Neurocomputing 259 (2017) 21–38

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A multimodal framework for sensor based sign language recognition

Pradeep Kumar[a,*], Himaanshu Gauba[a], Partha Pratim Roy[a], Debi Prosad Dogra[b]

[a] Department of Computer Science & Engineering, IIT Roorkee, India
[b] Department of Computer Science & Engineering, School of Electrical Sciences, IIT Bhubaneswar, India

ARTICLE INFO

Article history:
Received 22 February 2016
Revised 17 July 2016
Accepted 4 August 2016
Available online 21 February 2017

Keywords:
Sign language recognition
Gesture recognition
Multimodal framework
Sensor fusion

ABSTRACT

In this paper, we propose a novel multimodal framework for isolated Sign Language Recognition (SLR) using sensor devices. Microsoft Kinect and Leap motion sensors are used in our framework to capture finger and palm positions from two different views during gesture. One sensor (Leap Motion) is kept below the hand(s) while the other (Kinect) is placed in front of the signer for capturing horizontal and vertical movement of fingers during sign gestures. A set of features is next extracted from the raw data captured with both sensors. Recognition is performed separately by Hidden Markov Model (HMM) and Bidirectional Long Short-Term Memory Neural Network (BLSTM-NN) based sequential classifiers. In the next phase, results are combined to boost-up the recognition performance. The framework has been tested on a dataset of 7500 Indian Sign Language (ISL) gestures comprised with 50 different sign-words. Our dataset includes single as well as double handed gestures. It has been observed that, accuracies can be improved if data from both sensors are fused as compared to single sensor-based recognition. We have recorded improvements of 2.26% (single hand) and 0.91% (both hands) using HMM and 2.88% (single hand) and 1.67% (both hands) using BLSTM-NN classifiers. Overall accuracies of 97.85% and 94.55% have been recorded by combining HMM and BLSTM-NN for single hand and double handed signs.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Sign language is the primary way of communication for hearing impaired people. The communication is performed through the use of hand gestures, movements of arm/body, facial expressions, and lip movements. In recent years, success of automatic Sign Language Recognition (SLR) system has opened-up a new way of Human-Computer-Interaction (HCI) that can convert sign gestures into text/speech to enable communication between people who do not understand sign language. Researchers have used various sensors such as data gloves [1], digital camera [2], accelerometer [3], depth camera [4], Kinect [5] or Leap motion controller [6] to capture sign gesture inputs. The vision based approaches generally focus on feature extraction from 2D images and videos. If the gestures are recorded using camera, automatic segmentation and recognition becomes difficult due to various environmental noises such as self occlusion, background noise or illumination variation. Especially, signs that are typically represented by 3D gestures, are more prone to such noises.

Introduction of low cost depth sensors such as Kinect or Leap motion has made capturing of 3D data more convenient. The Kinect sensor comes with its associated Software Development Kit (SDK) and openNI library that enables to acquire the 3D skeleton of the whole body and is currently used in various appl ications including gaming, robotics, and gesture recognition. However, the device is not able to localize small body parts such as fingers or palm and the system is limited to recognize motion gestures like waving, move up/down, left/right, and forward/backward. In order to extract the hand and finger movements, SDK developers have already provided interfaces that can be combined with machine learning algorithms during feature extraction [7]. Some of the existing work that use Kinect for similar tasks are as follows. Raheja et al. [8] have developed a method to track fingertips and the center of the palm using Kinect. The system is highly accurate (nearly 100%) fingertip detection when fingers are extended, and the detection of palm center is around 90%. However, they did not experiment and evaluate their work for gesture recognition. Similar approaches for palm and fingertip detection with convex hull and contour features extracted from the segmented hand, is presented in [9] and [10]. A 14-patch color glove based approach is presented in [11] for hand detection and pose estimation. In our work, we have used Candescent NUI [7] library that provides information on fingertip movement. We modified the library to store the 3D view of palm and finger movements during sign language inputs.

Leap motion controller [12] can acquire 3D data with a millimeter level precision at a sampling rate of 120 frames/s. It has been

* Corresponding author.
E-mail addresses: pradeep.iitr7@gmail.com (P. Kumar), gauba.himanshu@gmail.com (H. Gauba), proy.fcs@iitr.ac.in (P. Pratim Roy), dpdogra@iitbbs.ac.in (D. Prosad Dogra).

http://dx.doi.org/10.1016/j.neucom.2016.08.132
0925-2312/© 2017 Elsevier B.V. All rights reserved.

designed especially for hand and finger gestures. The device comes with inbuilt interface and SDK that provides direct access to various features such as fingertip points, palm position, palm orientation, etc. It has already been applied in serious gaming [13], palm and finger spine rehabilitation [14], upper limb rehabilitation [15], stroke rehabilitation [16], HCI [17], etc. However, the device has a small field of view as compared to Kinect. Therefore, its usage is limited to acquire hand and finger movements.

Our observation is that, the recognition systems developed using a single sensor, are not adequate to accurately recognize every symbols of a given sign language dictionary. This is because, it may not be possible to capture all articulations or movements using a single sensor. Though, there exists a few research work that have used multiple sensors to capture the sign input [18–20], however, their recognition performance can further be improved. Also, we have not found any multi-sensors based work applicable for recognition of Indian Sign Language (ISL) gestures. Thus, we have proposed an efficient multimodal framework to recognize ISL gestures with the help of Kinect and Leap motion sensors. Both sensors simultaneously capture the input signs from two different views and we fuse them to recognize a gesture. The Kinect sensor is placed in front of the user to capture finger and palm movements visible in *Coronal* plane and a Leap motion controller is placed below the hand to capture movements on *Axial* plane. It has been shown in this work that, both inputs complement each others and we have fused them to obtain final recognition outputs.

2. Motivation of the work

The existing SLR systems are inadequate to acquire different types of movements that may take place during sign language communication. These systems usually fail to recognize the input signs when the position of the palm is not directly facing the camera (Kinect) or sensor or when performing gestures such that one hand/finger covers the other (Leap Motion). In such situations, it becomes hard to estimate the input sign. Some examples are depicted in Fig. 1(a) and (c). It may be observed that, a Kinect based system fails to detect the movement of fingers when the hands become parallel to the optical axis of camera or occlusion occurs. Similarly, Leap motion sensor fails to detect the sign input when one hand or finger covers another as shown in Fig. 1(b) and (d).

Since both Kinect and Leap motion sensors share common characteristics (capturing depth information from input), a combination of them can be handy and seems promising as a new multimodal framework for gesture recognition. The framework can combine data from both devices and we can build a robust SLR system. The framework seems to be robust to occlusion and other environmental noises which is important while designing system to recognize dynamic sign gestures.

2.1. Contributions of the paper

We have made the following contributions while developing the above mentioned multimodal framework.

1. Our first contribution is combination of features extracted using Kinect and Leap motion sensors to describe gestures representing various words of ISL. We have shown through rigorous experiments that, the proposed feature vector is robust against occlusion and other types of environmental noises.
2. Our second contribution is combining Hidden Markov Model (HMM) & Bidirectional Long Short-Term Memory Neural Network (BLSTM-NN) classifiers for improving the accuracy of the recognition phase. We have carried out several experiments to validate our claim using a large dataset comprises with 7500 independent words representing Indian sign gestures.

Rest of the paper is organized as follows. In Section 3, an in-depth review on existing methods, is presented. We present the proposed system design including preprocessing, feature extraction, and feature combination in Section 4. Recognition results are presented in Section 5. Finally, we present conclusion and future work in Section 6.

3. Related work

Research in SLR systems has gradually shifted from image based 2D to 3D using depth analysis. Most of the 2D approaches are camera based and rely on computer vision algorithms for extracting meaningful information from the image. However, 3D hand gesture recognition is a novel way and it has increased the interest of researchers in various applications like SLR, robotics, daily assistance and gaming [21]. The use of emerging 3D depth sensors such as Kinect and Leap motion controller has opened-up several challenging research problems in various fields. In this section, we present some of the existing gesture recognition methods. A summary of the existing work is given in Table 1.

3.1. 2D Image based gesture recognition

There exist a few work that accomplish gesture recognition using 2D images. Here, sign inputs are captured in 2D using a single camera. The classification of gestures rely on the information (features) extracted using various image processing techniques such as shape detection, segmentation, contour modeling, color, and motion. In [22], the authors have proposed a real time hand tracking approach with a colored-glove using single camera. The authors have employed a k-Nearest Neighbor (k-NN) approach to identify the color pattern imprinted on the glove. Chen et al. [23] have proposed a continuous gesture recognition approach using background modeling, skin color, edge and motion based features. The approach did not require any colored-glove and HMM based classification has been used for gesture recognition. However, their approach requires continuous hand motion. In [24], the authors have proposed a vision-based SLR system using video sequences. Hand segmentation and motion tracking was performed by skin color segmentation. Background modeling for removing static areas from the video sequence was done on pixel level by calculating the median of all pixels over time. The authors have extracted geometric features from the segmented images where an accuracy of 87.8% was recorded on 18 sign gestures when tested with HMM. In [25], the authors have proposed a skin-color segmentation based SLR system for Japanese Sign Language (JSL). At each frame, they tracked the face and hands using the skin color and elbows by matching the template of elbow shape. Overlapping of hands and face were distinguished using texture matching of previously tracked regions of hands and face. The authors have extracted geometric features and the system was tested on 65 JSL signs with HMM classifier. Starner et al. [26] have proposed a sentence level continuous SLR system using a single camera. The authors have used skin color for hand segmentation and hand-blobs were thus extracted from the scene. A feature vector of 16 dimension was computed from hand-blobs and HMM classifier was used to model each sign word. A total 40-word lexicon were used in their experiment and a recognition rate of 92% was recorded when tested with a dataset of 500 sentences. In [27], a SLR system for ISL was proposed using video sequences. The authors have extracted direction histogram based feature vectors with 18 and 36 bins. The system was tested on 22 ISL gestures performed using both hands where a recognition rate of 92.29% was achieved when tested with k-NN classifier. In [28], the authors have proposed a video sequence based SLR system for ISL alphabets. The authors have used skin color modeling for hand segmentation and

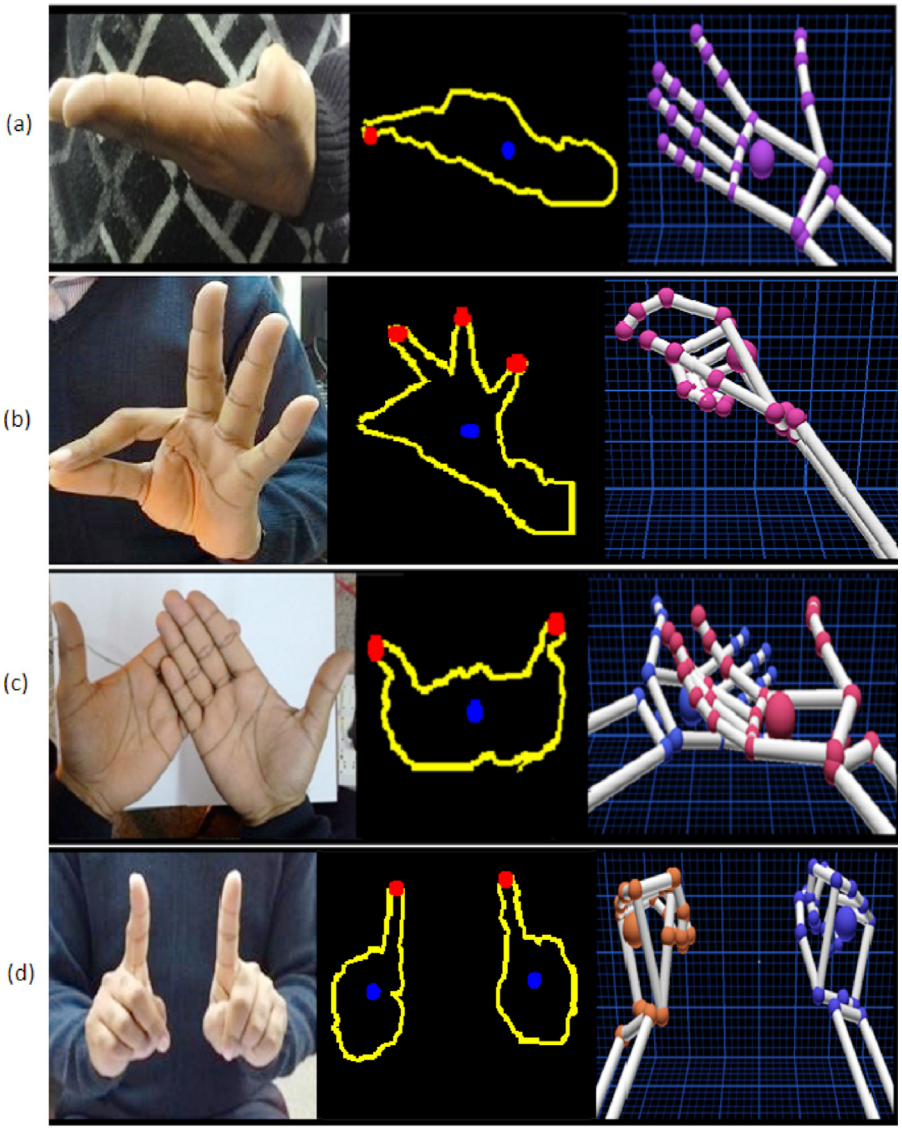


Fig. 1. Gesture recognition using single and double hand: (a) A single hand gesture ("*welcome*") that was hard to estimate in Kinect, however, captured by Leap motion sensor (b) A single hand gesture ("*office*") is captured better by Kinect (c) A double hand gesture ("*thanks*") that was not correctly decoded by Kinect (d) A double hand gesture ("*where*") that was not correctly decoded by Leap motion.

detection. Hand shape, texture and finger counts features were computed using Principle Curvature Based Region (PCBR), Wavelet Packet Decomposition (WPD-2) and convexity defects algorithm, respectively. The recognition process was performed on 26 ISL alphabets with k-NN, Dynamic Time Warping (DTW) and Support Vector Machine (SVM) classifiers where the best recognition rate of 91.3% was achieved using multiclass SVM classifier. Premaratne et al. [29] have proposed an approach for hand gesture tracking and recognition system. The authors have implemented Lucas–Kanade algorithm for hand tracking and motion invariant features were extracted for recognition of hand gestures. The system was tested on 8 hand gestures using SVM and Artificial Neural Network (ANN) classifiers.

Grzeszcuk et al. [30] have proposed a stereo-camera based gesture recognition technique. The authors modeled the users arm as a 3D line and found the arm's orientation and the location of hand. This information was used for modeling color-based hand segmentation. The recognition process was based on 2D geometric moments and Mahalanobis distance metric where a recognition rate of 96% was recorded on 6 gestures. In [31], a hand gesture recognition system using stereo color image sequence was proposed. The authors have used Gaussian model for skin color detection whereas a blob analysis was performed for hand detection and tracking. The system was tested on alphabets and numeric hand gestures where an accuracy of 94.72% was obtained when tested with HMM classifier.

3.2. Sensor data gloves based recognition

The data gloves based approaches make use of sensor devices such as accelerometer or gyroscope for capturing hand and finger movements. A framework for SLR has been proposed in [3] using combination of multi-channel electromyography (EMG) and 3D accelerometer sensors. The gesture segments were detected through EMG signals. The framework has been tested on 72 sign inputs using a combined HMM and Decision Tree based classification. In [1], the authors have proposed a sensor glove based SLR system. They have applied regression analysis for dimensionality reduction and to extract relevant features from the acquired data. The recognition of gestures has been done using k-NN based classification. A two-hand based SLR system is proposed in [32] using data gloves to capture 100 sign inputs from two signers. A Principal Component

Table 1
Summary of related work.

Method	Author name and year	Approach (feature and classifier)	Dataset and accuracy
2D Image and stereovision	Wang et al., 2009 [22]	k- Nearest neighbor (k-NN)	10 K image pairs, N.A.
	Starner et al., 1998 [26]	Skin color modeling, hand blobs using HMM	40-word lexicon, 500 sentences; 92%.
	Rekha et al., 2011 [28]	Skin color modeling, hand shape, texture and finger count using SVM	26 ISL gestures; 91.30%.
	Chen et al., 2003 [23]	Skin color, edge detection, motion based features using HMM	20 different gesture of TSL ; 90%.
	Grzeszcuk et al., 2000 [30]	Hand position, 2D geometric features using distance metric	6 gestures; 96%.
	Elmezain et al., 2008 [31]	Gaussian model, blob analysis using HMM	Alpha-numeric gestures; 94.72%.
Sensor data glove	Zhang et al., 2011 [3]	Velocity and statistical features using HMM	72 CSL words; 92.50%.
	Assaleh et al., 2012 [1]	Statistical features using k-NN	10 ArSL samples; 92.50%.
	Mohandes et al., 2013 [32]	PCA, statistical features using SVM	100 ArSL samples; 99.60%.
	Tubaiz et al., 2015 [33]	Statistical features using k-NN	80 words of ArSL; 98.90%.
Kinect	Lai et al., 2012 [5]	Statistical features using k-NN	8 simple gesture; 97.27%.
	Zafrulla et al., 2011 [35]	3D points of joints using HMM	60 phrases of ASL; 74.83%.
	Dominio et al., 2014 [56]	Distance, curvature and palm area features using SVM	10 gestures 100%, 12 ASL signs 97.60%
	Biswas et al., 2011 [36]	Motion feature using SVM	8 gesture; 82.52%.
	Ren et al., 2013 [38]	Distance metric using template matching	10 gestures; 93.20%.
	Pugeault et al., 2011 [41]	Gabor features using random forest	26 ASL alphabet; 75%.
	Lang et al., 2012 [42]	3D hand joints using HMM	25 GSL gesture; 97%.
	Li et al., 2012 [43]	3D hand joints using HMM	9 sign gesture; 84 %.
	Pedersoli et al., 2014 [44]	Gabor features using SVM	24 ASL alphabets; 70%.
	Almeida et al., 2014 [4]	Shape, movement and position of hands using SVM	34 BSL sign; 80%.
	Mehrotra et al., 2015 [45]	3D skeleton points, angular and distance features using SVM	37 ISL sign; 86.16%.
	Keskin et al., 2012 [39]	Scale invariant features using RDF	24 ASL sign; 84.30%.
	Kirac et al., 2014 [40]	3D hand points, regression forest and dynamic programming	12 ASL sign; 57.60% pose estimation.
Leap motion	Potter et al., 2013 [48]	3D finger points using ANN	26 AuSL alphabets; N.A.
	Chuan et al., 2014 [6]	Velocity, palm normal, pitch strength using k-NN and SVM	26 ASL alphabets; 72.78% and 79.83%.
	Elons et al., 2014 [49]	3D finger position, finger distance using MLP-NN.	50 ArSL gesture; 88%.
	Xu et al., 2015 [51]	Directional features	3D Chinese character recognition; N.A.
	Nigam et al., 2014 [52]	HOOF features using SVM	3D signature recognition; 91%.
Multimodal	ElBadawy et al., 2015 [18]	3D finger movement, body movement using ANN	20 ArSL gestures; 95%.
	Marin et al., 2014 [19]	Fingertips angle, distance, elevation, curvature and Correlation using SVM	10 ASL gesture; 91.28%.
	Fok et al., 2015 [20]	3D finger movements using HMM	ASL digits; N.A.
	Rossol et al., 2015 [12]	3D palm points, palm normal, hand direction etc. using SVM	3 different hand poses; 90.80%.
	Marin et al., 2015 [57]	3D fingertips, curvature and correlation using SVM	10 ASL gestures; 96.50%.
	Mihail et al., 2012 [54]	3D point cloud, PCA, angular features using k-NN	10 digit gestures; 90%.
	Hongo et al., 2000 [2]	Directional feature using LDA	Stone, paper, scissor gesture; 97.20%.

Analysis (PCA) based approach has been used for dimensionality reduction whose outcome was fed to SVM classifier for recognition. In [33], the authors have collected 40 sign sentences using a data glove and a camera has been used for data labeling. The recognition process has been performed using k-NN based classifier. However, the approach requires manual labeling of gestures for identifying the word boundaries which essentially leads to heavy training burden.

3.3. Kinect based gesture recognition

Kinect 1.0, released in 2010, is a popular device for sensing depth data in computer vision [34]. The sensor is able to provide skeletal view of the body which has been used by the authors of [5] to recognize 8 hand gestures in real time with an accuracy of 99%. Zafrulla et al. [35] have developed a hand gesture based automatic SLR system for deaf children. However, the performance of the system degrades due to tracking errors especially when the the user is seated. In [36], the authors have proposed a depth image based hand gesture recognition system that has been tested on 8 hand gestures. The depth images were used to remove the background of the users image using depth histogram techniques and to extract the hand position from rest of the body. Ren et al. [37,38] have developed a hand pose detection and gesture recognition system using depth and color information from Kinect. The authors have used a Finger-Earth Movers Distance metric (FEMD) for identifying dissimilarities between different hand shapes obtained from Kinect sensor. The system is insensitive to hand-pose variations and is able to operate in uncontrolled environments. The hand segmentation is done using depth stream and requires a specially designed black bracelet which is to be worn by the user. Keskin et al. [39] have proposed hand gesture recognition and pose estimation system using depth images. The authors have extracted scale invariant features from depth images and used Randomized Decision Forest (RDF) to perform per pixel classification where the final class label was determined by majority voting based scheme. The system was tested on 24 static sign gestures of ASL alphabets where a recognition rate of 84.3% was recorded. In other work, Kirac et al. [40] have proposed a 3D hand pose estimation technique using depth images. The authors have extracted the hand skeleton by finding the location of each joint through mode selection method and then applied dynamic programming for selecting the skeletal configuration as the final hand pose.

Pugeault et al. [41] have used Kinect to develop an interactive real-time American Sign Language (ASL) finger spelling system. The hand is segmented using the depth and color streams and Gabor features are extracted. A random forest based classification is used to distinguish letters of the alphabets with an accuracy of 75%. The authors address the ambiguity between certain letters by integrating English dictionary that allow user to choose between plausible letters. Lang et al. [42] have presented a framework for isolated sign recognition using NITE skeletal tracking. The framework has been used on 25 signs of German Sign Language (GSL) using HMM with an accuracy of 97% using depth-camera specific features. Li et al. [43] have developed a Kinect based hand-poses detection system that is able to identify fingertips. The system has been tested for gesture recognition to recognize 9 sign gestures that have been executed using single and double hands with accuracies of 84% and 90%, respectively. In [44], the authors have developed an open source framework for static and dynamic gesture recognition.

Dynamic gesture recognition is based on the angular feature of trajectories present in the depth stream. The framework has been tested on 16 geometrical uni-stroke dynamic hand gestures with an accuracy of 70%, whereas an accuracy of 90% has been recorded for 24 static gestures. Almeida et al. [4] have developed a Brazilian Sign Language (BSL) recognition system using Kinect. The authors have used seven vision based features computed using shape, movement and position of hands. The system has been tested on 34 BSL sign input and an accuracy of 80% was recorded with SVM based classification. Mehrotra et al. [45] have proposed a SLR system based on 3D skeleton points captured using Kinect. The authors have utilized upper body skeleton joints for extracting angular and distance based features. Using this, they have captured 37 non-continuous sign gesture of ISL and the recognition was performed on SVM classifier with an accuracy of 86.16%. In [46], the authors have proposed a feature learning approach for RGB-D inputs using sparse auto-encoder (SAE) with a Convolutional Neural Network (CNN). The learned features from RGB and depth channels were concatenated and fed into multilayer PCA for final feature. The approach was tested on 24 ASL alphabets where an accuracy of 99.05% was reported when tested with SVM classifier. Wang et al. [47] have proposed a SLR system using RGB-D data captured using Kinect sensor. Hand motions and hand postures were represented using 3D hand and HOG features, respectively. The recognition was performed using template matching where both hand postures and 3D hand trajectories were matched against the gallery of sign templates.

3.4. Leap motion based gesture recognition

Several researchers have used Leap motion to capture sign input for developing SLR, gaming, robotics or other HCI systems. Unlike Kinect, the device is used to track hand and finger joints only and easy to use because of its small size. Potter et al. [48] have presented a preliminary work on SLR system using Leap motion. The authors have used basic sign inputs of Australian Sign Language (AuSL) to analyze the performance. ANN has been trained for recognition purpose. However, their system is not able to recognize complex sign inputs. A 26 alphabet based sign input recognition system for ASL is presented in [6]. The authors have used SVM and k-NN based classification for recognition of sign inputs. The device is used for Arabic Sign Language (ArSL) recognition in [49] where 28 Arabic alphabet signs were recognized using Naive Bayes Classifier and Multilayer Perceptron Neural Network (MLP-NN). The device has also been used for 3D text recognition. The authors of [50] have presented a character and word recognition system using Leap motion. The inputs are treated as a time series data of 3D positions for DTW algorithm to recognize characters in real time. In [51], the device has been used for Chinese character recognition. The authors have used directional features in combination with 3D motion trajectories for recognition purpose. A Leap motion based signature authentication system is proposed in [52]. The recognition of signature is performed by computing features based on Histogram of Oriented Optical Flow (HOOF) and Histogram of Oriented Trajectory. A 3D text segmentation and recognition system has been proposed in [53] using Leap motion. The segmentation of text is done by finding large gap between connected components of a continuous stroke. They have used HMM based sequential classification techniques for recognition of segmented words.

3.5. Multimodal approaches

In order to capture sign inputs for SLR system, some researchers have proposed hybrid frameworks/models by combining inputs of more than one devices/sensors. In [18], a hybrid architecture of SLR has been proposed using Leap motion and two digital cameras. The Leap motion is configured to capture finger movements while the digital cameras are used for capturing body movements and facial expressions. The system has achieved an accuracy of 95% when tested with 20 signs of ArSL. A combined approach using Leap motion and Kinect device [19] has been used to capture ASL alphabet sign input. They have computed features based on the position and orientation of fingertips that are directly fed into SVM classifier for the recognition of the performed gestures. However, their dataset consists of static gestures and the recognition scheme specifically targets Leap motion data. A real time recognition system of ASL digits has been proposed in [20] using a multisensory approach. They have used two Leap motion sensors for capturing sign input by keeping the devices at different angles and the combined data of both the sensors is fed directly into HMM for recognition. Another multisensor approach used in [12] for SLR using two Leap motion sensors. The authors have used a trained SVM model for pose estimation which uses a subset of feature vector of each sensor. In [54], the authors employ two Kinect devices for static hand gesture recognition system. The algorithm is used to assist arthritis patients who face difficulty in using standard keyboard and mouse. Two Kinect devices have been used to provide 3D point cloud to indicate the gesture inputs. A k-NN classifier with combination of a majority voting scheme has been applied on multiple descriptors to recognize hand gestures. However, the approach do not combine the two sensors data. A multi-camera system has been proposed in [2] for face and hand gesture recognition using four camera setup. Two of the cameras are used to record stereo data for estimation of hand and face positions, while the other two cameras are used for detection and tracking. Four directional features have been used for face and hand gesture recognition using Linear Discriminant Analysis (LDA).

As per our understanding, majority of the existing approaches for SLR are designed either for static sign inputs [19,55] or their performance is not satisfactory on dynamic sign inputs [32,33]. We have noted that, single sensor does not capture hand gesture properly throughout the process. In this paper, we have proposed a multimodal framework for SLR system for dynamic gestures using Leap motion and Kinect sensors.

4. Proposed system

In this section, we present details of the proposed multimodal framework for capturing the sign input data and their recognition. A single sensor fails to deliver good view of the sign when it is performed on a plane parallel to the optical axis of the camera or sensor. To overcome this problem, we propose a multimodal approach using Leap motion and Kinect devices by keeping them at different views. Two devices complement each other while capturing the data. In our framework, if one device fails to acquire a sign input properly, the other sensor can acquire the data, thus recognition performance improves significantly. Experimental setup of our framework is shown in Fig. 2 and our framework in presented in Fig. 3. Fig. 2 shows the positions of two sensors for capturing inputs and corresponding display interface showing the outputs. Leap motion controller is kept below the user's arm facing upward while the Kinect having large field of view, is kept in front of the user.

4.1. Data acquisition

Sign gestures are simultaneously recorded by both devices. However, it is not possible to synchronize the frame rate because of hardware limitations. The frame rate of Leap motion controller is around 100 fps, whereas it is 30 fps for Kinect device. To deal

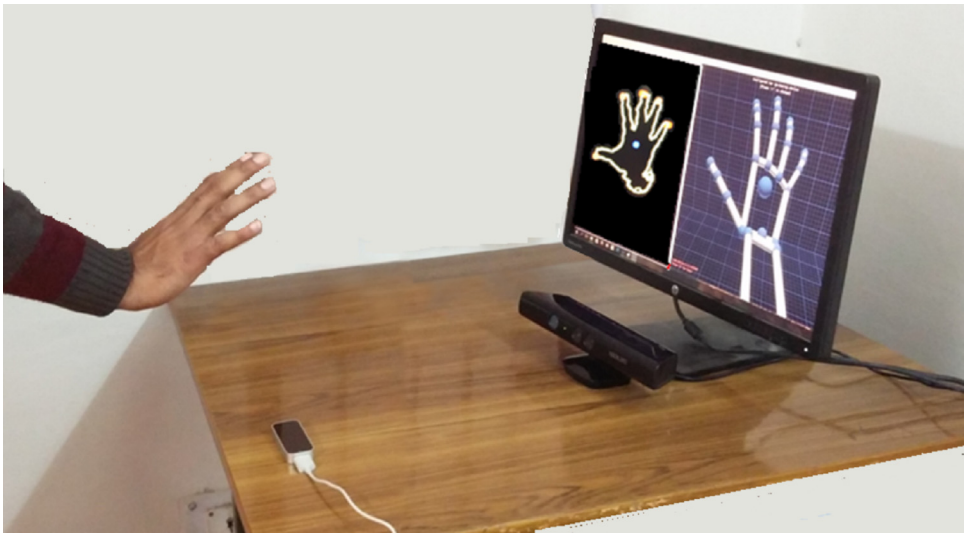


Fig. 2. The proposed multimodal setup used in our sign language recognition framework.

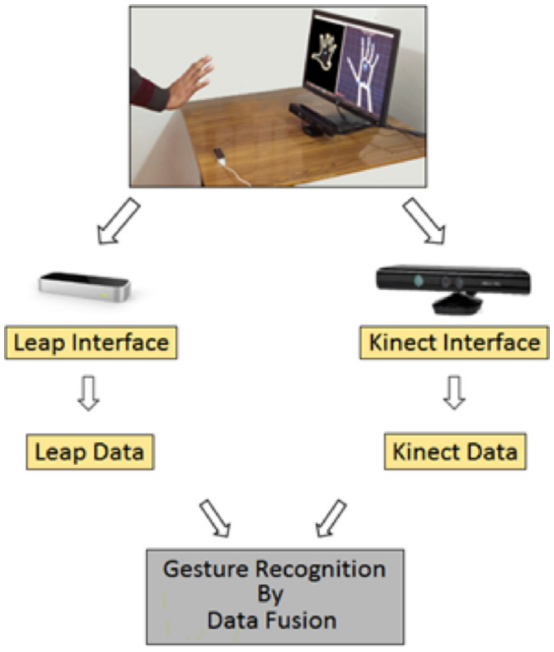


Fig. 3. Block diagram of the proposed system. Data is captured using Leap motion and Kinect independently and fused for recognition purpose.

with this, we have used a heuristic based preprocessing to synchronize the processing frame rate.

For Kinect, we have used an open source Candescent NUI library [7]. This library provides an easy interface for 3D hand and finger movement along with corresponding fingerIDs provided fingers appear in the field-view of the Kinect sensor. The fingerID information will not be available if the fingers are self-occluded and do not appear during gestures. To overcome this problem the 3D fingertips of missing/occluded fingers are approximated using palm center in our framework. This assumption works well because tips of the hidden fingers during a sign usually reach near to the palm center. Hence, during gesture, fingertips of five fingers of a single hand are always obtained and these information are used for gesture recognition. Data acquisition using Leap motion controller is done through Application Programming Interface (API) of the device. It provides a 3D point cloud of all fingertips. We have used the Leap motion controller API to extract fingertip positions and palm center in 3D. The acquired data is then preprocessed and relevant features are extracted.

4.2. Preprocessing

In our multimodal SLR framework, both the sensors simultaneously capture sign inputs. Since the recording speed of the devices vary, researchers have used correlation [58], re-sampling [59] and software synchronization [57] based techniques for the synchronization of two unequal temporal sequences. In this work, a re-sampling technique based on linear interpolation has been used for synchronization of frames. Sign data captured by both sensors may have different scaling factors. Thus, a z-score based normalization is used to normalize the coordinate system to $[-1,1]$. An example of preprocessing applied on a single-handed sign gesture '*college*' captured using Kinect, is shown in Fig. 4. Fig. 4(a) shows the 3D plot of the raw data corresponding to the word, while Fig. 4(b) shows 3D plot after re-sampling and normalization.

4.3. Feature extraction

We have extracted 11 dynamic features i.e. $(f_1, \ldots, f_5, \ldots, f_{11})$ for single-handed sign words and 22 dynamic features e.g. $(f_1, \ldots, f_{11}, \ldots, f_{22})$ for double-handed signs. These features are extracted independently from data acquired using each device. The details of the features are described below.

4.3.1. Fingertip position

The Leap motion API provides a point cloud of 3D position vectors of each finger, whereas the Kinect API provides access to 3D position of only those fingers that appeared during sign gestures.

Fingertip positions represent 3D point vectors. For a single-handed gesture, we have extracted five features (f_1 to f_5) as shown in Fig. 5(a) where each point in a feature consists of $< x_i, y_i, z_i >$ vectors in 3D space. We also consider position of the palm center, e.g. $< c_x, c_y, c_z >$ as feature f_6. The position of palm center is used to approximate the location of fingertips that are not extended during a sign in Kinect. An example is shown in Fig. 5(b) where fingertip positions of two fingers, which are not extended for gesture, are approximated with palm center position.

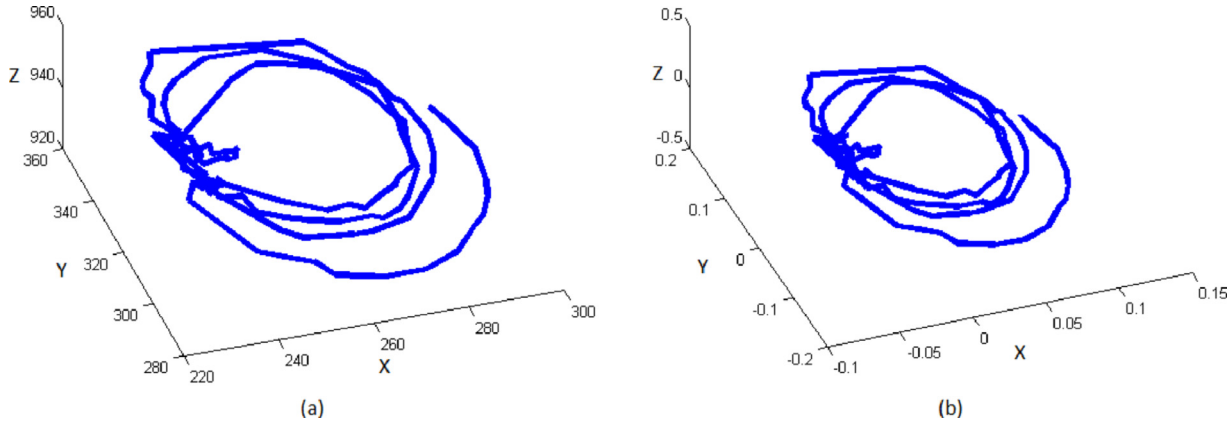


Fig. 4. Preprocessing of sign gesture 'college' (single handed) (a) 3D plot of raw data (b) 3D plot after re-sampling and normalization.

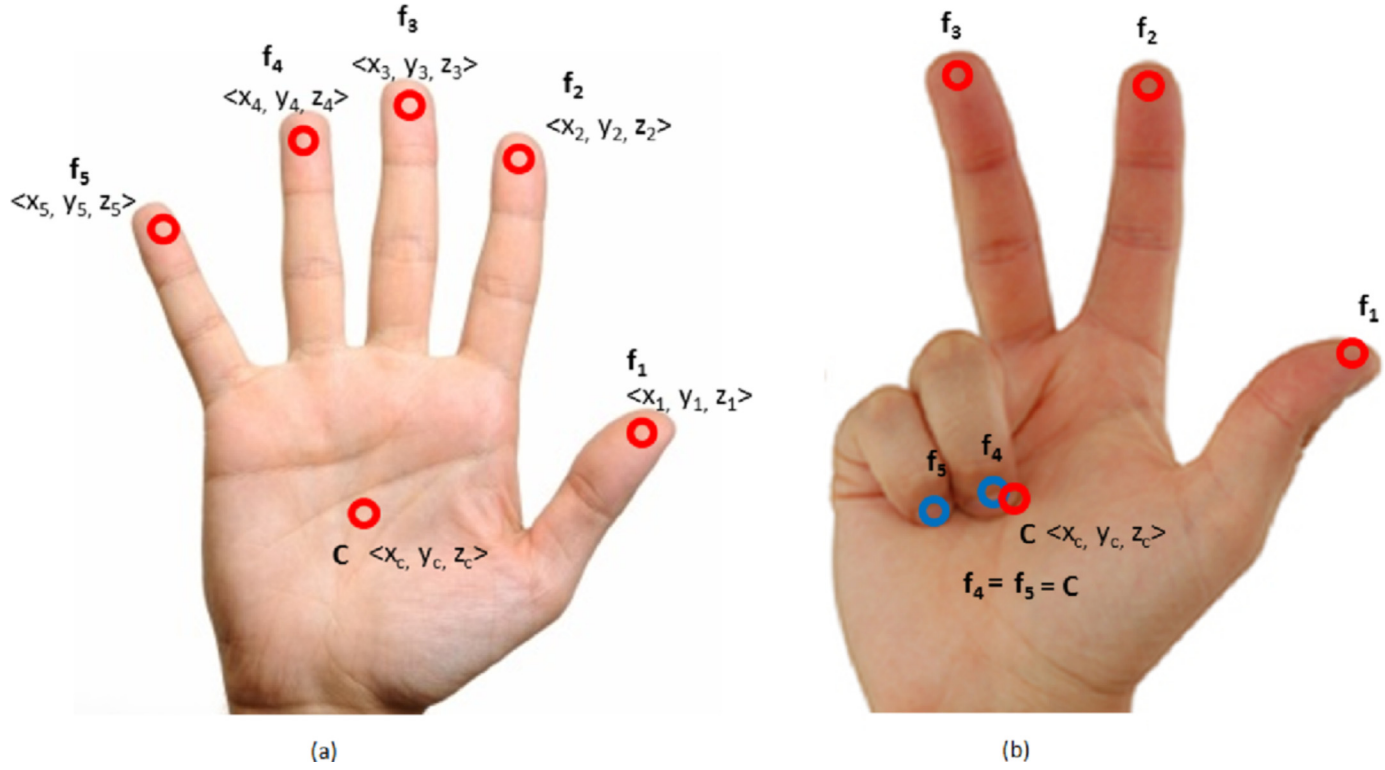


Fig. 5. Fingertip points: (a) 3D Fingertip and palm center coordinates and (b) Fingertip coordinates approximation using palm center.

4.3.2. Fingertip direction

Fingertip direction can be important in SLR systems as shown by various researchers [50,51,56,57]. In this study, we compute the fingertip direction in 3D with respect to the hand orientation. Direction of a 3D point $P(x, y, z)$ can be estimated by two neighbor points of a sequence, i.e. $P_1(x_1, y_1, z_1)$ and $P_2(x_2, y_2, z_2)$. Fig. 6 shows the fingertip movement from right to left to perform a gesture and it forms a vector $\overrightarrow{P_1P_2}$ and makes α, β and γ angles with the coordinate axes. These angles can be calculated using (1) and (2). We have computed five 3D direction features denoted by f_7 to f_{11}.

$$\overrightarrow{v} = \overrightarrow{OR} = < v_x, v_y, v_z >, |\overrightarrow{v}| = \sqrt{v_x^2 + v_y^2 + v_z^2} \quad (1)$$

$$cos(\alpha) = \frac{v_x}{|\overrightarrow{v}|}, cos(\beta) = \frac{v_y}{|\overrightarrow{v}|}, cos(\gamma) = \frac{v_z}{|\overrightarrow{v}|} \quad (2)$$

Same set of features is extracted for both hands when double-handed gestures are considered and the feature vector dimension is extended to 22.

4.4. Gesture classification

In this section, we present the details of HMM and BLSTM classifiers used in sign gestures recognition. The classifiers are well studied by various researchers in the context of gesture recognition.

4.4.1. HMM based sign language recognition

HMM has been successfully used by researchers to model the temporally correlated data streams in speech and handwriting recognition systems [53]. The model is a discrete state-space stochastic model which can be effectively used for complex hand gesture recognition such as SLR systems [3,23,44]. The model represents a statistical behavior of the observed sequence. An HMM

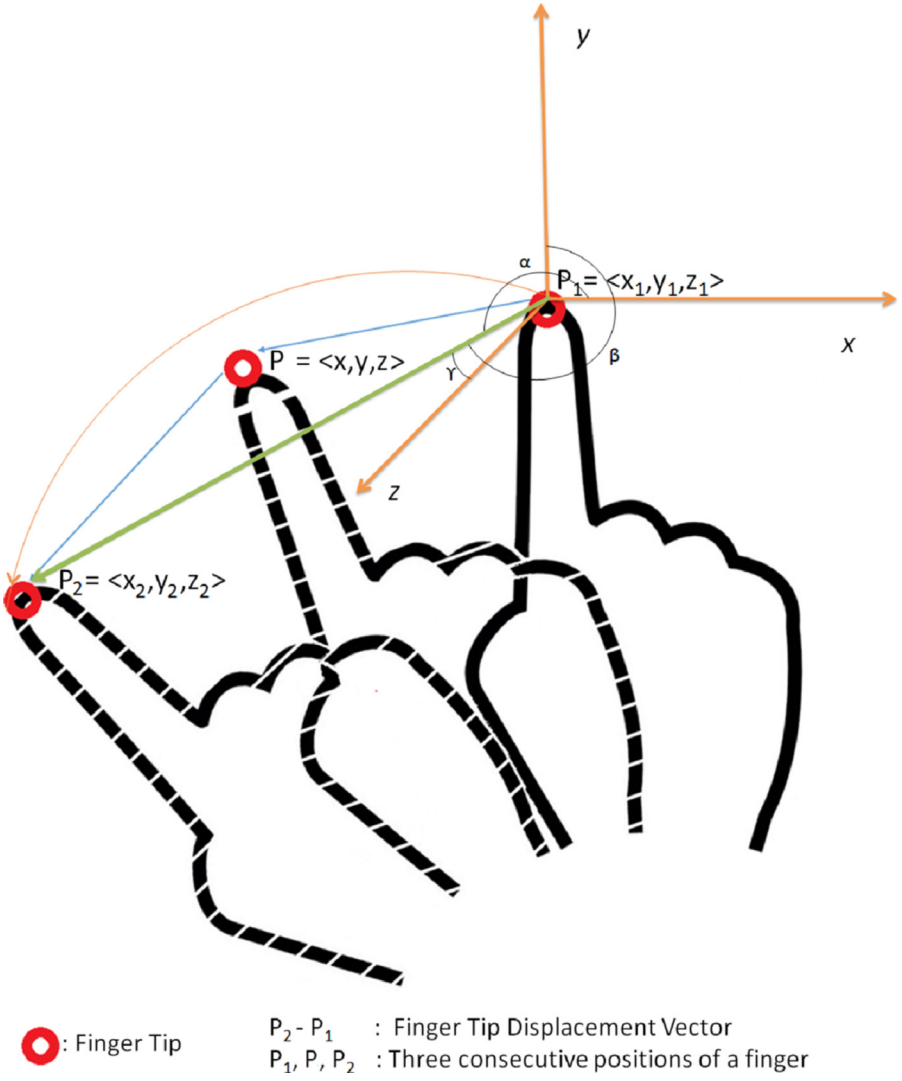


Fig. 6. Fingertip direction computation for a sign gesture using two neighbor points from the gesture sequence.

can be defined by initial state probabilities π, state transition matrix $A = [a_{ij}]$, $i, j = 1, 2, \ldots, N$, where the transition probability is denoted by a_{ij} from state i to state j and the observation probability $b_j(O_k)$ modeled with the continuous output probability density function. The density function is denoted by $b_j(x)$, where x represents the k dimensional feature vector. A Gaussian Mixture Model (GMM) is defined separately for each state of the model. Formally, the output probability density of state j can be defined using (3),

$$b_j(x) = \sum_{k=1}^{M_j} c_{jk} \aleph(x, \mu_{jk}, \Sigma_{jk}) \tag{3}$$

where M_j is the number of Gaussian assigned to j and $\aleph(x, \mu, \Sigma)$ denotes a Gaussian with mean μ and co-variance matrix Σ, and c_{jk} is the weight coefficient of the Gaussian with component k of state j. For a model λ, if O is an observation sequence, e.g. $O = (O_1, O_2, \ldots, O_T)$ and is assumed to be generated by a state sequence $Q = Q_1, Q_2, \ldots, Q_T$ of length T, we can calculate the probability of observation or likelihood using (4), where π_{q_1} denotes the initial probability of the start state.

$$P(O, Q|\lambda) = \sum_{Q} \pi_{q_1} b_{q_1}(O_1) \prod_{T} a_{q_{T-1} q_T} b_{q_T}(O_T) \tag{4}$$

In the training phase, the original sequence of the sign input and their corresponding feature vector sequences are fed together to train each sequence model. Re-estimation of the initial output probability distributions of $b_i(O)$ is done using Baum–Welch algorithm until the likelihood of the training set is maximized. The recognition is performed through Viterbi decoding algorithm that finds the signed sequence having the best likelihood using a given feature vector sequence.

4.4.2. BLSTM-NN based sign language recognition

Neural network based approaches can also be successfully used for modeling temporal data in hand gesture recognition systems [49,55]. Recently, it has been shown that, BLSTM-NN [60] performs better for speech and handwriting recognition applications. Since HMM based recognition system assumes the probability of each observation on the current state, it makes the prediction difficult while modeling contextual information. Moreover, BLSTM-NN process the input sequence in both directions [61] because of its two hidden layers. One layer processes the input sequence in forward direction and the other layer processes the input in backward direction. Both hidden layers are connected to the same output layer, thus providing access to both direction (i.e. forward and backward) of every point of the input sequence. In this work, we are using BLSTM-NN based sequence classification

Algorithm 1. Classifier combination algorithm.

1: **procedure** MAX PROBABILITY
2: Classifier (C_i) from C=$C_1, C_2 \ldots\ldots C_n$ assigns a probability (P_i) to a sequence 'x'
s.t. $P_i(x) \geqslant 0$ and x ϵL_k where L=$L_1, L_2 \ldots\ldots L_m$ number of classes.
3: $P_{max} = argmax_{i=1}^{n}(P_i)$ //(P_{max}) is the maximum probability
4: $L_j \ \epsilon L$ corresponding to classifier C_i selected by P_{max} will assign to 'x'.
5: **end procedure**

approach for recognition of sign input captured by Leap motion and Kinect device. For this, we have trained the network by using Cross Entropy Error based objective function. The output layer has the softmax function ensures that the network outputs are normalized between 0 and 1 which sums to 1 within each time step as a standard for 1 of K classification [61]. There are K output units, one for each class of the gesture sequence.

The cross entropy error (E) for K classes can be computed using (5).

$$E = -\sum_{(x,z)\epsilon S}\sum_{k=1}^{K} z_k \ln y_k \tag{5}$$

where x is the input sequence and z is the target sequence which collectively form the input pair (x, z) from the training set S and y is the probability defined in (6) such that the input belongs to a particular class.

$$p(z|x) = \prod_{k=1}^{K} y_k^{z_k} \tag{6}$$

4.5. Gesture recognition using multimodality

As we discussed earlier, we have used combination of Leap motion feature and Kinect feature to improve the recognition performance. Two different strategies; feature combination and classifier combination are used in our framework. These are discussed as follows:

4.5.1. Feature combination

In feature combination, the feature vectors of Leap motion (f_L) and Kinect (f_K) are combined. It results into a new feature vector f_T and is defined using (7),

$$f_T = f_L + f_K \tag{7}$$

where f_L and f_K represent feature vectors for Leap and Kinect, respectively. Hence f_T results into a feature vector of 22 dimensions for single hand sign gestures. The newly generated f_T is then used for sign gesture recognition. The fused feature f_T is better than individual feature vectors f_L or f_K. This has been verified through experiments.

4.5.2. Classifier combination

In classifier combination module, each classifier C_i, $i \epsilon (1, 2, \ldots, n)$ is trained with the given feature set F_j $(j = 1, 2, , \ldots, z)$. The recognition of a unknown sign gesture is decided using a probabilistic framework as described in Algorithm 1. A combined probability score using max-rule is calculated for a given test sequence 'x' and assigns a class L_k, $k \epsilon (1, 2, \ldots, m)$ to which the sequence belongs. The assigned class (L_k) is decided by selecting the maximum probability (P_{max}) among P_i assigned by C_i.

Table 2
Subset of words selected for sign gestures.

S. no.	Single hand gesture	Double hand gesture
1	Child	Born
2	Cool	Close
3	Daily	Enjoy
4	Day	House
5	Good	Open
6	Office	Please
7	Morning	Shop
8	No	Where
9	See	Which
10	Stand	Want

5. Experiment results and discussion

In this section, we describe the dataset and present results of experiments. The performance of the proposed framework has been evaluated on the collected dataset in two phases. First, we computed the recognition performance using independent sensor. In the next phase, we evaluate our proposed multimodal framework.

5.1. Dataset collection

We have collected a large dataset containing sign gestures of ISL. These are captured simultaneously by Leap motion and Kinect sensors. The dataset consists of 50 dynamic sign gestures that are performed by 10 different signers, 8 of the signers were male and rest were females candidates. The signers are the students of '*Anushruti*' (an intermediate school for hearing impaired people at Indian Institute of Technology, Roorkee, India). Every sign input was repeated 15 times by each signer which makes 750 different gestures associated to a single signer, thus a total of 7500 word samples have been recorded. The dataset consists of both single and double hand dynamic sign gestures. Out of these 50 dynamic sign gestures, 28 words were performed by single hand (right hand only) and rest of the signs (i.e. 22) were performed using both hands. All collected gestures in the dataset were performed in the viewing field of both the sensors, hence gestures that leave the viewing field of either sensor e.g. head, shoulder, ear and nose touching were not considered in this work. The dataset is also made available online[1] for further research in this direction. Descriptions of some of the gestures, are shown in Table 2.

In our dataset, we have included basic sign gestures that are mostly used in daily life activities. These gestures were selected based on the discussion with hearing-impaired students. From the discussion, we noted that there also exist some gestures which are performed to construct certain words such as "*cool*", "*wow*",

[1] https://sites.google.com/site/iitrcsepradeep7/.

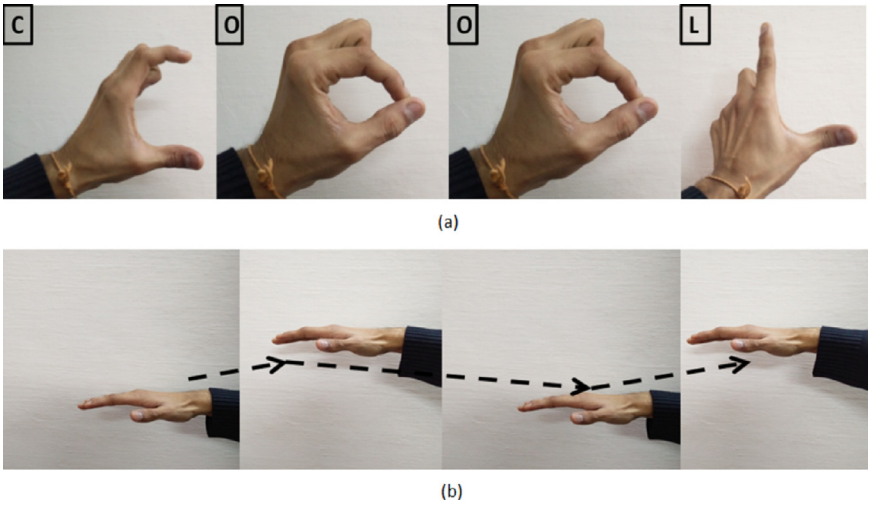


Fig. 7. Sign gesture representation (a) for the word 'cool' using finger spelling (b) for the word 'child' using sign language.

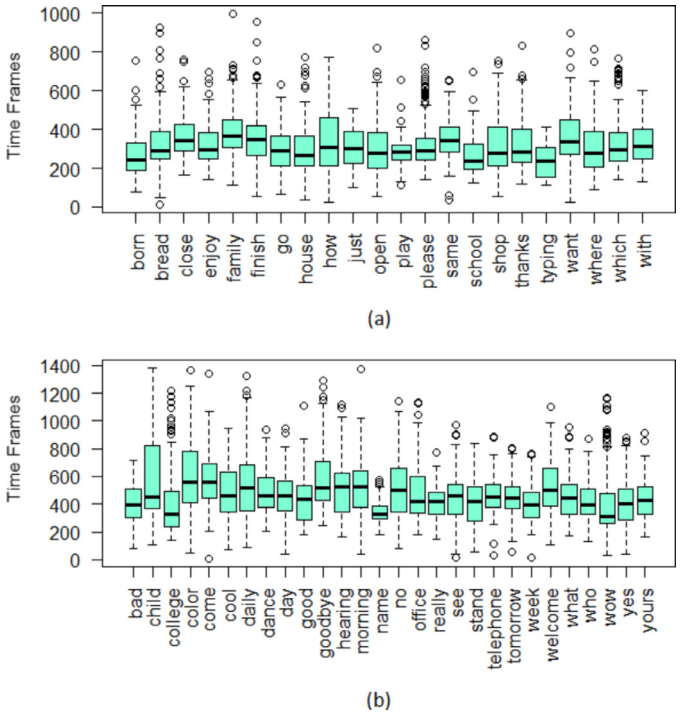


Fig. 8. Box Plots of sign gestures captured using Leap motion: (a) Single hand sign gestures and (b) Double hand sign gestures.

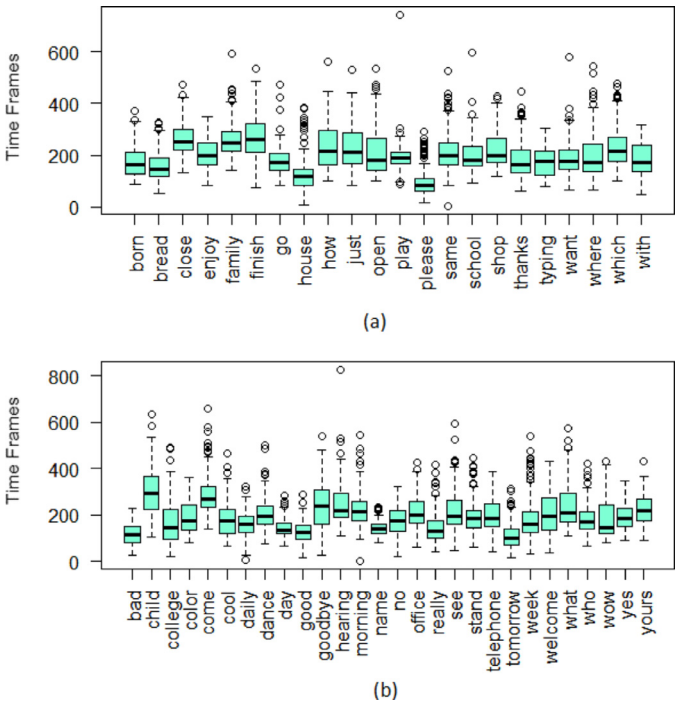


Fig. 9. Box Plots of sign gestures captured using Kinect: (a) Single hand sign gestures and (b) Double hand sign gestures.

etc. using finger-spelling. These finger-spelled words are also used frequently in their daily-life activities. In our dataset, such finger-spelled sign gestures are included with the rest of the ISL gestures. Fig. 7(a) shows the sign representation of the word "*cool*", where the word is represented using English alphabets. Fig. 7(b) shows the sign representation of the word "*child*". Here, the word "*child*" is expressed using gesture of hand.

Large variation in data is measured using a statistical analysis technique known as box-plot. It uses median, approximate quartiles, and lowest / highest data points to convey the level, spread, and symmetry of distribution of the data. We have shown example box-plots for single and double hand gestures, in Fig. 8(a) and (b), respectively. These gestures were captured using Leap motion sensor. Similarly, box-plots of the sign gestures captured using Kinect, are shown in Fig. 9(a) and (b) for single and double hand signs, respectively.

5.2. Details and visualization of signed gesture

Since a sign gesture sequence is performed in air, thus, can be visualized using 3D plot that shows the variation within the same sign when performed by different users. Single-handed signs representing the word "*child*" captured by Kinect and Leap motion, are shown in Fig. 10(a) and (b), respectively. The figure has been obtained by plotting the palm center points during the whole sequence. Each color in the plot is used to discriminate a signer from others which shows the variation in the input sequences. Another example of both-hand based sign gesture "*close*" is shown in Fig. 11. The sign was completely sensed by Kinect, hence it shows

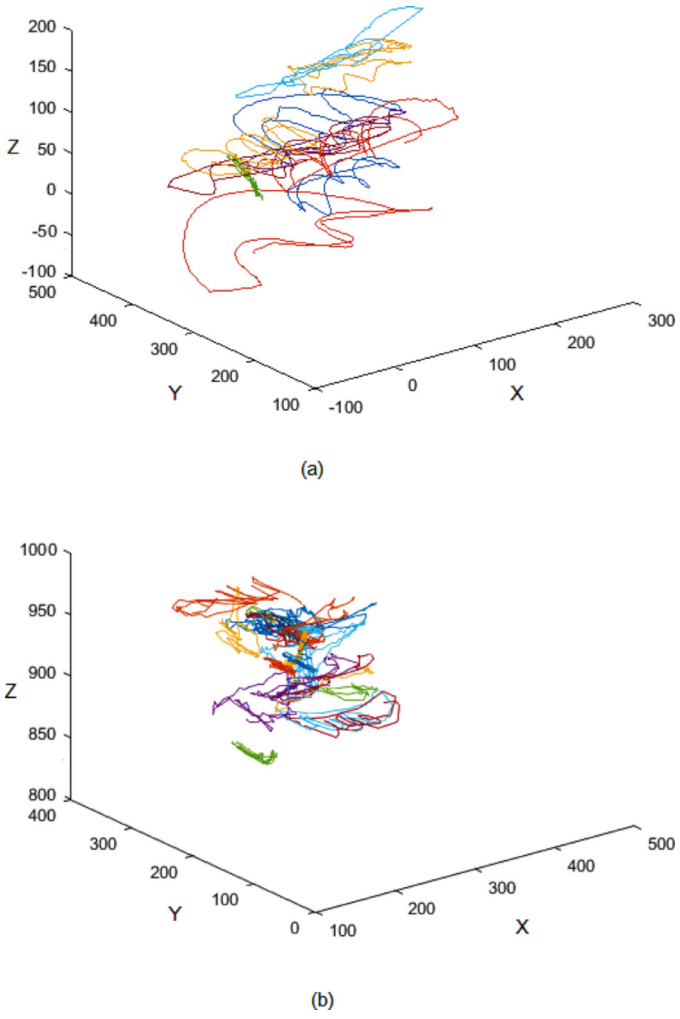


Fig. 10. A single hand based 3D sign representation for the word "*child*" by different signers (a) captured by Leap motion and (b) Captured by Kinect.

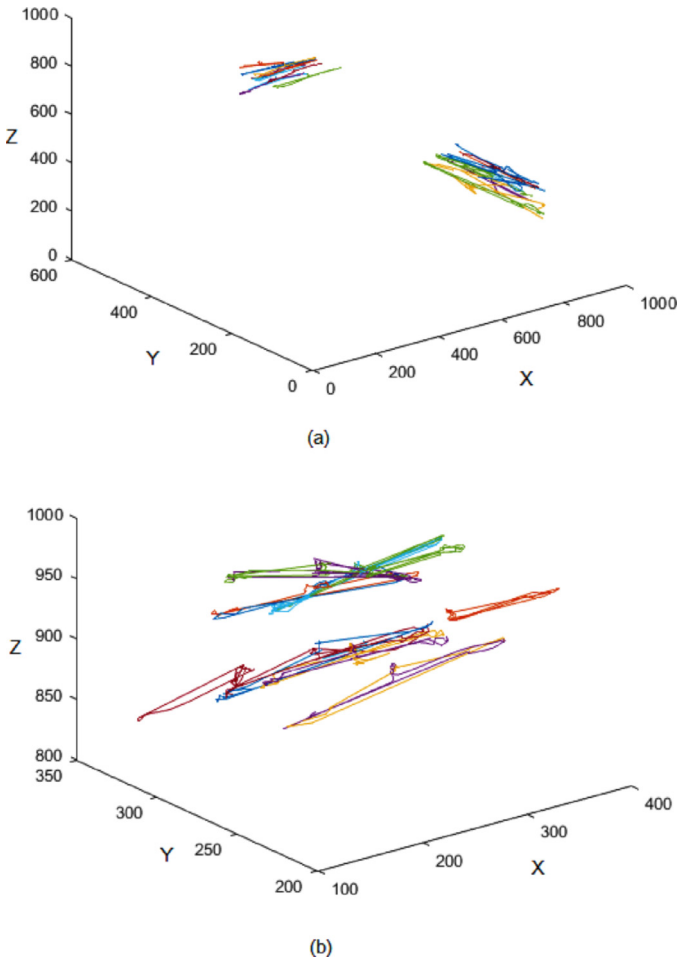


Fig. 11. 3D sign representation for the word "*close*" by double hand by different signers (a) captured by Leap motion and (b) Captured by Kinect.

variations in the plot, while Leap motion was not able to sense the sign completely, thus shows very small variation in the plot.

5.3. Gesture recognition results

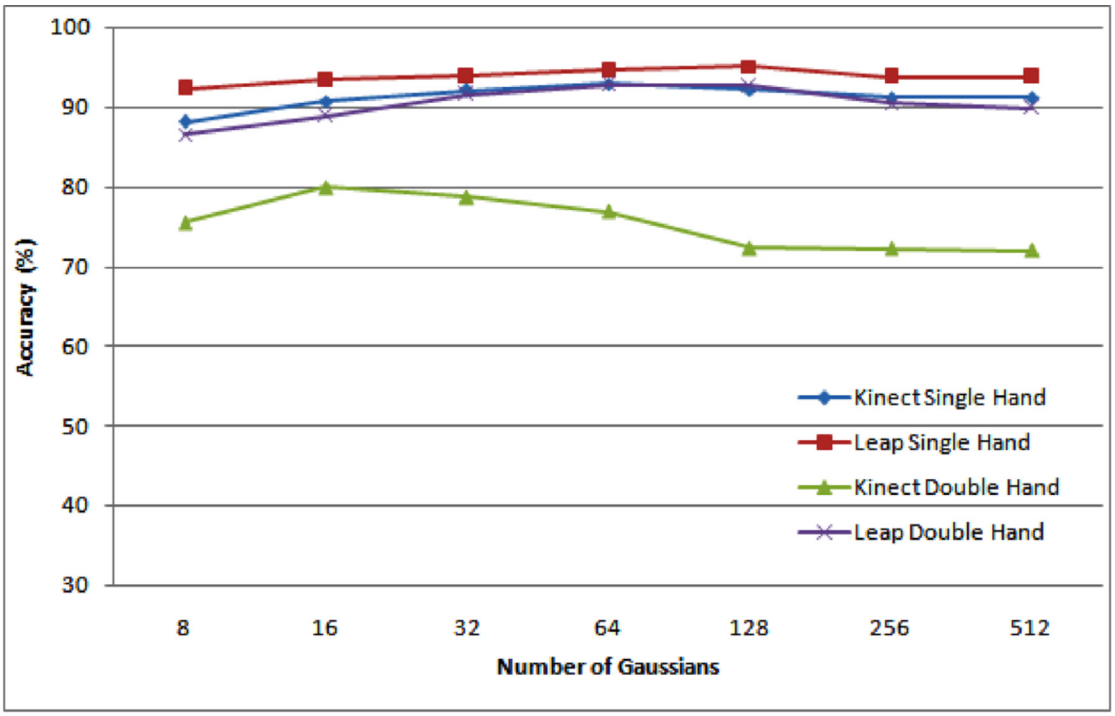
We have divided the collected dataset into five folds for both single and double hand gestures, out of which, four sets have been used in training and one in testing. The recognition process was carried out using two different classifiers on Leap motion and Kinect datasets, separately. Finally, a combined approach using the fused feature, has been adopted.

5.3.1. HMM guided gesture recognition

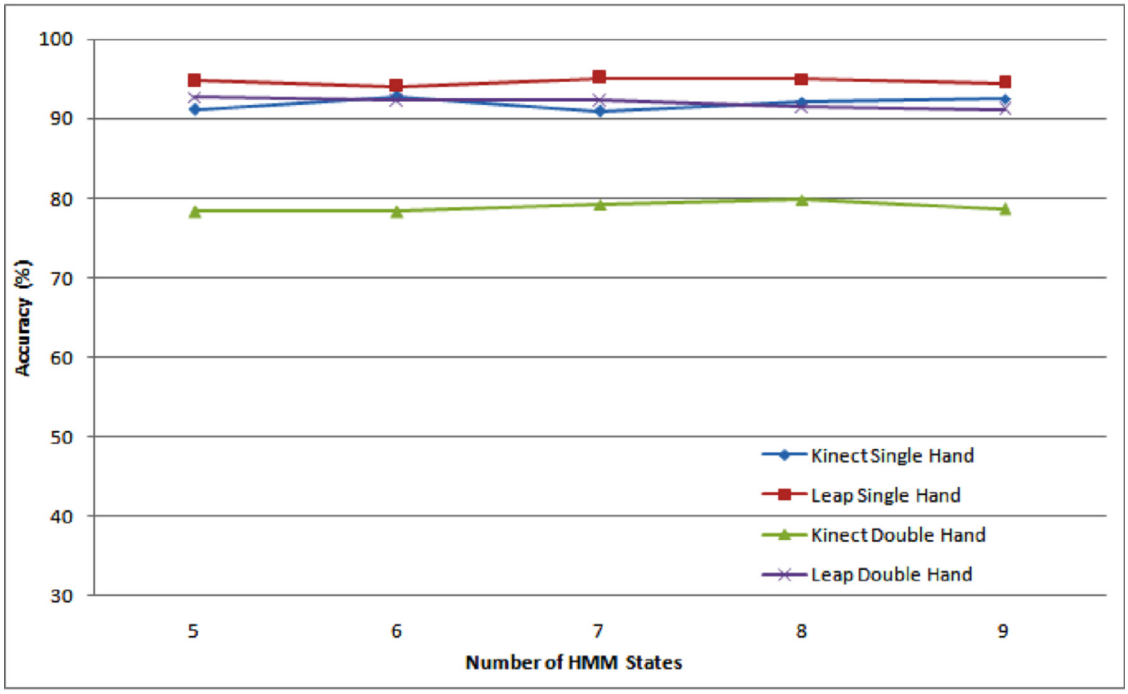
We have trained the HMM classifier through systematic tuning of the parameters. The experiments were performed by varying HMM states, $S_t \in \{5, 6, 7, 8, 9\}$ and by varying the number of mixture components per state i.e. from 1 to 512 with an incremental step of power of 2. The performance of the recognition is recorded on data acquired by Leap motion and Kinect, respectively. Single-hand sign gesture recognition performance can be seen in Fig. 12(a), where an overall accuracy of 95.12% has been recorded using Leap motion data with the help 128 Gaussian distributions, the accuracy is reduced to 92.86% with the help of 64 Gaussian distributions, when Kinect sensor data is used. Similarly for both hand gestures, Leap motion's accuracy with 64 GMMs has been recorded as 92.73% and Kinect sensor's accuracy with 16 GMMs has been recorded as 79.85%, respectively. The recognition is also carried out by varying HMM states as shown in Fig. 12(b). Overall accuracies of 95.12% (with 7 HMM states) and 92.86% (with 6 HMM states) have been recorded for single hand gestures on Leap and Kinect dataset, respectively. Similarly, an accuracies of 92.73% (with 5 HMM states) and 79.85% (with 8 HMM states) have been recorded for double hand gestures on Leap and Kinect dataset.

5.3.2. BLSTM-NN guided gesture recognition

The BLSTM-NN network was trained with initial weights chosen from flat random distribution range [-0.1,0.1], a learn-rate of $1e-4$ and a momentum of 0.9. The order of the training set was chosen randomly at the start of each training epoch and the weight updates were carried out at the end of each gesture sequence separately i.e. dataset belongs to Leap motion, Kinect and joint calibrated approach. In single hand based sign gestures, accuracies of 84.74% and 79.41% have been recorded for Kinect and Leap motion datasets, respectively. In two hand gestures, recognition rates of 72.3% and 81.82% have been recorded using Kinect and Leap motion datasets, respectively. Learning curves are shown in Fig. 13(a) and (b) representing single and double hand signed gestures, respectively. The learning curves show a decrease in training and validation error on our dataset. It may be noted that, the learning curve in Fig. 13(a) does not change after 68 training epochs, and there is very small change in validation network and classification error beyond this many training epochs. This has been marked as the best network in the figure. Similarly, the learning curve for two



(a)



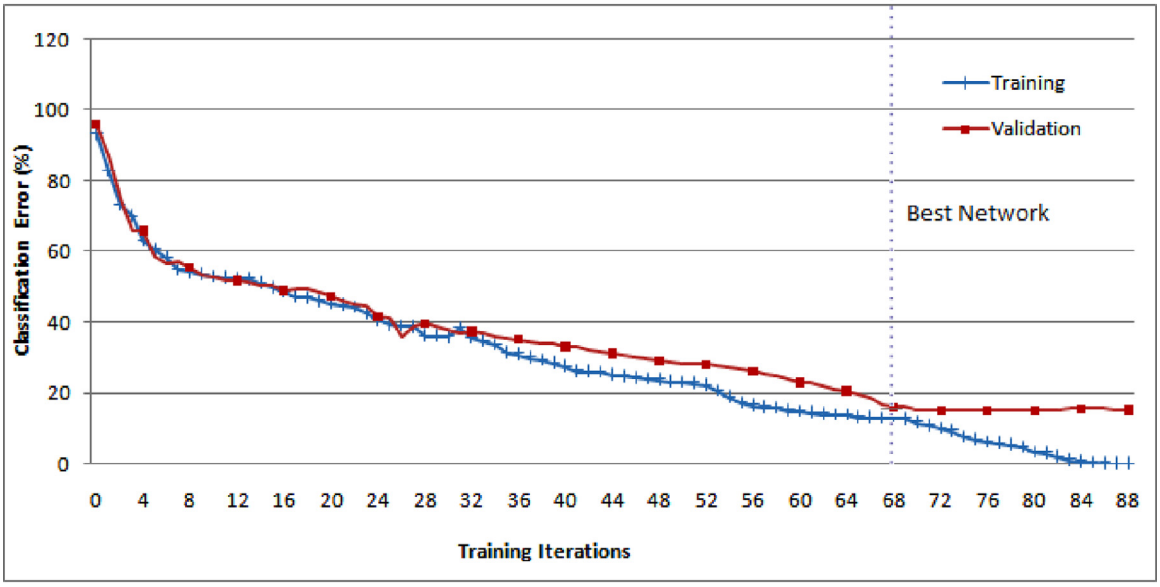
(b)

Fig. 12. HMM based sign gesture recognition: (a) Gaussian distribution per state and (b) State number.

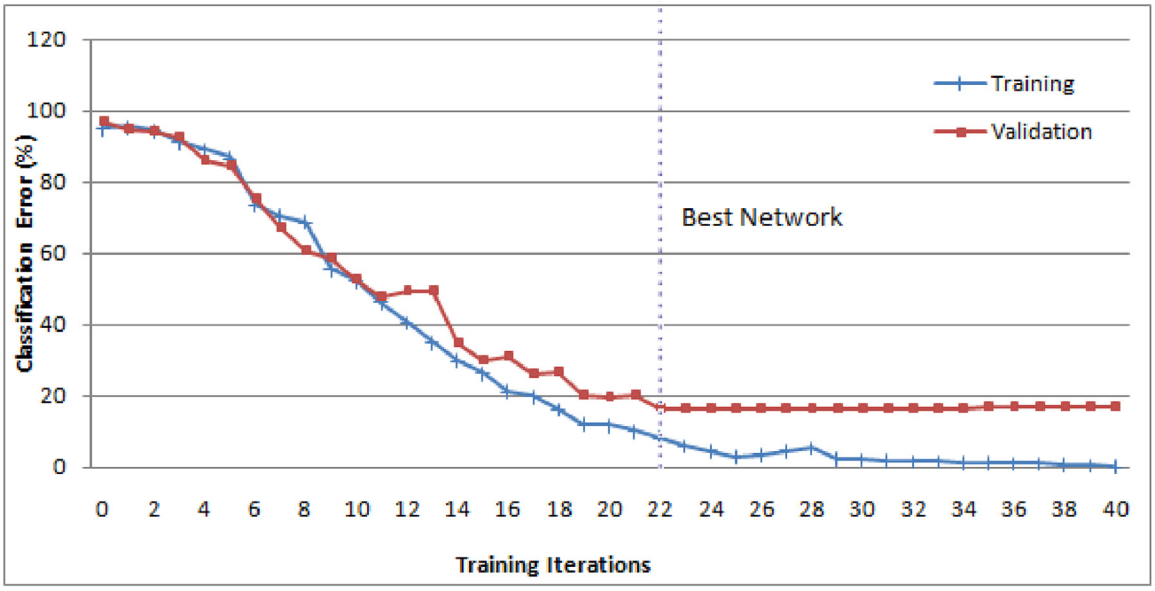
hand sign gestures is shown in Fig. 13(b). It can be observed that, classification error does not change after 22 training epochs, thus, it has been marked as the best network in the figure.

5.3.3. Gesture recognition using feature combination

Feature combination based recognition is done by combining the feature vector f_L and f_K extracted from Leap motion and Kinect, respectively as discussed in Section 4.5.1. We show in Table 3 some examples of the sign gestures which has been correctly recognized after feature combination step. For instance, the first two entries in Table 3 representing single and double hand gestures, were wrongly classified by HMM when tested with separate feature vector of Leap and Kinect, however, after feature combination, correct classification was obtained. It is because certain sign



(a)



(b)

Fig. 13. Variation of error in sequence classification with respect to training size: (a) for single hand gestures and (b) for double hand gestures.

gestures were partially captured by both devices. After feature combination the partially captured sign sequences were classified properly which proves the robustness of the multimodal framework for gesture recognition. The approach increases the overall accuracy of gesture recognition for single and double hand sign gestures as recorded in Table 4. Hence an overall gains of 3.0% and 5.91% were obtained on all gestures (combined single and double-handed gestures) with combined features when tested with HMM and BLSTM-NN classifiers, respectively. Fig. 14 shows confusion matrix of the recognition performance of all 50 sign gestures using HMM classifier with feature combination. However, a gains of 2.26% (0.91%) and 2.88% (1.67%) were obtained for single (double) hand gestures with the help of HMM and BLSTM-NN classifiers, respectively.

After combining f_L and f_K using (7), the dimension of the feature vector (f_T) becomes of 132 dimensions. In our experiment, we have performed a PCA [62] based dimension reduction approach on f_T to check the effectiveness of the reduced dimensionality of the feature vector. PCA is a popular dimensionality reduction technique and it has been used as a feature selection by researchers in gesture recognition [32,46,54]. The recognition results with different reduced features is shown in Fig. 15, where the best

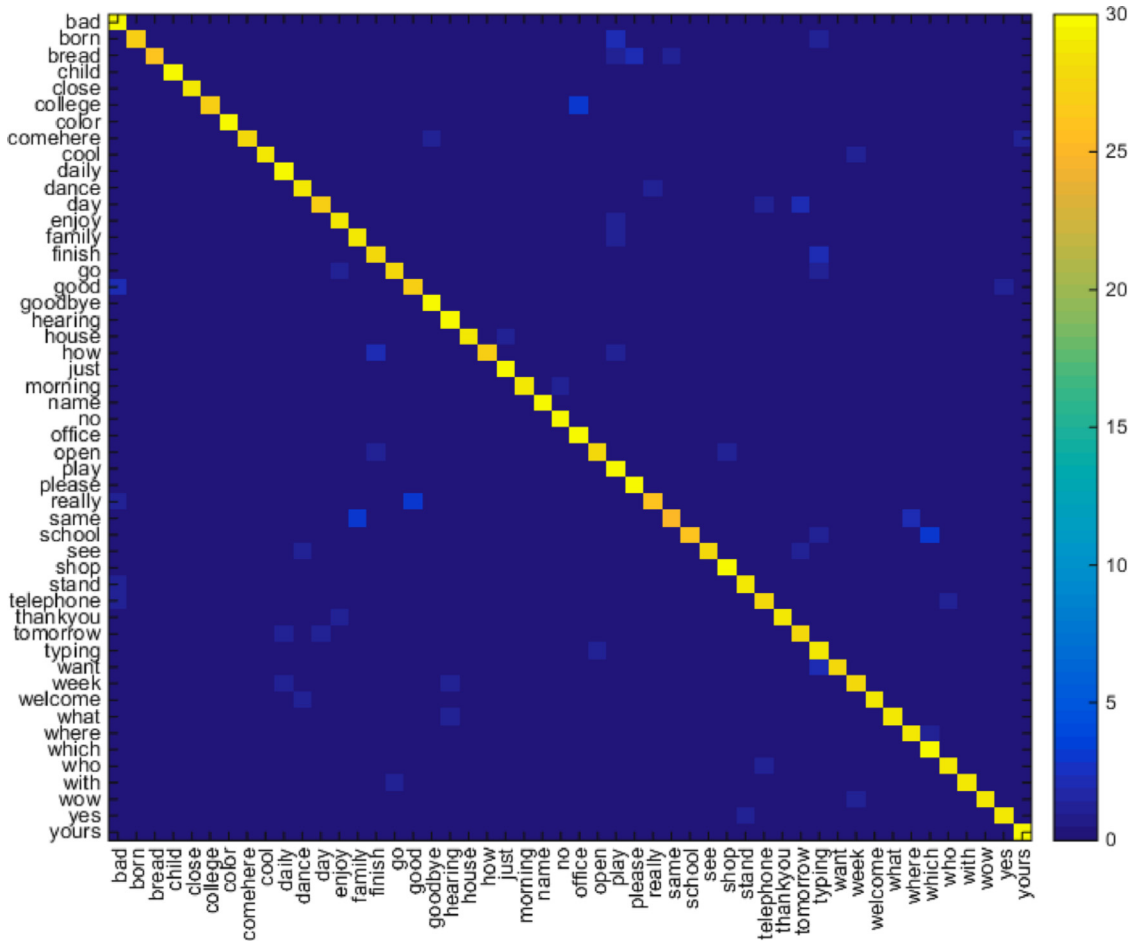


Fig. 14. Confusion matrix of all 50 sign gestures using feature combination.

Table 3
Feature combination for single/double hand sign gestures.

S. no.	Sign gesture	Leap	Kinect	Leap-kinect
Single handed sign gestures				
a.	Morning	× Who	× No	√ Morning
b.	Stand	× Telephone	× Bad	√ Stand
c.	Come	√ Come	× Telephone	√ Come
d.	Child	× Come	√ Child	√ Child
Double handed sign gestures.				
e.	Enjoy	× Play	× School	√ Enjoy
f.	How	× Family	× Finish	√ How
g.	Born	√ Born	× Play	√ Born
h.	Bread	× How	√ Bread	√ Bread

Table 4
Recognition accuracy of single/double hand gestures after feature combination.

S. no.	Classifier	Leap motion	Kinect	Leap motion+ kinect
Single handed sign gestures (%)				
a.	HMM	95.12	82.86	97.38
b.	BLSTM-NN	79.41	84.74	87.62
Double handed sign gestures				
c.	HMM	92.73	79.85	93.64
d.	BLSTM-NN	81.82	72.30	83.49
All sign gestures (Combined)				
e.	HMM	92.60	84.26	95.60
f.	BLSTM-NN	72.39	78.66	84.57

recognition rate of 95.56% has been recorded with top 60 principal components. Note that the best result obtained after PCA is less than without dimension reduction based approach.

5.3.4. Gesture recognition using classifier combination

In this section, we present the results obtained using classifier combination as discussed in 4.5.2. The approach is based on probability based framework in which the input was f_T (combined feature from Leap motion and Kinect devices), we have calculated the probability of a given sign sequence using HMM and BLSTM classifiers. The word with the highest probability is chosen as the resulting sign. A subset of the sign gestures recognized using classifier combination, is shown in Table 5. A recognition rate of 96.33% has been recorded on all gestures when tested with the probability based framework. Whereas accuracies of 97.85% and 94.55% have been recorded in single and double hand sign gestures, respectively. Hence a gain of 0.73% in recognition rate was recorded for all gestures in comparison to the feature combination based recognition process. Similarly, gains of 0.47% and 1.37% in recognition rate were recorded for single and double hand gestures, respectively.

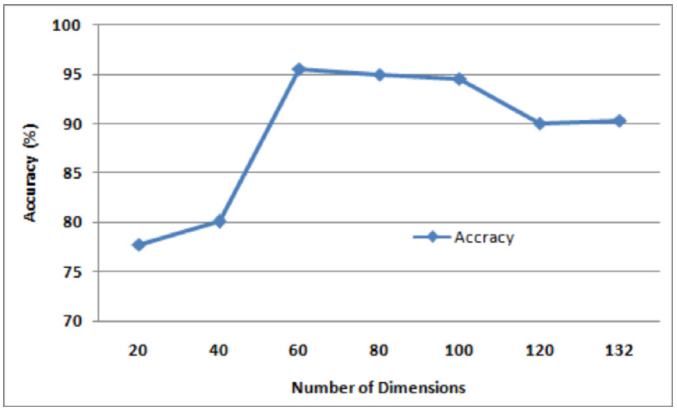


Fig. 15. PCA based recognition performance of all 50 sign gestures by varying Eigenvectors.

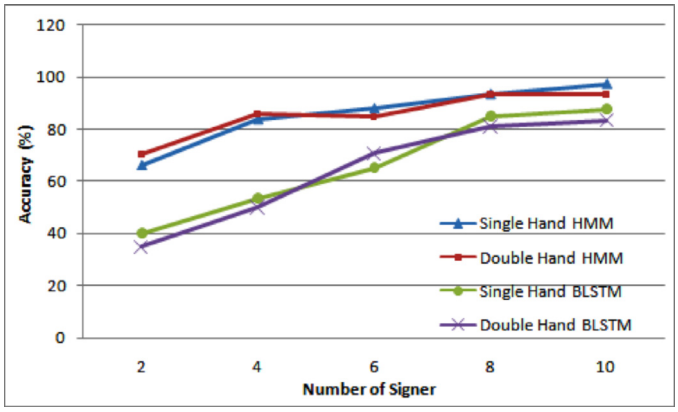


Fig. 16. Scalability test for single and double hand sign gestures.

Table 5
Classifier combination for single/double hand sign gestures based on probability.

S. no.	Sign gesture	HMM Probability	BLSTM Probability	Recognized Sign gesture
Double handed sign gestures				
a.	See	√ See	× Day	√ See
b.	Name	× Hearing	√ Name	√ Name
c.	Telephone	× Color	× Who	× Who
Double handed sign gestures				
d.	Family	√ Family	× Shop	√ Family
e.	Thanks	× Enjoy	√ Thanks	√ Thanks
f.	Same	× Family	× Finish	× Family

5.3.5. Oracle approach

We have manually calculated the recognition results of classifier combination by checking the ground truth of test data. It was observed that, in single hand gestures (840 samples), 22 gestures have been wrongly classified by HMM out of which 13 gestures have been correctly recognized by BLSTM-NN. Thus, an overall accuracy of 98.93% can be achieved. In double hand signs, our test set contained 660 gestures sequences. 42 gestures have been wrongly classified by HMM, out of which 15 gestures have been correctly recognized by BLSTM-NN. Thus, an accuracy of 95.91% can be achieved. A comparative analysis of the results obtained using our proposed framework against the oracle experiment can be found in Table 6.

5.4. Scalability test

Here, we show the scalability of our approach using training data from number of users. The system has been trained by varying number of signer's i.e. (2, 4, 6, 8, 10) and keeping the testing data fix for all experiments. The testing consists with 840 (single hand) and 660 (both hands) number of gestures, respectively. Results of scalability using HMM and BLSTM-NN is shown in Fig. 16. We have observed that, the accuracy of the system keeps on increasing as the number of users increases for both types of gestures.

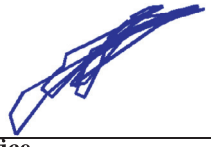
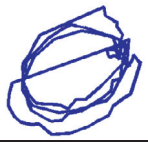
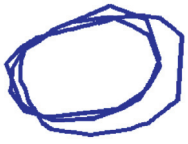
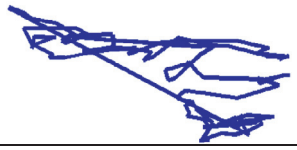
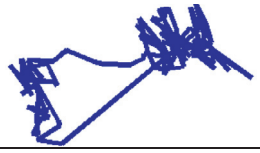
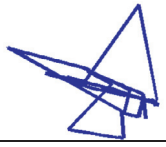
5.5. Error analysis

Finally, an error analysis is performed to analyze the drawback of our framework by consulting the confusion matrix shown in Fig. 14. We noted that some words share similar gesture characteristics in terms of number of fingers, finger movement, and orientation of the hands. Some of the similar looking words along with their error percentage are shown in Table 7. The error percentage column shows the share of sign gestures among total wrongly classified gestures. We have 18 and 36 wrongly classified sign gestures after classifier combination in both single and double handed gestures, respectively. First row of the table (e.g. row a) depicts the word "*come*" that is wrongly recognized as "*yours*". This happens because both words share five number of fingers and similar wrist movement. The difference between these gestures is as follows. In "*come*" gesture, the palm is oriented toward the signer, whereas during "*yours*" gesture, it is opposite. In (b), the sign gesture "*college*" is found to be closely matching with "*office*" because both gestures share a similar looking circular movement of the palm during the sign. The user performs a circular gesture for character '*c*' of "*college*", while the character '*o*' of "*office*" is also represented by a circular movement. Thus, the classifier gets confused during

Table 6
Performance comparison for single/double hand sign gestures.

S. no.	Approach	Accuracy (%)
Single handed sign gestures		
a.	Oracle approach	98.93
b.	HMM based feature combination and recognition	97.38
c.	BLSTM-NN based feature combination and recognition	87.62
d.	HMM + BLSTM-NN based classifier combination and recognition	97.85
Double handed sign gestures		
e.	Oracle approach	95.91
f.	HMM based feature combination and recognition	93.64
g.	BLSTM-NN based feature combination and recognition	83.49
h.	HMM + BLSTM-NN based classifier combination and recognition	94.55

Table 7
Error analysis for single/double hand sign gestures.

S.No.	SIGN GESTURE PLOT (3D)	SIGN GESTURE PLOT (3D)	Error (%)
SINGLE HAND SIGN GESTURES			
a.	come 	yours 	11.11
b.	college 	office 	22.22
c.	telephone 	who 	16.66
DOUBLE HAND SIGN GESTURES			
d.	bread 	house 	8.33
e.	with 	same 	5.55
f.	same 	family 	5.55

recognition. Similarly, the sign gesture for "*telephone*" got wrongly matched with the sign word "*who*". In case of double hand gestures, the word "*bread*" is confused with "*house*" because both gestures share same number of fingers (two) and similar finger movement and style. In (d), "*with*" has been found to be confused with "*same*" because both share similar hand movement as both hands becomes closer to each other. The difference between these two gestures is as follows. The word "*with*" uses both hands' thumb while the word "*same*" uses both hands' index finger. Therefore, both hands get closer to each other that creates a confusion during classification. Likewise, the sign gesture for the word "*same*" was wrongly classified as "*who*" during classification.

5.6. Comparative study

We have performed a comparative study with popular DTW algorithm. DTW is a dynamic programming technique and this algorithm has been used by various researchers for gesture, speech and handwriting recognition [63,64]. The algorithm temporally aligns two gesture sequences, a query and a model sequence. A matching score is then computed in form of distance at each time step between query and model sequence which is used as a similarity measure for classifying the query sequence. In our dataset, the recognition of query sign gesture was done by using k-NN approach in which, the similarity score was computed by Euclidean distance and 1-NN was used. An accuracy of 40.23% was recorded for all sign gestures, whereas, recognition rates of 30.48% and 41.73% were recorded in single and double handed sign gesture sequences, respectively. In our system, the signal-data obtained from each gesture is of certain duration and for each gesture we obtained 3D raw-data from each finger. Due to noise and duration of these signals (from fingers), DTW-matching based algorithms are not efficient for similarity finding. Therefore, a low gesture recognition rate was recorded with DTW based matching technique.

6. Conclusion and future scopes

In this paper, we have proposed a multimodal framework for SLR systems using Leap motion and Kinect sensor. The proposed model is robust as it captures sign inputs efficiently using two sensors than a typical single-sensor based data acquisition method. We have collected a large dataset consisting with 50 dynamic ISL words. These signs have been captured simultaneously by both devices. The recognition of sign gestures have been performed independently using Leap motion and Kinect datasets. Finally, combining data acquired using both sensors, we have achieved better accuracy. With the help of a classifier combination and fused features, we have obtained better results. In future, the method will be extended towards complete sign sentence recognition with the help of lexicon of sign words. In addition to that, the framework can be extended by capturing facial expressions and lip movements for further improvement.

References

[1] K. Assaleh, T. Shanableh, M. Zourob, Low complexity classification system for glove-based arabic sign language recognition, in: Proceedings of the International Conference on Neural Information Processing, 2012, pp. 262–268.

[2] H. Hongo, M. Ohya, M. Yasumoto, Y. Niwa, K. Yamamoto, Focus of attention for face and hand gesture recognition using multiple cameras, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition, 2000, pp. 156–161.

[3] X. Zhang, X. Chen, Y. Li, V. Lantz, K. Wang, J. Yang, A framework for hand gesture recognition based on accelerometer and EMG sensors, IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. 41 (6) (2011) 1064–1076.

[4] S.G.M. Almeida, F.G. Guimarães, J.A. Ramírez, Feature extraction in Brazilian sign language recognition based on phonological structure and using RGB-D sensors, Expert Syst. Appl. 41 (16) (2014) 7259–7271.

[5] K. Lai, J. Konrad, P. Ishwar, A gesture-driven computer interface using kinect, in: Proceedings of the Southwest Symposium on Image Analysis and Interpretation, 2012, pp. 185–188.

[6] C.-H. Chuan, E. Regina, C. Guardino, American sign language recognition using leap motion sensor, in: Proceedings of the 13th International Conference on Machine Learning and Applications, 2014, pp. 541–544.

[7] S. Stegmueller, "Hand and finger tracking with Kinect depth data." [2012-10-25], http://candescentnui.codeplex.com (2011).

[8] J.L. Raheja, A. Chaudhary, K. Singal, Tracking of fingertips and centers of palm using kinect, in: Proceedings of the 3rd International Conference on Computational Intelligence, Modelling and Simulation, 2011, pp. 248–252.

[9] Y. Wen, C. Hu, G. Yu, C. Wang, A robust method of detecting hand gestures using depth sensors, in: Proceedings of the International Workshop on Haptic Audio Visual Environments and Games, 2012, pp. 72–77.

[10] J. Suarez, R.R. Murphy, Hand gesture recognition with depth images: a review, in: Proceedings of the IEEE International Workshop on Robot and Human Communication RO-MAN, 2012, pp. 411–417.

[11] Y. Yao, Y. Fu, Contour model-based hand-gesture recognition using the kinect sensor, IEEE Trans. Circuits. Syst. Video Technol. 24 (11) (2014) 1935–1944.

[12] N. Rossol, I. Cheng, A. Basu, A multisensor technique for gesture recognition through intelligent skeletal pose analysis, IEEE Trans. Hum. Mach. Syst. 46 (3) (2016) 350–359.

[13] O. Cho, S. Lee, A study about honey bee dance serious game for kids using hand gesture, Int. J. Multimed. Ubiquitous Eng. 9 (6) (2014) 397–404.

[14] K. Vamsikrishna, D.P. Dogra, M.S. Desarkar, Computer-vision-assisted palm rehabilitation with supervised learning, IEEE Trans. Biomed. Eng. 63 (5) (2016) 991–1001.

[15] D. Charles, K. Pedlow, S. McDonough, K. Shek, T. Charles, An evaluation of the leap motion depth sensing camera for tracking hand and fingers motion in physical therapy, in: Proceedings of the International Conference on Interactive Technologies and Games, 2013.

[16] M. Khademi, H. Mousavi Hondori, A. McKenzie, L. Dodakian, C.V. Lopes, S.C. Cramer, Free-hand interaction with leap motion controller for stroke rehabilitation, in: Proceedings of the Conference on Human Factors in Computing Systems, 2014, pp. 1663–1668.

[17] J. Palacios, C. Sagues, E. Montijano, S. Llorente, Human-computer interaction based on hand gestures using RGB-D sensors, Sensors 13 (9) (2013) 11842–11860.

[18] M. ElBadawy, A.S. Elons, H. Sheded, M.F. Tolba, A proposed hybrid sensor architecture for arabic sign language recognition, in: Proceedings of the Intelligent Systems, 2015, pp. 721–730.

[19] G. Marin, F. Dominio, P. Zanuttigh, Hand gesture recognition with leap motion and kinect devices, in: Proceedings of the International Conference on Image Processing, 2014, pp. 1565–1569.

[20] K.-Y. Fok, C.-T. Cheng, N. Ganganath, Live demonstration: a HMM-based real-time sign language recognition system with multiple depth sensors, in: Proceedings of the International Symposium on Circuits and Systems, 2015, p. 1904.

[21] J. Han, L. Shao, D. Xu, J. Shotton, Enhanced computer vision with Microsoft kinect sensor: a review, IEEE Trans. Cybern. 43 (5) (2013) 1318–1334.

[22] R.Y. Wang, J. Popović, Real-time hand-tracking with a color glove, ACM Trans. Graph. 28 (3) (2009) 63.

[23] F.-S. Chen, C.-M. Fu, C.-L. Huang, Hand gesture recognition using a real-time tracking method and hidden Markov models, Image Vis. Comput. 21 (8) (2003) 745–758.

[24] J. Zieren, K.-F. Kraiss, Robust person-independent visual sign language recognition, in: Proceedings of the Conference on Pattern Recognition and Image Analysis, 2005, pp. 520–528.

[25] N. Tanibata, N. Shimada, Y. Shirai, Extraction of hand features for recognition of sign language words, in: Proceedings of the International conference on Vision Interface, 2002, pp. 391–398.

[26] T. Starner, J. Weaver, A. Pentland, Real-time american sign language recognition using desk and wearable computer based video, IEEE Trans. Pattern Anal. Mach. Intell. 20 (12) (1998) 1371–1375.

[27] A. Nandy, J.S. Prasad, S. Mondal, P. Chakraborty, G.C. Nandi, Recognition of isolated indian sign language gesture in real time, in: Proceedings of the Information Processing and Management, 2010, pp. 102–107.

[28] J. Rekha, J. Bhattacharya, S. Majumder, Shape, texture and local movement hand gesture features for indian sign language recognition, in: Proceedings of the 3rd International Conference on Trendz in Information Sciences & Computing, 2011, pp. 30–35.

[29] P. Premaratne, S. Ajaz, M. Premaratne, Hand gesture tracking and recognition system using Lucas–Kanade algorithms for control of consumer electronics, Neurocomputing 116 (2013) 242–249.

[30] R. Grzeszcuk, G. Bradski, M.H. Chu, J.-Y. Bouguet, Stereo based gesture recognition invariant to 3D pose and lighting, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 1, 2000, pp. 826–833.

[31] M. Elmezain, A. Al-Hamadi, B. Michaelis, Real-time capable system for hand gesture recognition using hidden Markov models in stereo color image sequences, J. WSCG 16 (1) (2008) 65–72.

[32] M.A. Mohandes, Recognition of two-handed arabic signs using the cyberglove, Arab. J. Sci. Eng. 38 (3) (2013) 669–677.

[33] N. Tubaiz, T. Shanableh, K. Assaleh, Glove-based continuous arabic sign language recognition in user-dependent mode, IEEE Trans. Human Mach. Syst. 45 (4) (2015) 526–533.

[34] Z. Zhang, Microsoft kinect sensor and its effect, IEEE Multimed. 19 (2) (2012) 4–10.

[35] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, P. Presti, American sign language recognition with the kinect, in: 13th International Conference on Multimodal Interfaces, 2011, pp. 279–286.

[36] K.K. Biswas, S.K. Basu, Gesture recognition using Microsoft kinect®, in: Proceedings of the 5th International Conference on Automation, Robotics and Applications, 2011, pp. 100–103.

[37] Z. Ren, J. Meng, J. Yuan, Z. Zhang, Robust hand gesture recognition with kinect sensor, in: Proceedings of the 19th International Conference on Multimedia, 2011, pp. 759–760.

[38] Z. Ren, J. Yuan, J. Meng, Z. Zhang, Robust part-based hand gesture recognition using kinect sensor, IEEE Trans. Multimed. 15 (5) (2013) 1110–1120.

[39] C. Keskin, F. Kıraç, Y.E. Kara, L. Akarun, Hand pose estimation and hand shape classification using multi-layered randomized decision forests, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 852–863.

[40] F. Kirac, Y.E. Kara, L. Akarun, Hierarchically constrained 3D hand pose estimation using regression forests from single frame depth data, Pattern Recogn. Lett. 50 (2014) 91–100.

[41] N. Pugeault, R. Bowden, Spelling it out: real-time ASL fingerspelling recognition, in: Proceedings of the International Conference on Computer Vision Workshops, 2011, pp. 1114–1119.

[42] S. Lang, M. Block, R. Rojas, Sign language recognition using kinect, in: International Conference on Artificial Intelligence and Soft Computing, Springer, Berlin Heidelberg, 2012, pp. 394–402.

[43] Y. Li, Hand gesture recognition using kinect, in: Proceedings of the 3rd International Conference on Software Engineering and Service Science, 2012, pp. 196–199.

[44] F. Pedersoli, S. Benini, N. Adami, R. Leonardi, Xkin: an open source framework for hand pose and gesture recognition using kinect, Vis. Comput. 30 (10) (2014) 1107–1122.

[45] K. Mehrotra, A. Godbole, S. Belhe, Indian sign language recognition using kinect sensor, in: Proceedings of the International Conference Image Analysis and Recognition, 2015, pp. 528–535.

[46] S.-Z. Li, B. Yu, W. Wu, S.-Z. Su, R.-R. Ji, Feature learning based on SAE–PCA network for human gesture recognition in RGB-D images, Neurocomputing 151 (2015) 565–573.

[47] H. Wang, X. Chai, X. Chen, Sparse observation (SO) alignment for sign language recognition, Neurocomputing 175 (2016) 674–685.

[48] L.E. Potter, J. Araullo, 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, 2013, pp. 175–178.

[49] A. Elons, M. Ahmed, H. Shedid, M. Tolba, Arabic sign language recognition using leap motion sensor, in: Proceedings of the 9th International Conference on Computer Engineering & Systems, 2014, pp. 368–373.

[50] S. Vikram, L. Li, S. Russell, Handwriting and gestures in the air, recognizing on the fly, in: Proceedings of the Conference on Human Factors in Computing Systems, 13, 2013, pp. 1179–1184.

[51] N. Xu, W. Wang, X. Qu, Recognition of in-air handwritten chinese character based on leap motion controller, in: International Conference on Image and Graphics, Springer International Publishing, 2015, p. 168.
[52] I. Nigam, M. Vatsa, R. Singh, Leap signature recognition using hoof and hot features, in: Proceedings of the International Conference on Image Processing, 2014, pp. 5012–5016.
[53] C. Agarwal, D.P. Dogra, R. Saini, P.P. Roy, Segmentation and recognition of text written in 3D using leap motion interface, in: Proceedings of the 3rd Asian Conference on Pattern Recognition, 2015, pp. 539–543.
[54] R. Mihail, N. Jacobs, J. Goldsmith, Static hand gesture recognition with 2 kinect sensors, in: Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition, 2012, p. 1.
[55] H. Hasan, S. Abdul-Kareem, Static hand gesture recognition using neural networks, Artif. Intell. Rev. 41 (2) (2014) 147–181.
[56] F. Dominio, M. Donadeo, P. Zanuttigh, Combining multiple depth-based descriptors for hand gesture recognition, Pattern Recogn. Lett. 50 (2014) 101–111.
[57] G. Marin, F. Dominio, P. Zanuttigh, Hand gesture recognition with jointly calibrated leap motion and depth sensor, Multimed. Tools. Appl. (2015) 1–25.
[58] I. Zubrycki, G. Granosik, Using integrated vision systems: three gears and leap motion, to control a 3-finger dexterous gripper, in: Proceedings of the Recent Advances in Automation, Robotics and Measuring Techniques, 2014, pp. 553–564.
[59] F.R. Al-Osaimi, M. Bennamoun, A. Mian, Spatially optimized data-level fusion of texture and shape for face recognition, IEEE Trans. Image Process. 21 (2) (2012) 859–872.
[60] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, J. Schmidhuber, A novel connectionist system for unconstrained handwriting recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (5) (2009) 855–868.
[61] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (5) (2005) 602–610.
[62] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1–3) (1987) 37–52.
[63] P. Doliotis, A. Stefan, C. McMurrough, D. Eckhard, V. Athitsos, Comparing gesture recognition accuracy using color and depth information, in: Proceedings of the 4th International Conference on Pervasive Technologies Related to Assistive Environments, 2011, pp. 1–7.
[64] S. Celebi, A.S. Aydin, T.T. Temiz, T. Arici, Gesture recognition using skeleton data with weighted dynamic time warping., in: Proceedings of the International Conference on Computer Vision Theory and Applications, 2013, pp. 620–625.

Pradeep Kumar is pursuing Ph.D in Department of Computer Science at IIT Roorkee, India. His research interest includes Gesture Recognition and Machine Learning.

Himaanshu Gauba is studying B.Tech at Indian Institute of Technology (IIT), Roorkee.

Partha Pratim Roy received his Ph.D. degree in computer science in 2010 from Universitat Aut$_5$noma de Barcelona, (Spain). He worked as postdoctoral research fellow in the Computer Science Laboratory (LI, RFAI group), France and in Synchromedia Lab, Canada. Presently, Dr. Roy is working as Assistant Professor at Indian Institute of Technology (IIT), Roorkee. His main research area is Pattern Recognition.

Debi Prosad Dogra is presently working as Assistant Professor in the School of Electrical Sciences of IIT Bhubaneswar. Earlier, he worked with various R&D organizations in India and abroad. He has obtained his doctorate degree in Computer Science & Engineering from IIT Kharagpur. His research interest includes video object tracking, visual surveillance, gesture recognition, augmented reality, and computer vision guided healthcare automation.