Designing a Sign Language Translation System using Kinect Motion Sensor Device

Shihab Shahriar Hazari*, Asaduzzaman, Lamia Alam and Nasim Al Goni
Dept. of Computer Science & Engineering
Chittagong University of Engineering & Technology,
Chittagong, Bangladesh.
Email: shihab2555@gmail.com, asad@cuet.ac.bd, lamiacse09@gmail.com and shourav1208@gmail.com

Abstract— Hand gesture is one of the most natural and expressive ways for the hearing impaired. However, because of the complexity of dynamic gestures either static gestures, postures, or a small set of dynamic gestures are focused by most researchers. In this paper, Kinect Motion Sensor Device is used to recognize the gesture of the user. But the gesture of each user of a particular word will be slightly different. As real-time recognition of a large set of dynamic gestures is considered, some efficient algorithms and models are needed. Here an efficient algorithm is used to recognize the gesture and translate them. In this paper, in order to recognize the gesture in both training and translation mode a grid view algorithm has been used. We evaluated our system by translating gesture by various people for twelve different words and experimental results reveals that our proposed system has about eighty percent success rate in translating gestures.

Keywords— Kinect, sign language, gesture, depth data, translation

I. INTRODUCTION

In present world communication system plays a vital role in the development of mankind. The development of mankind is heavily dependent on communication system. Though there are some shortcomings in this present communication system. One of the shortcomings is the communication gap between deaf and mute people with general people. Deaf and mute people use sign language to communicate between them, which general people can't understand. On the other hand general, people use oral language which is inconsolable to deaf and mute people. To decrease this communication gap this paper proposed a method. In this system user will give a gesture as input, the gesture will be translated and the translated output will be shown to another user. The Kinect motion sensor device is used to recognize gestures by which deaf and mute people deliver sign language. Also an output screen is used where the translated word will be shown.

We intended to develop a system that can translate the sign language of a deaf user so that an ordinary person can understand. A deaf user will express his gesture in front of Kinect device. Then the system will detect the respective sign language for the gesture and will give the output in monitor as text. Thus the ordinary user will understand the sign language. The specific aims of this work are given below:

- Use data provided by the Microsoft Kinect device.
- Recognize a list of basic signs. This list will contain key words. Using these words, the deaf user will be able to transmit what he/she needs and the communication between deaf and ordinary user will be possible.
- By executing a combination of several signs in a row, the user will be able to construct some basic sentences that will make the communication more suitable.
- Design an interactive user interface so that the user will be able to run the application without any previous knowledge.
- The system must work on real time and give an instantaneous output once the sign is executed.

II. RELATED WORK

A lot of research activities have been conducted on designing sign language translator.

Different project has being done using different types of device. A work had been done from University of Tennessee which used Kinect motion sensor device and Hidden Markov Model (HMM) to translate from American Sign Language (ASL) [1]. Discriminative word model is being used in another work to translate sign language from ASL [2]. Another work is being done to translate sign language using HMM [3]. In this case robotic gloves have being used.

In [4-7] different neural networks have been used for ASL translation. But these approaches require considerably more time and data to train.

In [8] Support Vector Machines and k-Nearest Neighbors were employed to characterize each color channel after background subtraction and noise removal and attain an accuracy of 62.3% . Sharma et al. proposed a work using a contour trace, which is an efficient representation of hand contours [9].

Most of the previous work used expensive camera, robotic arms or gloves. In contrast to that, our proposed system uses a comparatively cheaper sensor (i.e., Kinect) that will work in non-laboratory conditions too. Here as depth data is considered as input thus the project is independent of environment constraints such as lighting condition. Here a framework is

978-1-5090-5627-9/17/$31.00 ©2017 IEEE

being proposed which can capture the gesture as input and translate it with the help of database and developed algorithm.

III. Proposed Methodology

Sign language is expressed by gesture. Each gesture contains different word. By combining those words or gesture a full, meaningful sentence is found. In this project the machine translates each gesture, word by word. Here user can also train the machine to learn gesture for new word which is not available in the database. Then that word will be included in the database and user can use it later. Here, two features is added: translation and training.

User first has to determine which feature he/she will use. After selecting user have to stand in a pre-defined specific position in front of Kinect Motion Sensor Device. Then, the initial position of human will be detected based on the depth image. After that frame by frame position is updated according to the depth data which results in tracking the human. After detecting the position correctly system will instruct the user to provide the gesture. Than the system will capture the gesture according to the algorithm and will show complete message after capturing the gesture. Here, fig. 1 shows the proposed framework for the training mode while fig. 2 describes the translation mode framework.

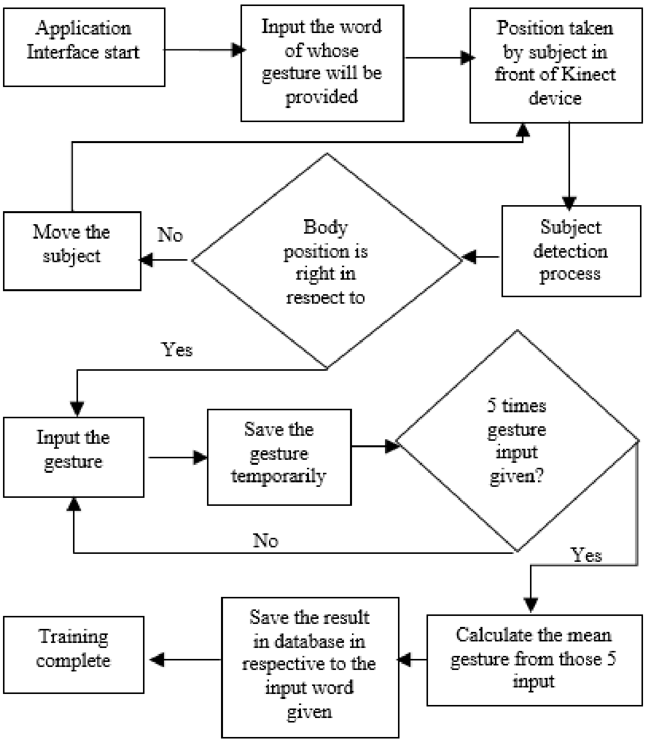

Application Interface start
Input the word of whose gesture will be provided
Position taken by subject in front of Kinect device
Subject detection process
Body position is right in respect to
No
Move the subject
Yes
Input the gesture
Save the gesture temporarily
5 times gesture input given?
No
Yes
Calculate the mean gesture from those 5 input
Save the result in database in respective to the input word given
Training complete

Fig. 1. Proposed framework for training method

A. Subject Detection Process

Human detection is one of the internal features of Kinect Motion Sensor Device. The machine does it by capturing the depth information of human body. The infrared (IR) emitter and infrared (IR) depth sensors of Kinect are the key point to capture the depth information of human. A depth image of human is produced by Kinect in front of it. Both IR emitter and the IR depth sensor use a combined function to obtain the X, Y and Z co-ordinate values on specific point of the detected human. Fig. 3 shows the methodology of determining a human in front of the machine. Here 10 joint of subject is considered. Those are Head and Neck and Shoulder, Elbow, Wrist and Palm of both hands.

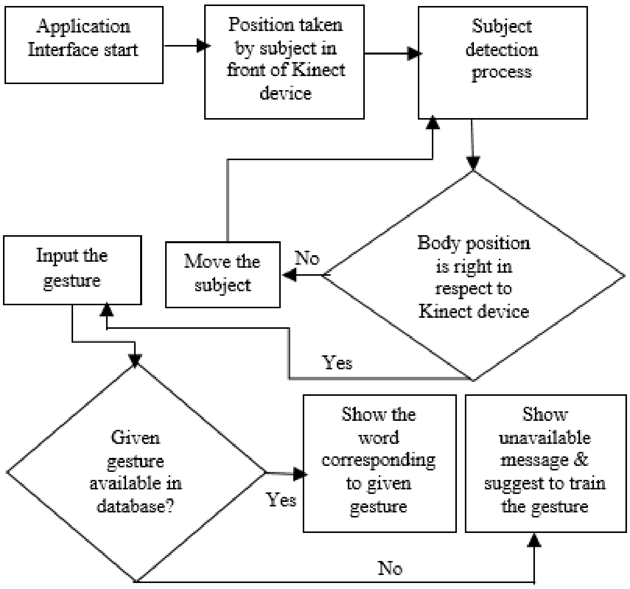

Application Interface start
Position taken by subject in front of Kinect device
Subject detection process
Body position is right in respect to Kinect device
No
Move the subject
Input the gesture
Yes
Given gesture available in database?
Yes
Show the word corresponding to given gesture
Show unavailable message & suggest to train the gesture
No

Fig. 2. Proposed frameworks for translation method

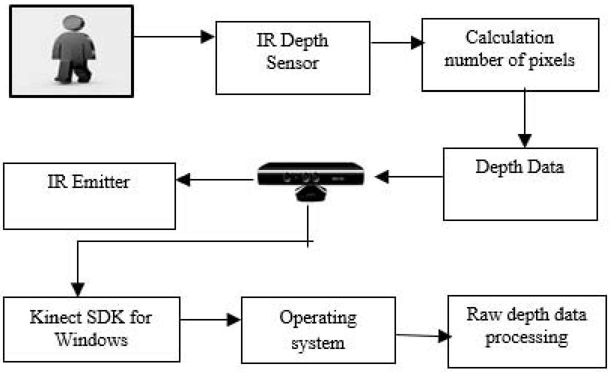


Fig. 3. Methodology of human detection by Kinect Motion Sensor Device

B. Depth Data Processing

A model based method is used for processing the depth data. Here, human is detected using depth information obtained by Kinect using a 2-stage head detection process. It includes a 2D edge detector along with a 3D shape detector. Fig. 4 shows the overview of processing the depth data which results the human detection.

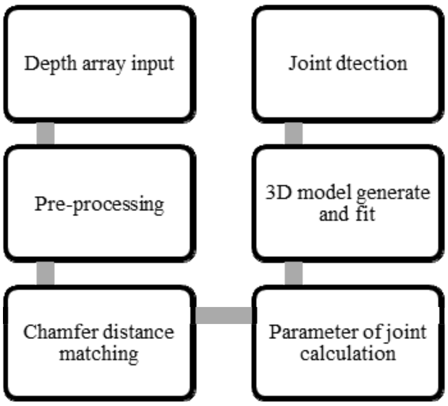

Depth array input
Joint dtection
Pre-processing
3D model generate and fit
Chamfer distance matching
Parameter of joint calculation

Fig. 4. Proposed framework for translation method

2D chamfer for distance matching algorithm is used. Whole image is scanned here and possible reason of containing people given. Each of these regions using a 3D head model is examined here. In this project total 10 joints are detected. Each of the detection contains the same process. The parameters of a joint from the depth array is executed and the parameter is used to build a 3D model. Then the 3D model is matched against all the detected regions and final estimation is made. From the detection result of 2D chamfer matching, the depth of the joint can be found. If head is considered as a sample detection point of human body, in order to generate virtual 3D model of sphere the only parameter that is required is radius of the sphere has conducted a regression test that have arrived at a cubic equation that gives the relationship between the depth value and the approximate height of it. The cubic equation that fits the geometrical height of the head is computed as Eq. 1.

$$-y = -1.3835\times 10^{-9}x^{3} + 1.8435\times 10^{-5}x^{2} - 0.09140x + 189.38 \quad (1)$$

By equation Eq. (1), the standard height of the head is calculated in this depth. Then search for the head within a certain range that is defined. Standard height of the head is defined by the Eq. (2).

$$R = 1.33\,\frac{h}{2} \quad (2)$$

Here, h is the height of the head, R is the search radius. Next the height from millimeter is converted into pixel unit as the matching works pixel by pixel.

$$Radius\ (pixel) = \left(\frac{1}{1.3}\right) \times Radius\ (mm) \quad (3)$$

At the same way the other joint rather than head is detected.

C. Human position and gesture detection

Next part is gesture detection and the user have to stand straightaway to the Kinect sensor. The correct position of human follows two conditions. 1st one is user have to stand in between 2-3 meters from the sensor. This distance can be measured by Z co-ordinate calculated by the machine. 2nd condition is the neck joint of the user should always be inside of a yellow box on the screen. The position of yellow box is pre-defined and fixed in the screen. The program can't translate or train if the neck joint remain outside the yellow square box. In fig. 5 a sample image is given to show the correct position of a human body in respect to Kinect Motion Sensor Device.

Once the user stood at the correct position the yellow box will disappear and a 15*15 virtual square graph will appear behind the body in the screen. One of the most important thing is that that the graph size varies upon the height of the user shown in fig. 6 For the user have longer height the graph size will be larger comparing to the person have smaller height. A person's height can be approximately measured by the distance between the end of middle finger of his/her two hand when he/she stands by stretching out his/her both hand on two side. In this project this concept is used to determine the height of human using the following equation.

$$L = (s1 \sim s2) + (s1 \sim p1) * 2 \quad (4)$$

Here, L=Length of human.

s1~s2= Distance between the two shoulder

s1~p1=Distance between shoulder and palm of any hand

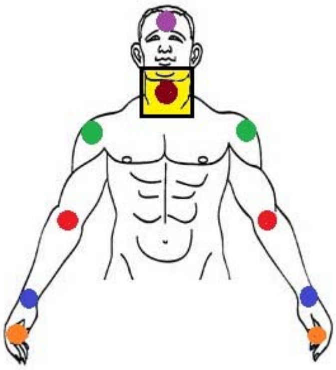


Fig. 5. Neck joint should always be inside in a pre-defined yellow box

After measuring the height the horizontal/vertical length of the whole graph and length of the each square in the graph is calculated. This calculation is only valid for this specific user. For other user the graph length will vary following the same process. Equation 5 and 6 defines the horizontal/vertical length of total graph and square length respectively.

$$G1 = L/15 * 16 \quad (5)$$

$$G2 = G1/15 \quad (6)$$

Here, L= Length of human

G1= Horizontal/vertical size of human

G2= Square length

The graph will appear such a way that the neck position will remain at the middle of the graph. That is in 8th position in both horizontally and vertically. Now user can provide the gesture. This program can detect sign language word by word. Even though for every different word of sign language gestures are different but the initial and final position of the body for each gesture should always be same for this project which is shown in fig. 7. When the device detects the body at this position it will consider it the starting point of a gesture. Next time it detect the body at the same position it consider it as the ending point of that gesture. Than the program translate that gesture and show the corresponding translated word in the output screen. In case of training it will continue the process for 5 times for each gesture and then save the mean value of that gesture in respect to the input word.

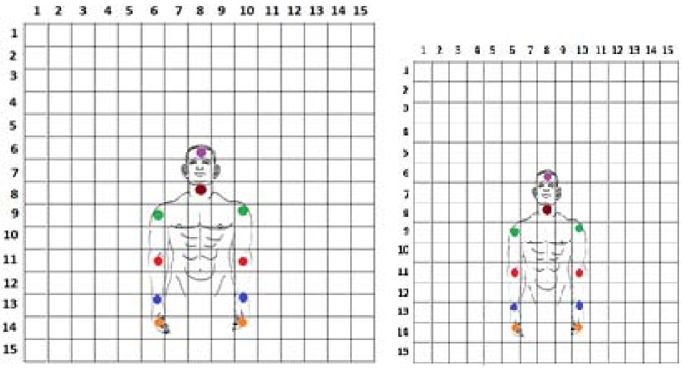


Fig. 6. Graph size is different for different person

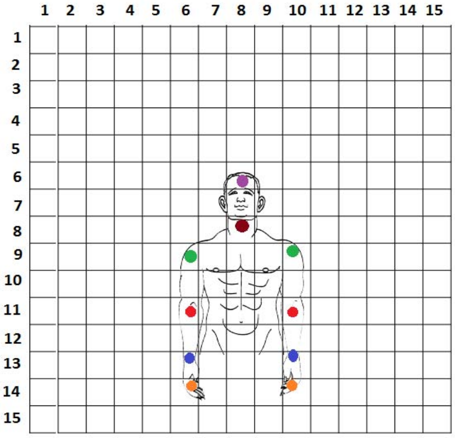


Fig. 7. Position of human body at starting/ending point of a gesture

D. Database Management

The database contains the word with its sign language or gesture. In case of translation mode the input gesture given by the user is matched from the database and shows the respective word. In case of training mode the gesture input by the user is saved in the database corresponding to the given word. Here the sequence of each joint except head and neck is taken in consideration for creating the database. In training mode a user completes a gesture the sequence movement is saved in two matrixes in the database. These two matrixes are for separate hands. Each matrix consists of four row and variable number of column. One row contains the movement sequence of one specific joints. The four row contains movement sequence of shoulder, elbow, wrist and palm from top to bottom. In database for every word two matrix is stored. During storing the position of a joint both co-ordinate of x and y are combined by using following equation.

$$M = X * 100 + Y \tag{7}$$

For example if the position of joint is captured in co-ordinate (6, 14) the combine position will be considered as 614. In fig. 8 and 9 a sample movement of palm of both hand and the corresponding matrix in respect to the movement of the palm is showed. Here red highlighted mark for the left hand and orange for the right hand.

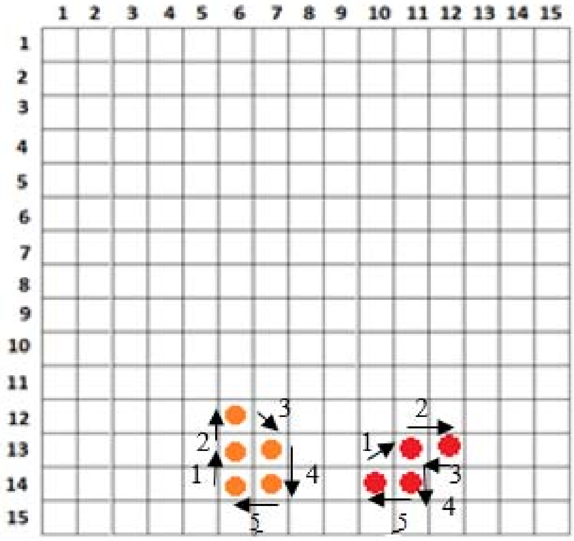


Fig. 8. Movement sequence of palm of both hand of a gesture

$$\mathbf{R} = \begin{bmatrix} 614 & 613 & 612 & 713 & 714 & 614 \end{bmatrix}$$

$$\mathbf{L} = \begin{bmatrix} 1014 & 1113 & 1213 & 1113 & 1114 & 1014 \end{bmatrix}$$

Fig. 9. Corresponding matrix of palm for the gesture shown in Fig 8

In this way the sequence for shoulder, elbow and wrist is saved in row 1, 2 and 3 respectively in the matrix for both hands. The two matrixes are saved corresponding to the input word given.

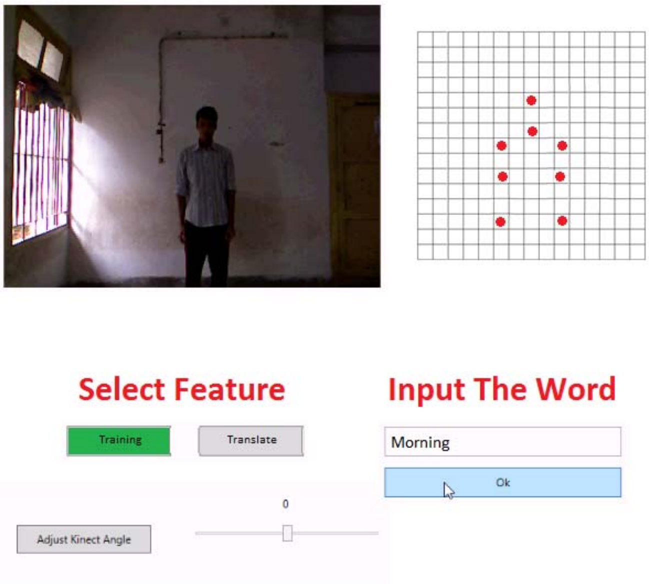
IV. Experimental Results

In order to evaluate the proposed framework, data has being taken in several conditions. Here 6 volunteer has delivered their gesture for twelve different words. Each word consists of different gesture including simple to complex gesture. Human height and shape has a vital impact on this project. Thus the volunteer being chosen has different height and shape.

A. Output of the System

We experiment and analyze the system with different human as input. We found significant results for most of the cases. There are two main features of this program, training and translation. In the software interface we showed the RGB or video of the user with gesture along with the depth data corresponding to the gesture. User need to select first about which feature he is going to use. The interface of the software changes according to the feature selected. Beside there is an automated Kinect feature named "Adjust Kinect Angle". By this feature user can set the Kinect camera angle if needed.

Fig. 10 shows that user has selected training mode. In this case user needs to input the word first whose corresponding gesture he/she is going to input.



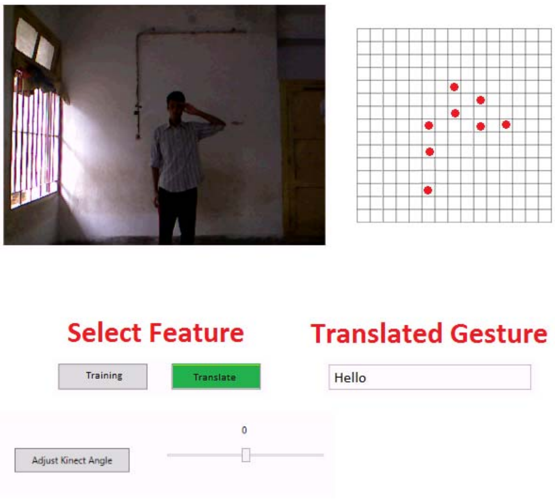
Select Feature

Input The Word

Morning

Fig. 10. Training feature selected by user

Fig. 11 shows that user has selected translate feature. After selecting the feature user has to input the gesture which need to translate. Here the program successfully translated the gesture and showed it.



Select Feature

Translated Gesture

Hello

Fig. 11. Translate feature selected by user

B. *Experimental Setup*

Before experimenting with the volunteer the gesture for those twelve experimented word is already being given and saved in database. Than the volunteer are asked to deliver their gesture for those twelve words separately. The experiment is being done in several days and in several locations to find the dependency of the project upon the environment. After completing the experiment its being found that the success rate varied for different volunteer and also for different gesture. Though each of them delivered same gesture for 1 word the machine failed to translate some gesture. Table I shows the data analysis which contain if the machine could successfully translate the gesture delivered by the volunteer or not.

TABLE I. SUCCEED/FAILED TRANSLATION FOR DIFFERENT GESTURE FOR DIFFERENT PERSON

Word	Person 1	Person 2	Person 3	Person 4	Person 5	Person 6
Hello	Yes	Yes	Yes	Yes	Yes	Yes
Bye	Yes	Yes	Yes	Yes	Yes	No
End	Yes	Yes	Yes	No	No	Yes
Bag	No	Yes	No	Yes	No	Yes
Morning	Yes	Yes	No	No	Yes	No
Evening	Yes	No	Yes	No	Yes	No
Sign	Yes	Yes	Yes	No	Yes	No
Come	No	No	Yes	No	No	No
Later	Yes	No	Yes	No	No	No
Picture	Yes	No	Yes	Yes	Yes	No
Umbrella	Yes	Yes	Yes	No	Yes	Yes
Here	Yes	Yes	No	No	No	No

C. *Data Analysis*

The gesture given by different person for the same word is slightly different. This is because the human structure of different persons is different. The system treat person of different height in different way. Though human structure doesn't only depend on height rather it also depends on shape, weight etc. Considering these facts the system became unsuccessful to translate some gesture correctly even though the person provided correct gesture. In most cases the machine became successful to translate the word correctly. Fig. 12 shows the translation success rate (%) for different person depending upon the gesture for those twelve experimental words.

Data analysis evaluation not only done on person but also it's being done for those twelve experimental words. The reason behind for some unsuccessful translation is not only the human structure but also the type gesture delivered and gesture detection. There are two main reasons behind this drawback-machine error and type of gesture. For a few moments Kinect Motion Sensor Device failed to capture the real depth data. During this time machine will also capture a gesture in a wrong way. Thus the program won't be able to translate or process the gesture. Another main reason is the complexity of input gesture. The success rate of this program is more for the simple gesture comparing to the complex gesture. The gesture become more complex the success rate decreases. Here simple to complex gesture is taken as experimental data. Fig. 13 shows the success rate (%) for different gesture input provided by the volunteers.

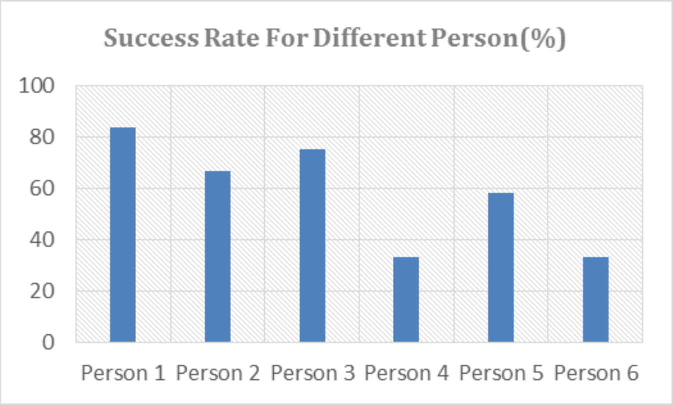


Fig. 12. Success rate for different person

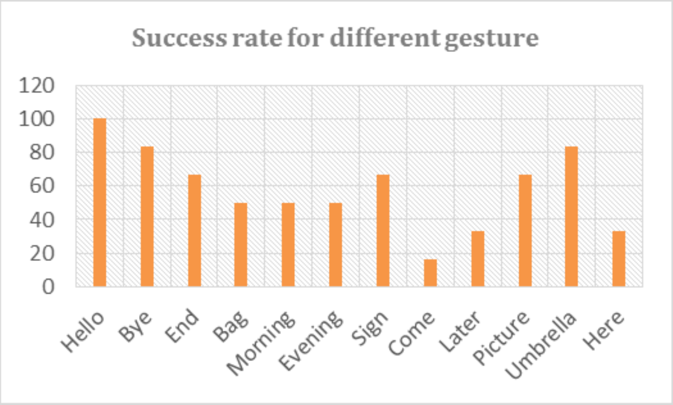


Fig. 13. Success rate for different gesture

V. Conclusions

Sign language is a visual way of communicating using hand gestures, facial expressions and body language. Visual communication methods have been around for thousands of years and nowadays there are hundreds of different types of sign languages in use across the world. It is a language used in visual-spatial modality. It is a natural language as sophisticated and complex as any speech language. It has been primarily developed and used by culturally Deaf people. In this paper, mainly three things have been done: Capture the gesture, Storing in the database or comparing with database and show the output. Though, in some cases the system cannot translate the gesture even though the input gesture is correct. In those cases, algorithm can be slightly modified according to the decreasing the error rate. In current system during training mode user has to write the input word. Future plan is to introduce Text-to Speech (TTS) system. That means in future user don't have to write the word during training mode. Rather he/she will provide input using his/her voice and the system will take his/her speech as input. Than the system will convert his/her speech into text and will translate those text as like current system.

References

[1] D. M. Capilla, “On Sign Language Translator Using Microsoft Kinect XBOX 360,” M. Sc. Thesis, University of Tennessee, Knoxville, USA.

[2] A. Farhadi, D. Forsyth, “Aligning ASL for statistical translation using a discriminative word model,” University of Illinois.

[3] S. Young, “ HTK: Hidden Markov Model Toolkit VI.5,” Cambridge Univ. Eng. Dept., Speech Group and Entropie Research Lab. Inc., Washington DC.

[4] P. Mekala, Y. Gao, J. Fan and A. Davari, "Real-time sign language recognition based on neural network architecture," in Proc. 2011 IEEE 43rd Southeastern Symposium on System Theory, Auburn, AL, 2011, pp. 195-199.

[5] Y.F. Admasu, and K. Raimond, “Ethiopian Sign Language Recognition Using Artificial Neural Network,” in proc. 10th International Conference on Intelligent Systems Design and Applications, 2010. 995-1000.

[6] J. Atwood, M. Eicholtz, and J. Farrell, “American Sign Language Recognition System. Artificial Intelligence and Machine Learning for Engineering Design,” Dept. of Mechanical Engineering, Carnegie Mellon University, 2012.

[7] L. Pigou, S. Dieleman, P.-J. Kindermans and B. Scharauwen, “Sign Language Recognition Using Convolutional Neural Networks,” in proc. European Conference on Computer Vision, 6-12 September, 2014

[8] D. Aryanie, Y. Heryadi, “American Sign Language-Based Finger-spelling Recognition using k-Nearest Neighbors Classifier,” in Proc. 3rd International Conference on Information and Communication Technology, 2015, pp. 533-536.

[9] R. Sharma, Y. Nemani, S. Kumar, L. Kaneand P. Khanna, “Recognition of Single Handed Sign Language Gestures using Contour Tracing descriptor,” in Proc. World Congress on Engineering (WCE 2013), vol. 2, July 3 - 5, 2013, London, U.K.