LEAP for DEAF: A Sign Language To Speech Translator

Bhushan R. Kadam[1] Komal K. Supekar[2] Harshad K. Chikhale[3] Priyanka B. Salve[4]
[1,2,3,4]NDMVP's KBT COE, Nashik

Abstract— Sign language is vital for facilitating communication between hearing impaired and the rest of society. Researchers in Sign language recognition had tailored completely different sensors to capture hand signs. Gloves, digital cameras, depth cameras and Kinect were used instead in most systems. Owing to signs closeness ,input accuracy is a terribly essential constraint to achieve a high recognition accuracy. Our aim is to design a Sign Language to Speech Translation system for ISL(i.e. Indian Sign Language) based on a brand new digital 3D motion detector referred to as Leap Motion, that consists of two inbuilt cameras and three infrared sensors that capture 3-D dynamic hand gestures. The palm-sized Leap Motion sensor provides far more portable and economical solution than Cyber glove or Microsoft Kinect employed in existing studies . This sensor tackles the major issues in vision-based systems like skin color, lighting etc... The planned system will make use of DTW algorithm as a classifier for converting hand gestures into an appropriate text as well as an audible speech.

Key words: Leap Motion Sensor, Gesture Recognition, Indian Sign Language, Human Computer Interaction, Machine Learning

I. Introduction

Disabled individuals face non-ending difficulties after they wish to touch upon the new technologies: the utilization of a computer, the access to information, editing and printing a text, etc. Reading a document can be an especially complicated task, despite their simplicity for traditional user. Today, the extraordinary progress of the new technologies, absolute to the data processing and the internet, offers exceptional opportunities to bring a much better quality of life to people who endure handicap and disabilities. The target community addressed by the project has its own language, referred to as Sign language. A Sign language could be a language that uses manual communication rather than sound to convey meaning, simultaneously at the same time combining hand shapes, orientation and movement of the hands, arms or body, and facial expressions to fluidly express a speaker's thoughts. The Sign language remains however a fully-fledged language, with its own constructional methodology of the sentences. The Indian sign language (ISL) faces an occasional diffusion level among the deaf as well as the hearing communities. To be effective, the communication of a deaf person needs the data of ISL, not solely on the speaker's side, however conjointly on the listener's. It's conjointly essential that each speak a similar sign language. Owing to the shortage of information or lack of availability of the Sign language, it's vital to enhance it, permitting this group to possess access to information in their initial language, i.e., the Sign language.

There are a lot of sensors and technologies present today which detect motion and gestures. But the accuracy and speed of these sensors are not much [2][1]. In existing analysis for Sign language recognition, the image recognition of color pictures, depth images and hand shapes are used. Since it must be taken with colored gloves, the glove worn isn't convenient. The image recognition needs long computation time to detect the hand and fingers. Within the case of the recognition with Kinect [7], massive space is needed for skeletal tracking. It's tough to recognize the sign gestures anyplace with Kinect. Therefore, Sign language recognition is needed employing a compact device which will directly acknowledge the shape of fingers or hands anyplace.

In this paper, we tend to propose a Indian sign language (ISL) recognition technique using Leap Motion Controller [8][3]. Putting hands and fingers over a Leap Motion controller, the ISL recognition is performed. The main advantage of using LEAP motion sensor is that, it is very accurate and very fast. It is even fast enough to detect handwriting. Leap Motion has skeletal tracking that acknowledges the framework of fingers to get a extremely accurate varied data such as the position of finger bones and the degree of the thumb and index finger. Additionally, the utilization of Leap Motion permits ISL recognition without any physical contact.

In this project, users will able to identify 26 alphabets of the ISL. The recent trend for learning sign language is via an instructor or through typed instructions. But, after this technology, the delay for learning the language can be minimized. There will be no more wait for further instructions. This can reduce the learning time and greatly aid the differently abled.

The rest of the paper is constructed as follows. Section II provides related works of Sign language recognition thoroughly. Section III explains a ISL recognition methodology. In Section IV, we perform experiments using the proposed technique.

II. Related works

Most of the scrutiny works in Sign language recognition is performed on Sign languages aside from ISL. Of recent, this area is a gaining vogue among analysis professionals. Computer vision and pattern identification techniques [6], involving object perception, feature distillation, clustering, and classification, have been successfully utilized for several sign-gesture recognition systems. Image processing techniques like analysis and detection of structure, texture, color, motion, optical flow, image improvement, segmentation, and contour modeling, have additionally been found to be effective. Connectionist approaches, incorporating Multi-layer Perception (MLP), Time Delay Neural Networks (TDNN), and radial basis function networks (RBFN), have been manipulated in gesture identification as well. Whereas static gestures (pose) recognition can generally be accomplished by stencil matching, standard pattern identification, and neural network, the dynamic gesture recognition problem incorporates the utilization of techniques like time compressing template, dynamic time warping (DTW), HMMs, FSM etc. [7].

The earliest research works on Sign language recognition is especially based on data glove based strategies. These strategies aren't user friendly and are more costly. Later some systems are developed for vision based recognition objective. They compared the proficiency of glove based systems as well as vision based systems and realize that vision based systems are more economical[8]. Several researchers conducted surveys that focused on the research works in vision based sign language recognition [10]. Here data acquisition, feature distillation and classification techniques were applied for the analysis of sign language gestures.

Xiaodong Yang and YingLi Tians paper on Eigen Joints [11] utilized for activity recognition helped us in diagnosing local features as the distance from the palm centroid to each fingertip. Yang Mingqiang, Kplama Kidiyo and Ronsin Joseph [9] have classified different strategies by which a curve will be segmented and used for feature distillation. Here strategies just like the quad tree depiction of curves and bag of features technique have been mentioned. Geetha M and Manjusha UC [8] have manifest the use of B spline curves for the identification of static hand gestures. They have made use of the Chain code technique by dividing the space into eight octant and plotting the maximum Curvature Points (MCP) within this space.

Many of the works focus on recognition of a small subset of signs. As already mentioned, for most of them difference in viewpoint (or alternatively, different rotation, position and scaling of a hand), environment and subject appearance, represents significant difficulty. Although some of the previous research trials accomplished a relatively high recognition rate, this rate dramatically degrades in real environment due to change in capturing speed, distance, skin color, etc.... the traditional sensors such as digital cameras are not sufficient to build up a realistic translation system. Some recent sensors such as Microsoft Kinect and Leap motion appeared for game playing issue. Researchers started to think to customize them in sign language recognition systems.

Microsoft Kinect is one of the modern sensors used to recognize sign language. It produces a live stream with depth information, body motion and skeletal movements. As Kinect produces a live stream of the tracked object, it provides information about the full body parts while moving such as: the parts position, velocity and direction. And the main advantage of this sensor is that the data is independent of light condition, as it uses infrared light to detect surrounding objects. Many researchers used kinect to recognize Sign language. In 2013 Billiet et al [5] built a rule based system on 8 different hand postures Kinect images and accomplished 96% recognition rate [10, 11]. Although Kinect facilitates body and hands tracking, it does not support hand shape recognition.

Among the above studies, within the case of coloured glove based recognition, the user might feel a difficult to wear them and a photograph image must be taken with wearing the coloured glove, that isn't very convenient. In the case of Kinect based recognition, an oversized area is needed to obtain depth further as and color image for the skeletal tracking, that isn't simple for ordinal use. In the case of Leap Motion based recognition, the machine learning needs giant computation for a replacement person.

III. System Overview

In this paper, a Leap Motion device is utilized to capture the hand signs in 3D digital data. The input data is exposed to a proposed recognition system for ISL recognition using DTW classifier based on Machine Learning approach.

A. *Leap Motion Controller*

The leap motion controller (LMC) is a device recently developed by Leap Motion Company. It detects and tracks hands, fingers and finger-like objects reporting distinct position and motion. It operates in an exceedingly close proximity at a rate of 200 frames per second [13]. The LMC field of view is an inverted pyramid of about 8 cubic feet centered on the device. The effective range of the LMC extends from about 1 inch to 2 feet on top of the device. The LMC uses 2 high exactitude infrared cameras and 3 infrared LEDs as shown in Fig.1, to capture hand data within its active range. However, it doesn't offer pictures of detected pictures. Its driver software package processes the acquired data, extracts positions and other alternative information like shown in Fig.2 using complex mathematics [14].

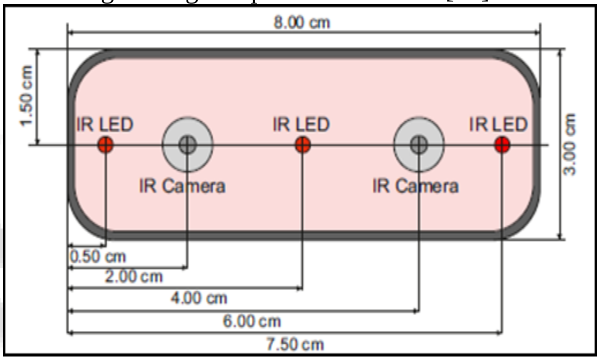

Fig. 1: Schematic view of leap motion controller (LMC)

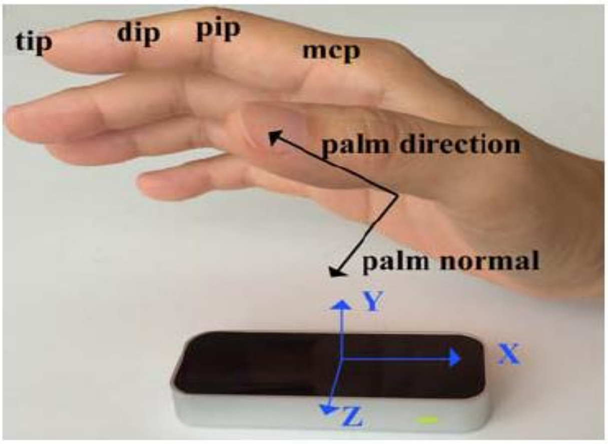

Fig. 2: Leap motion sensor with its co-ordinates and data supported by APIs

When it detects hand and fingers, the Leap Motion software assigns it a novel ID tag. The ID remains constant as long as that entity remains visible among the devices field of view. If tracking is lost and regained, the software might assign for it a replacement ID [4].

B. System Design

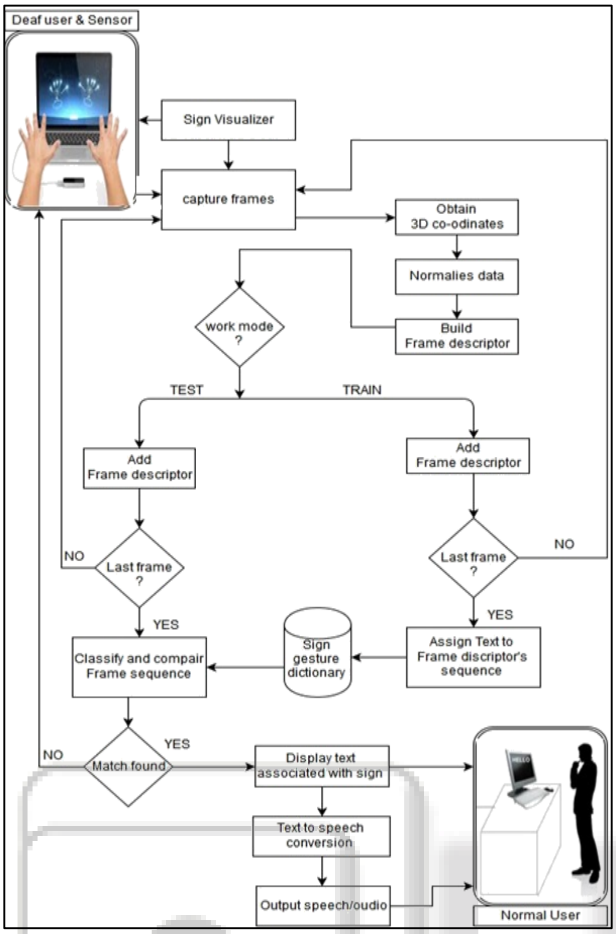

Fig. 3: System Architecture

As per Fig.3, the deaf user performs a sign over Leap motion controller or getting ready to do so. With a frame rate of 50 fps, a replacement frame is obtained and therefore the video stream is updated with the skeleton of the users hand overlapped onto it. At that time, if the user needs to record a sequence (otherwise, the system asks the camera to induce successive frame), three main blocks are executed: the primary block consists of getting the frame sequence generated by Leap Motion API, the second block consists of normalizing these information, and the third one consists of building the frame descriptor. Then, if the operating mode is set to training (meaning that the user is adding a new sign to the training set), the frame descriptor is added to the correspondent file of the sign gesture dictionary. Otherwise, if the mode is set to TESTING (meaning that the user needs to translate the sign that's been done), the frame descriptor is added to the present test sample. Then, the system checks if the present frame is the last frame of the sign. once a sign is finished and if the operating mode is TESTING, the test sign is compared employing a classifier with the signs from the dictionary and the corresponding speech as well as relative text output is displayed so the normal user can recognize the corresponding relative word in the spoken language. After that, the system keeps going with successive frame and the flow of the block diagram is repeated again.

IV. METHODOLOGY

A. Feature Extraction

	Palm		Fingers	
	Name	Type	Name	Type
Hand features[a]	Normal	vector	Direction	vector
	Position	vector	Length	in mm
	Velocity	vector in mm/sec	Tip position	vector
	Confidence	a float in [0, 1]	Tip velocity	mm/sec
	Pinch strength	a float in [0, 1]	Dip position	vector
	Grab strength	a float in [0, 1]	Pip position	vector
	Sphere center	vector	Mcp position	vector
	Sphere radius	in mm		

[a] Features were obtained using API version 2.0.2.

Fig. 4: Features obtained from the API

Fig. 4 lists the hand and finger features obtained from the API during this study. Palm related features embrace palm normal (a unit directional vector), palm position (the center position of the palm), and speed (in millimeter per second). The API conjointly reports a float range representing the confidence of the data's accuracy. Grab strength is additionally a float range between zero and one, with zero indicating open hand and one for closed hand. Almost like grab strength, pinch strength indicates the openness in between thumb and any other finger of a similar hand. Sphere center and radius are calculated supported the estimated sphere placed as if the hand were holding a ball. We conjointly obtained features for each individual five fingers from the API. We collected finger's direction (a unit directional vector) and length (in millimeter). Positions of each finger joints between distal phalanges and intermediate phalanges (i.e. DIP), proximal phalanges and intermediate phalanges (i.e. PIP), metacarpals phalanges (i.e. MCP), and tip positions (i.e. TIP) were conjointly recorded as shown in Fig. 4. Additionally, we conjointly collected the tip velocity of every finger [1].

B. Normalization of the data

The normalization must take into consideration the position of the user's hand. The deaf users hand may be at completely different positions and consequently the information should be stored consequently to that position. A slight variation in depth will cause a substantial variation of the X and Y values. The distances between one joint and another one can drastically vary depending on the position of the user's hand [11]. Rather than directly storing the cartesian coordinates X,Y, and Z (which are often obtained using Leap Motion API), the proposal consists in normalizing all the joint coordinates with regard to the position of the Palm. This position remains continually constant along the sign frames and is that the right one to be used to make the system position-invariant. Rather than using the cartesian coordinates X,Y, and Z, the spherical coordinates considering Palm position as the origin are stored. In mathematics, a spherical coordinate system is a coordinate system for three-dimensional space where the position of a point is specified by three numbers: the radial distance of that point from a fixed origin (r), its polar angle measured from a fixed zenith direction (ϴ), and also the azimuth angle of its orthogonal projection on a reference plane that passes through the origin and is orthogonal to the zenith direction, measured from a fixed reference direction on that plane (ɛ) [2]. Fig. 5 shows these three numbers or values and also the correspondence of these three values in the system.

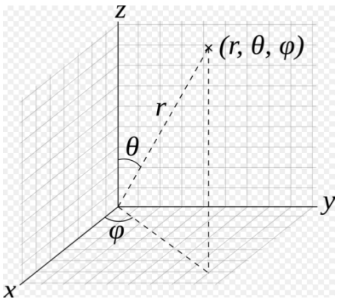

Fig. 5: Use of the spherical coordinates

The radial distance r will be expressed by d and defines a vector between the Palm and the correspondent finger joint.

Once the Finger joints data are obtained and normalized, successive step is building a descriptor for every sign. As shown in Fig.6, the descriptor must be able to describe a sign in a means that this descriptor will be distinctive and sufficiently different from alternative descriptor of dictionary [2].

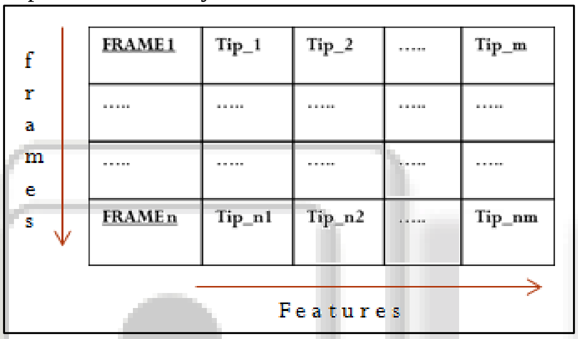

Fig. 6: Sign Frame Descriptor

When the Impaired User will perform the sign over the device, actions are captured and accordingly values are generated through distance calculation module. This numeric values get stored within the Sign Gesture dictionary i.e. when user add signs dynamically, actions are recorded and its values get stored within the text file. Whenever user performs any action, numeric values get appended into the dictionary. in this manner dictionary is maintained, that helps to spot the actions.

C. Classification

The classifier is the function that will out the corresponding word of the spoken language once the deaf user inputs the sign. Given an input sequence of frames, the classifier match it with sequence of frames from the default dictionary. The problem here is that two compared sequences doesn't share the same length. Thus to tackle this constraint, a DTW classifier is developed.

1) Dynamic Time Warping (DTW):

Dynamic time warping (DTW) was introduced in 1960s and it's an algorithm for measuring similarity between two sequences which may vary in time or speed. as an example, similarities in sign gesture patterns would be detected, even if in one video the person is performing a sign gesture slowly and if in another video he or she is performing the same sign more quickly, or even if there are accelerations as well as decelerations throughout the course of one observation. By using DTW, a computer will be able to notice an optimal match between two given sequences (i.e. signs) with certain restrictions. The two distinct sequences are warped non-linearly within the time dimension in order to determine a measure of their similarity independent of indisputable non-linear variations within the time dimension. in this project, DTW is satisfactorily used for sign gesture recognition purposes, coping in this means with sign executions speeds [2].

2) Sign Gesture Recognition Based on DTW

During the sign training process, a frame descriptor and threshold range for every class of sign is computed. Within the real-time recognition stage, the Dynamic Time Warping (i.e. DTW) algorithm measures the similarity between the input sign and also the sign dictionary. The input sign can either be accepted as a member of the class to which it has the minimum normalized aggregate warping cost distance, or if the similarity measurement doesn't match the threshold range, rejected as belonging to none of the categories [3].

A sign frame descriptor can be computed by recording a single or N training sample(s) for every class of gestures that must be recognized. The frame descriptor for every class of sign can be found from the recorded training samples by computing the distance between each of the N training samples. The training sample within the given category, that has the minimum normalized aggregate warping cost distance in comparison against the N–1 training samples, is recognized as the desired sign gesture for that class. The classification threshold range for every sign gesture is calculated by taking the average normalized warping cost distance between the frame descriptor and the alternative N−1 training examples for that gesture.

Using a classification threshold for every sign gesture overcomes the problem of false positives throughout the recognition stage, as unknown time series input is classified as a null class if no one match is found within the sign gesture dictionary. If a new sign gesture class is incorporated to an existing sign in the dictionary, or if an existing sign is removed, the sign recognition model doesn't need to be retrained. Instead, we need only train a new sign frame descriptor and threshold range for the new sign, that thus reduces the training time.

Once the DTW algorithm has been trained, an unknown multidimensional time-series input sign can be classified by calculating the normalized aggregate warping cost distance between the input and each of the frame descriptor within the sign gesture dictionary. The input gesture is then classified according to the sign class corresponding to the minimum normalized aggregate warping cost distance.

The DTW method [`12] is determined as follows. If X and Y are two time- series sign gesture sequences with distinct lengths, where $X = \{x_1, x_2, \ldots . x_n\}$ and $Y = \{y_1, y_2, \ldots . y_n\}$ a collective distance matrix D, which represents a mapping and alignment collaboration between X(i) and Y(j), is formed to compute the similarity between distinct sequences X and Y. Afterwards, a warping path $W = \{w_1, w_2, \ldots . w_p\}$ composed of the Local cumulative distances D(i, j) is computed. The length of warping path is:

$$max\ (n, m) \leq P < (n + m)$$

and the kth element of the warping path is specified by :

$$w_k = (x_i, y_i)$$

To enhance the proficiency of DTW, we oblige the warping path so that the utmost permitted warping path cannot diverge too far from the diagonal. The restrictions placed on the warping window speeds up the DTW computation. The constraints placed are as follows. The warping path must initiates at the start of each time series, *i.e., at* w_1 *= (1,1),* and end *at* w_p *= (n,m).* This assures that every index of both time series is manipulated in the warping path. The warping must be continual; *i.e.*, if w_k = *(i, j)*, then w_{k+1} = *(i, j)*, must equal either *(i, j),(i+1,j), (i,j+1),or(i+1,j+1).* The warping path must manifest uniformity behavior. The optimal warping path that minimizes the normalized aggregate warping cost distance is given by:

$$d(w) = \frac{1}{p}\sum_{k=1}^{p} d(w_{k_i}, w_{k_j})$$

Where $d(w_{k_i}, w_{k_j})$ is the Euclidean distance point *i* in time series X and *j* in time-series Y, is stated by w_k. The minimum optimal aggregate warping path can be effectively found by applying dynamic programming through the cumulative distance *D(i, j)* given by:

$$D(i, j) = d(x_i, y_i) + \min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\}$$

The DTW(X,Y) between the two distinct time series sequences is then calculated by finding the minimum normalized aggregate warping cost distance between X and Y. This is defined as:

$$DTW(x, y) = \frac{1}{p}\sum_{k=1}^{p} d(w_{k_i}, w_{k_j})$$

The proposed DTW-based recognition algorithm [12] enables early recognition of sign gestures and translation to relative text and speech according to learned sign gestures.

V. EXPERIMENTAL RESULT

In this section, the accuracy of the system for the implemented approaches and configuration of parameters is analyzed. The default training set contains a total of 780 different samples, which is the result after adding 10 different frames of 3 different samples for each of the 26 signs from the dictionary of ISL alphabets as shown in Fig.7.

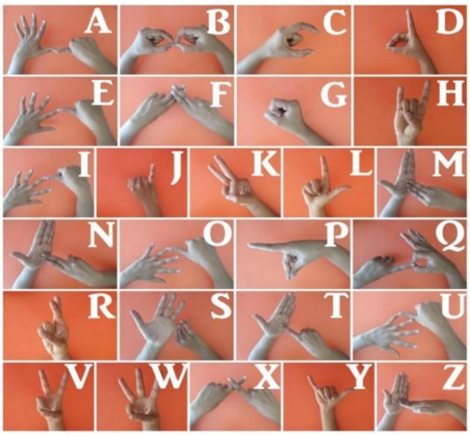

Fig. 7: ISL signs

Before classification is done, the classifier is trained with part of the data. This was done using five- fold cross validation. Cross validation provides a frame work for creating several train and test splits and guaranteeing that each data points appears in the test set at least once. The procedure is as follows:

Splits the data into n-equal sized groups, For i = 1 to n, a) Select group i to be the test set and all other (n-i) to be training set, b) Train the model on the training set and evaluate on the test set.

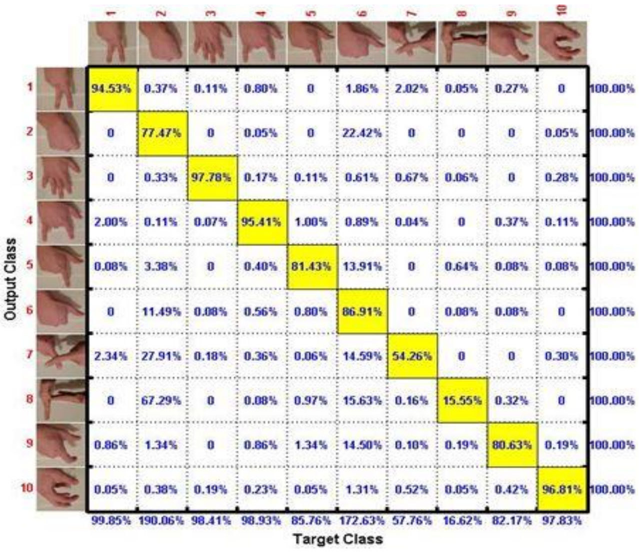

Fig. 8: Recognition rates of classes for feature set 10

Three samples of each of the 26 ISL alphabet signs were collected from a single signer. Ten frames were acquired from each sample letter sign, to provide a total of 780 frames of data. Ten features were selected from Fifteen features provided by the LMC for the representations of each frame in the coverage area of the LMC. As shown in Fig.8, the best accuracy of the sign recognition achieved using the DTW was about 97.78% while Analysis of the misclassified signs (18 out of 780 frames) revealed that not all misclassified signs are similar to the signs they are classified to. This indicates that the problem could be due to the fact that the LMC detects the hands and fingers movement from one side. Thus some of the fingers may be occluded by others which reduce the reparability of the some signs. In future work we will use two LMCs, one in the front and the other on the side of the area of sign articulations. Several methods for combining the features from both LMCs will be investigated to analyze. Further study on the use of two LMC for ISL word signs and full sentence recognition will be investigated in the future.

VI. CONCLUSION

This paper presents a proposed system to recognize 26 alphabetic signs from ISL using a new sensor, Leap motion. The sensor has the ability to track hands and finger positions and movements in frames and showed superiority on Kinect that can't distinguish between fingers. The proposed system builds a DTW Classifier to classify the input data into output sign meaning. System is working since the best configuration achieves 98.3% accuracy. Despite the fact that the defined signs do not belong to a specific official Sign Language, the idea of the project is to show that with basic descriptors and classifiers and the use of Leap Motion sensor, a wide number of signs could be recognized and the system has the potential to provide a computationally efficient design without sacrificing the recognition accuracy compared to other similar projects. This work can be extended by adding other features such as articulation and orientation. Also, adding a new leap motion device resolves the issue of hands orientation relative to the device. This new sensor can be a dramatic achievement for sign language recognition systems as it can track the manual features of the sign.

ACKNOWLEDGEMENT

This research work was support by Prof. V. S. Tajanpure, NDMVP'S KBT COE Nashik. We thank her for guiding us and providing insight which greatly assisted our research work. We also thank Prof. B. S. Tarle, H.O.D. NDMVP'S KBT COE Nashik for his constant motivation. We would also like to show our gratitude to Dr. Prof. Jayant T. Pattiwar, Principal NDMVP'S KBT COE Nashik and Management of NDMVP Samaj for providing all necessary facilities and their constant encouragement and support.

REFERENCES

[1] C. Chuan, E. Regina and C. Guardino, "American Sign Language Recognition Using Leap Motion Sensor", 2014 13th International Conference on Machine Learning and Applications, 2014.

[2] D. Martinez, "Sign Language Translator using Microsoft Kinect XBOX 360TM", VIBOT 5., Department of Computer Science, Computer Vision Lab, University of Tennessee, 2013.

[3] P.Karthick, S.Thanalaxmi, "Transforming Indian Sign Language into text using Leap Motion", International Journal of Innovative Research in Science Engineering and Technology (An ISO 3297), Vol. 3, 2014.

[4] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the leap motion controller", Sensors (Basel, Switzerland), vol. 13, no. 5, p. 6380, 2013.

[5] L. Billiet, J. A. Oramas Mogrovejo, M. Hoffmann,W. Meert, and L. Antanas, "Rule-based hand posture recogultlOn usingqualitative finger configurations acquired with the kinect", ln Proceedings of the 2nd International Conference on PatternRecoguition Applications and Methods, pages 1--4, Feb.2013.

[6] R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis", New York: Wiley, 1973.

[7] Sushmita Mitra, Senior Member, IEEE, and Tinku Acharya, Senior Mem-ber, IEEE, "Gesture Recognition: A Survey", IEEE Transactions on Systems, Man, And Cybernetics part C: Applications and Reviews, VOL. 37, NO. 3, MAY 2007.

[8] Geetha M and Manjusha U C, "A vision based Recogniton of Indian Sign language Alphabets and Numerals using Bspline approximation", INTERNATIONAL JOURNAL ON COMPUTER SCIENCE AND EN-GINEERING (IJCSE), VOL. 4, NO. 3, MARCH 2012.

[9] Yang Mingqiang, Kplama Kidiyo and Ronsin Joseph, "A survey of shape feature extraction Techniques", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICSPART C: APPLICATIONS AND REVIEWS, VOL. 30, NO. 2, MAY 2000.

[10] Sylvie C.W. Ong and Surendra Ranganath, "Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE.

[11] Xiaodong Yang and YingLi Tian, "Eigen Joints Based Action Recognition Using Nave Bayes Nearest Neighbour", THE CITY COLLEGE OF NEW YORK, NEW YORK.

[12] S. Patil, H. Chintalapalli, D. Kim, and Y. Chai, "Inertial sensor-based touch and shake metaphor for expressive control of 3D virtual Avatars," Sensors, vol. 15, no. 6, pp. 14435–14457, Jun. 2015.

[13] http://dartmouthbusinessjournal.com/2013/08/the-leap-motion-controller-and-touchless-technology.

[14] https://forums.leapmotion.com.

