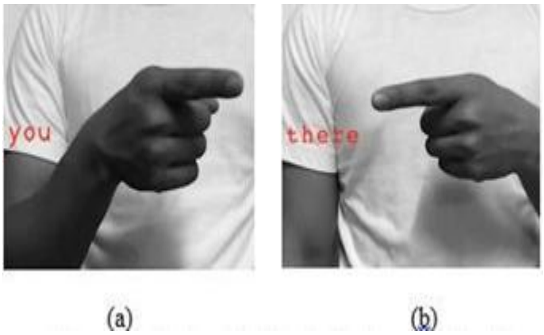

Figure 9: (a) sign for “Sign; (b) sign for “There”

There are also times when dynamic gestures were not able to recognize by the system due to the following. The system recognizes the first frame but not the last frame. Or the system did not recognize the first frame but recognize the last frame. As a result, the system cannot present a recognized dynamic gesture as shown below on Figure 10.

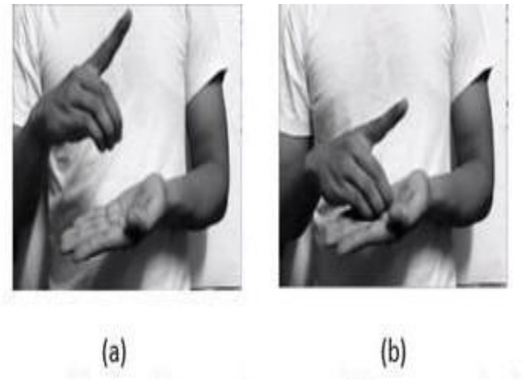

Figure 10: (a) First frame of “Doctor” and (b) Unrecognized “Doctor”

Table 3 shows the average recognition rate for every user in the test recognition they had performed, as well as for the reference thesis of this study. It can be seen that the highest average recognition rate is 96.92% for static and 94.62% for dynamic for User 1, but the reference got higher recognition rates of 97.18% and 98.97% for User 1. This user got the highest recognition rate due to the fact that this user’s templates were the ones used in the system; the other two users, despite not training the system, still got a high recognition rate. The recognition rates for Users 2 and 3 were lower than the rates of User 1, but it can be seen that the recognition rates of the new system for Users 2 and 3 are greatly improved. The overall system reliability for the system is 93.03% based on the results, compared to the previous system’s 87.82% overall reliability which shows that it has improved overall.

Table 3. Average Recognition Rate of every user for overall reliability

SYSTEM RECOGNITION OVERALL RELIABILITY				
User	**Data Gathered**		**Reference Data**	
	Average Recognition Rate (Finger Spelling)	**Average Recognition Rate (Hand Gestures)**	**Average Recognition Rate (Finger Spelling)**	**Average Recognition Rate (Hand Gestures)**
1	96.62%	94.62%	97.18%	98.97%
2	94.87%	90.77%	82.05%	81.79%
3	91.28%	89.74%	82.05%	84.87%
Average	**94.36%**	**91.71%**	**87.09%**	**88.55%**
Overall Reliability:	**93.03%**		**87.82%**	

4. CONCLUSIONS

The researchers developed a system that can recognize static and dynamic American Sign Language hand gestures using the application of Earth Mover's Distance algorithm despite differences in hand shape and orientation. Although the reliability of the system in recognizing the gestures was high, there were factors that affect the successful recognition of the system. These were the factors identified by the proponents that affect recognition: (1) Signs that had similar form have values of area that is close, causes confusion to the system since the signs were similar. (2) Overlapping fingers often mistaken as one finger instead of two crossed with each other. (3) There are also times when gestures are not being recognized by the system, because the first and last frame of the dynamic gesture was not recognized in sequence.

Regardless of the mentioned problems, the proponents conclude that the system can still recognize American Sign Language hand gestures despite changes in hand shape and hand orientation. The reliability of the system has improved the recognition rate in comparison to the results from the reference study. The difference of rates of static, dynamic and overall reliability can be seen through Table 1, 2, and 3. The system also solved the problems of the hand shape and orientation in recognizing gestures; therefore, it prevented the adding of new set of templates if there were more users

5. RECOMMENDATIONS

Based on the results and data gathered from the study, the proponents came up with the following recommendations for further development. The proponents recommend using other features or adding more features selected from the images to be used in comparing the distributions. This may help avoid the confusion between similar gestures that have the same positioning of fingers and hand. If additional features were added or enough selected features were made, it can help better identify the differences between the almost identical gestures. The proponents recommend using a feature where the system will be able to identify where the fingertips are and use them as classifiers. This may be very useful for signs and gestures where the fingers are overlapping or crossing one another because the system often confuses them as being one finger instead.

6. REFERENCES

Vaishali, S., & Kulkarni, P. (2010). Appearance based recognition of American Sign Language using gesture segmentation. *International Journal on Computer Science and Engineering*, 2(3), 560-565.

Yang, H.-D. (2015). Sign language recognition with the Kinect sensor based on conditional random fields. *Multidisciplinary Digital Publishing Institute*, 10(6), 123-135.

Abualola, H., Ghothani, H. A., & Eddin, A. N. (2016). Flexible gesture recognition using wearable inertial sensors. In *Circuits and Systems* (pp. 123-130). Abu Dhabi, United Arab Emirates.

Abhishek, K. S., Qubeley, L. C. F., & Ho, D. (2016). Glove-based hand gesture recognition sign language translator using capacitive touch sensor. In *Electron Devices and Solid-State Circuits* (pp. 200-205). Hong Kong, China.

Dong, C., Leu, M. C., & Yin, Z. (2015). American Sign Language alphabet recognition using Microsoft Kinect. In *Computer Vision and Pattern Recognition Workshops* (pp. 124-130). Boston, MA, USA.

Funasaka, M., Ishikawa, Y., Takata, M., & Joe, K. (2017). Sign language recognition using Leap Motion Controller. In *Parallel and Distributed Processing Techniques and Applications* (pp. 98-104). Nevada, USA.

Taskiran, M., Killioglu, M., & Kahraman N. (2018). A real-time system for recognition of American sign language by using deep learning. Proceedings of the 41st international conference on telecommunications and signal processing (pp. 1-5). Athens, Greece,

Bantupalli K. and Xie Y., (2018). "American Sign Language Recognition using Deep Learning and Computer Vision," *2018 IEEE International Conference on Big Data* (pp. 4896-4899). Seattle, WA, USA,

Bin, L.Y., Huann, G.Y., Yun, L.K.(2019) Study of Convolutional Neural Network in Recognizing Static American Sign Language, Proceedings of the 2019 IEEE International Conference on Signal and Image Processing Applications, ICSIPA 2019

Kadhim R.A., Khamees M., (2020). A real-time american sign language recognition system using convolutional neural network for real datasets, TEM Journal. Volume 9, Issue 3, Pages 937-943

Saleh Y., Issa G.F., (2020). Arabic sign language recognition through deep neural networks fine-tuning, International Journal of Online and Biomedical Engineering, Vol.16 Issue 5, pp. 71-83

Lee C.K.M., Ng K.H., Chen C.H., Lau H.C.W., Chung S.Y., Tsoi T., (2021). American sign language recognition and training method with recurrent neural network, Expert Systems with Applications, Vol.167, Article 114403