Multimed Tools Appl (2018) 77:8823–8846
DOI 10.1007/s11042-017-4776-9

A position and rotation invariant framework for sign language recognition (SLR) using Kinect

Pradeep Kumar[1] · Rajkumar Saini[1] · Partha Pratim Roy[1] · Debi Prosad Dogra[2]

Received: 11 October 2016 / Revised: 11 April 2017 / Accepted: 28 April 2017 /
Published online: 10 May 2017

Abstract Sign language is the only means of communication for speech and hearing impaired people. Using machine translation, Sign Language Recognition (SLR) systems provide medium of communication between speech and hearing impaired and others who have difficulty in understanding such languages. However, most of the SLR systems require the signer to sign in front of the capturing device/sensor. Such systems fail to recognize some gestures when the relative position of the signer is changed or when the body occlusion occurs due to position variations. In this paper, we present a robust position invariant SLR framework. A depth-sensor device (Kinect) has been used to obtain the signer's skeleton information. The framework is capable of recognizing occluded sign gestures and has been tested on a dataset of 2700 gestures. The recognition process has been performed using Hidden Markov Model (HMM) and the results show the efficiency of the proposed framework with an accuracy of 83.77% on occluded gestures.

Keywords Sign language · Depth sensors · Hidden Markov Model (HMM) · Occluded gestures

✉ Pradeep Kumar
pradeep.iitr7@gmail.com

Rajkumar Saini
rajkr.dcs2014@iitr.ac.in

Partha Pratim Roy
proy.fcs@iitr.ac.in

Debi Prosad Dogra
dpdogra@iitbbs.ac.in

[1] Department of Computer Science and Engineering, IIT Roorkee, Roorkee, India

[2] School of Electrical Sciences, IIT Bhubaneswar, Bhubaneswar, India

1 Introduction

Sign language is a form of visual language that uses grammatically structured manual and non-manual sign gestures for communication [45]. Manual gestures include hand shape, palm's orientation, location and movements, whereas non-manual gestures are represented by various facial expressions including head tilting, lip pattern, and mouthing [5, 39]. The language forms one of the natural means of communication among hearing impaired people. The goal of a Sign Language Recognition (SLR) system is to translate sign gestures into a meaningful text that helps persons without any speech or hearing disability can to understand sign language [25], hence, provides a natural interface for communication between humans and machines. A simple way to implement a SLR systems is based on tracking the position of hand and identifying relevant features to classify a given sign. This process of detecting and tracking the hand movements is relatively easier when compared with articulated or self occluded gestures and hand movements [20]. In literature, there exists a number of SLR systems proposed by various researchers for multiple sign languages including American [45], Australian [36], Indian [20], Spanish [13], and Greek [34], etc. However, most of the existing SLR systems require a signer to perform sign gestures in front of the capturing device, i.e., camera or sensor. These systems fail to recognize sign gestures correctly (i) when there is a change in the signer's relative position with respect to the camera or (ii) when the signer performs the gestures in a different plane leading to some change in Y-axis orientation. Such a scenario is depicted in Fig. 1, where a signer performs a sign gesture with a rotation along Y-axis in the camera coordinate system that results into self occlusion and a distorted view of the gestures being acquired. Therefore, pose and position invariant hand gesture tracking and recognition system can be very much helpful to improve the overall performance of the SLR systems and makes them usable for real life scenario including real time gesture recognition involving multiple signers, sign word spotting, etc.

With the advancement in low-cost depth sensing technology and emergence of sensors such as Leap motion and Microsoft Kinect, new possibilities in Human-Computer-Interaction (HCI) are evolving. These devices are designed to provide 3D point cloud of the observed scene. Kinect provides a 3D skeleton view of the human body through its rich Software Development Kit (SDK) [40]. 3D skeleton tracking can successfully address the

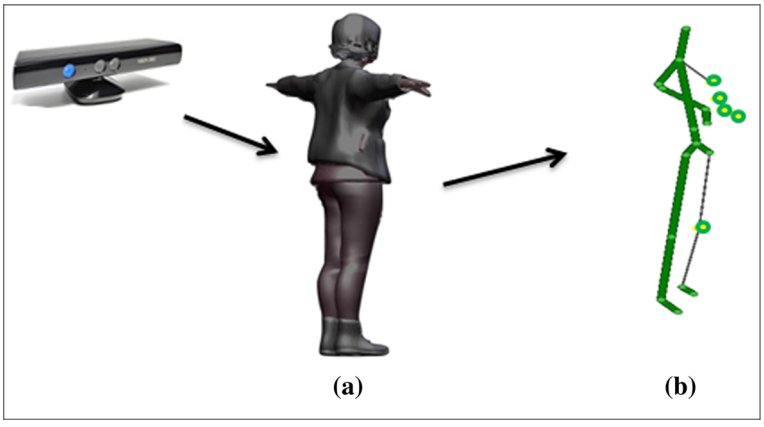


Fig. 1 Body occlusion occurs when a signer performs (**a**) sign gesture with rotation along Y-axis (**b**) a distorted human torso of the performed sign

body part segmentation problem, therefore, it is considered highly useful in hand gesture recognition. Illumination variation related problems that are usually encountered in images captured using traditional 2D cameras, can be avoided using such systems. Kinect has been successfully used in various applications including 3D interactive gaming [4], robotics [29], rehabilitation [14] and hand gesture recognitions [20, 26, 31]. Zafrulla et al. [45] have developed an automatic SLR system using Kinect based skeleton tracking. The authors have utilized 3D points of the upper body skeletal joints and fed these features to Hidden Markov Model (HMM) for recognition purpose. However, their system suffers from tracking errors when the users remain seated. In such cases, the main challenge is to extract self occluded, articulated or noisy sign gestures that can be used to perform recognition. In this paper, we propose a new framework for SLR using 3D points of body skeleton that can be used to recognized gestures independently irrespective to the signer's position or rotation with respect to the sensor. The main contributions to the paper are as follows:

(i) Firstly, we present a position and rotation invariant sign gesture tracking and recognition framework that can be used for designing SLR. Our system observes all gestures independently by transforming the 3D skeleton feature points with respect to one of the coordinate axes.
(ii) Secondly, we demonstrate the robustness of the proposed framework for recognition of self-occluded gestures using HMM. A comparative gesture recognition performance has also been presented using the HMM and SVM (Support Vector Machine) classifiers.

Rest of the paper is organized as follows. In Section 2, a chronological review of recent works in this field of study, is presented. System setup along the preprocessing and feature extraction are presented in Section 3. Experimental results are discussed in Section 4. Finally, we conclude with the future possibilities of the work in Section 5.

2 Related work

Hand gesture recognition is one of the basic steps of SLR systems. Handful of research work are being carried out in locating and extracting the hand trajectories. These work vary from vision-based skin color segmentation to depth-based analysis. To overcome the self-occlusion problem in hand gesture recognition, researchers have used multiple cameras to estimate 3D hand pose. Athitsos et al. [2] have proposed an estimation of 3D hand poses by finding the closest match between the input image and a large image database. The authors have used a database indexed with the help of Chamfer distance and probabilistic line matching algorithms by embedding binary edge images into a high dimensional Euclidean space. In [6], the authors have proposed a Relevance Vector Machine (RVM) based 3D hand pose estimation method to overcome the problem of self-occlusion using multiple cameras. The authors have extracted multiple-view descriptors for each camera image using shape contexts to form a high dimensional feature vector. Mapping between the descriptors and 3D hand pose has been done using regression analysis on RVM.

Recent development in depth sensor technology allows the users to acquire images with depth information. Depth cameras such as time-of-flight and Kinect have been successfully used by researchers for 3D hand and body estimation. Liu et al. [28] have proposed hand gesture recognition using time-of-flight camera to acquire color and depth images, simultaneously. The authors have extracted shape, location, trajectory, orientation and speed as

features from the acquired 3D trajectory. Chamfer distance has been used to find the similarity between two hand shapes. In [35], the authors have proposed a gesture recognition system using Kinect. Three basic gestures have been considered in the study using skeleton joint positions as features. Recognition of gestures has been performed using multiple classifiers namely, SVM, Backpropagation Neural Network (NN), Decision Tree and Naive Bayes where an average recognition rate of 93.72% was recorded in their work. Monir et al. [33], have proposed a human posture recognition system using Kinect 3D skeleton data points. Angular features of the skeleton data are used to represent the body posture. Three different matching matrices have been applied for recognition of postures where a recognition rate of 96.9% has been observed with priority based matching. Another study of hand gesture recognition using skeleton tracking has been proposed in [32] using torso-based coordinate system. The authors have used angular representation of the skeleton joints and SVM classifier to learn key poses, whereas a decision forest classifier has been used to recognize 18 gestures. Almeida et al. [1] have developed one SLR system for Brazilian Sign Language (BSL) using Kinect sensor. The authors have extracted seven vision-based features that are related to shape, movement and position of the hands. An average accuracy of 80% has been recorded on 34 BSL signs with the help of SVM classifier. In [27], the authors have proposed a covariance matrix based serial particle filter to track the hand movements in isolated sign gesture videos. Their methodology has been applied on 50 isolated ASL gestures with an accuracy of 87.33%.

Uebersax et al. [42] have proposed the SLR system for ASL using time-of-flight (TOF) camera. The authors have utilized depth information for hand segmentation and orientation estimation. Recognition of letters is based on average neighborhood margin maximization (ANMM), depth difference (DD), and hand rotation (ROT). Confidences of the letters are then combined to compute a word score. In [38], the authors have utilized Kinect sensor to develop gesture based arithmetic computation and rock-paper-scissors game. They have utilized depth maps as well as color images to detect the hand shapes. Recognition of gestures has been carried out using a distance metric to measure the dissimilarities between different hand shapes known as Finger-Earth Mover's Distance (FEMD). A Discriminative Exemplar Coding (DEC) based SLR system is proposed in [41] using Kinect. The authors have used background modeling to extract human body and hand segmentation. Next, multiple instance learning (MIL) has been applied to learn similarities between the frames using SVM, and AdaBoost technique has been used to select the most discriminative features. An accuracy of 85.5% was recorded on 73 sign gestures of ASL. Keskin et al. [18] have proposed a real time hand pose estimation system by creating a 3D hand model using Kinect. The authors have used Random Decision Forest (RDF) to perform per pixel classification and the results are then fed to a local mode finding algorithm to estimate the joint locations for the hand skeleton. The methodology has been applied to recognize 10 ASL digits, where an accuracy of 99.9% has been recorded using SVM. A Multi-Layered Random Forest (MLRF) has been used to recognize 24 static signs of ASL [24]. The authors have used Ensemble of Shape Function (ESF) descriptor that consist of a set of histograms to make the system translation, rotation and scale invariant. An accuracy of 85% has been recorded when tested on gestures of 4 subjects.

Chai et al. [5] have proposed the SLR and translation framework using Kinect. Recognition of gestures has been performed using a matching score computed with the help of Euclidean distance. The methodology has been tested on 239 Chinese Sign Language (CSL) words, where an accuracy of 83.51% has been recorded with top 1 choice. A hand contour

model based gesture recognition system has been proposed in [44]. Their model simplifies the gesture matching process to reduce the complexity of gesture recognition. The authors have used pixel's normal and the mean curvature to compute the feature vector for hand segmentation. The methodology has been applied to recognize 10 sign gestures with an accuracy of 96.1%. A 3D Convolutional Neural Network (CNN) has been utilized in [15] to develop a SLR system. The model extracts both spatial and temporal features by performing 3D convolutions on the raw video stream. The authors have used multi-channels of video streams, i.e., color information, depth data, and body joint positions, and these features have been fed as input to the 3D CNN. Multilayer Perceptron (MLP) classifier has been used to classify 25 sign gestures performed by 9 signers with an accuracy of 94.2%. In [43], the authors have used hierarchical Conditional Random Field (CRF) to detect candidate segments of signs using hand motions. A BoostMap embedding approach has been used to verify the hand shapes and segmented signs. However, their methodology requires a signer to wear a wrist-band during data collection. In [8], the authors have proposed one SLR system for 24 ASL alphabet recognition using Kinect. Per-pixel classification algorithm has been used to segment human hand into parts. The joint positions are obtained using a mean-shift mode-seeking algorithm and 13 angles of the hand skeleton have been used as the features to make the system invariant to the hand's size and rotational directions. Random Forest based classifier has been used for recognition of ASL gestures with an accuracy of 92%. In this work, we have used Affine transformation based methodology on 3D human skeleton to make the proposed SLR system position and rotation invariant.

To the best of our knowledge, all of the existing gesture recognition systems require the users to perform in front of the Kinect sensor. Therefore, such systems suffer when the users perform gestures that are recorded from side views. This creates occlusion, especially in the joints of the 3D skeleton. In the proposed framework, we present a solution to the self-occluded sign gestures. Our solution is position and rotation independent of the signer within the sensor's viewing field. A summary of the related work in comparison to the proposed methodology is presented in Table 1.

3 System setup

In this section, we present a detailed description of the proposed framework of the SLR system. It offers position and rotation independent sign gesture tracking and recognition. Since the displacement in signer's position with respect to the Kinect can change the origin of the coordinate system in the XZ-plane, it may cause difficulty in recognition. Similarly, when the signer performs a sign and the side view of the gesture is captured by the sensor, it may cause self-occlusion. A block diagram of the proposed framework is shown in Fig. 2, where the acquired 3D skeleton represents gesture sequences that undergoes Affine transformation. After transformation, the hands are segmented from the skeleton to extract gesture sequence and it is followed by feature extraction and recognition.

3.1 Affine transformation

After capturing the signer's skeleton information through Kinect, skeleton data are then processed through affine transformation. Affine transformation has been used to cancel out

Table 1 Summary of the related work in comparison to the proposed methodology

Author & Year	Approach	Number of Signers	Dataset	Accuracy (%)
Athitsos et al. [2], 2003	Chamfer distance, Edge orientation, line matching	n.a.	26 hand shapes	74.8%
Campos et al. [6], 2006	RVM regression, hand silhouette, Gaussian kernels	n.a.	1679 hand shapes	2-5% error
Liu et al. [28], 2004	Hamds shape, location, trajectory, orientation, speed, Chamfer distance	n.a.	10 signs	n.a.
Patsadu et al. [35], 2012	Kinect joint positions, SVM, Decision Tree, NN, Naive Bayes	6	3 gestuers	93.72%
Monir et al. [33], 2012	3D joint positions, angular features, priority matching	6	4 postures	96.9%
Miranda et al. [32], 2012	3D joint positions, joint angles, SVM, Decision forest	10	18 poses	94.84%
Almeida et al. [1], 2014	Hands shape, movement, position, SVM	n.a.	34 BSL signs	80%
Lim et al. [27], 2016	Serial particle filter, covariance matrix	3	50 ASL signs	87.33%
Uebersax et al. [42], 2011	Hand orientation, ANMM, DD, ROT	7	ASL alphabets	88%
Ren et al. [38], 2011	Depth maps, hand shapes, FEMD metric	n.a.	14 gestures	90.6%
Sun et al. [41], 2013	Background modeling, MIL-SVM, AdaBoost	9	73 ASL signs	85.5%
Keskin et al. [18], 2013	Hand joints, RDF, SVM	10	10 ASL	99.9%
Kuznetsova et al. [24], 2013	MLRF, ESF	4	24 ASL alphabets	85%
Chai et al. [5], 2013	3D motion data, Euclidean distance	n.a.	239 CSL signs	83.51%
Yao et al. [44], 2014	Hand contour, mean curvature, pixel normal	n.a.	10 gestures	96.1%
Huang et al. [15], 2015	color, depth, 3D joint data, 3D CNN, MLP	9	25 sign gestures	94.2%
Yang et al. [43], 2014	3D hand joints, hierarchical- CRF, BoostMap	n.a.	24 ASL signs	90.4%
Dong et al. [8], 2015	Per-pixel classification, 13 angles of hand skeleton, Random Forest	5	24 ASL alphabet	92%
Proposed Methodology	**3D hand joints, Affine Transformation, angular, curvature, velocity, dynamic features, HMM, SVM**	**10**	**30 ISL signs**	**83.77%**

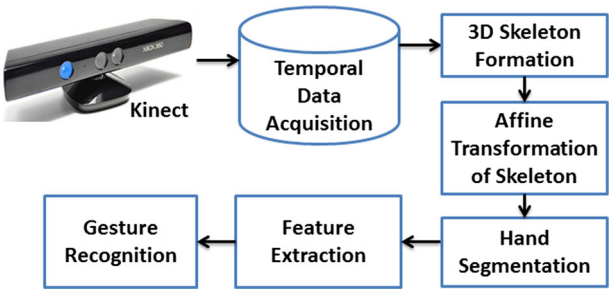


Fig. 2 Block diagram of the proposed framework for SLR

the effect of signer's rotation and position while performing the gestures. Two different 3D transformations , namely rotation and translation as given in (1) and (2), have been applied,

$$R_y^\theta = \begin{bmatrix} cos(\theta) & 0 & sin(\theta) & 0 \\ 0 & 1 & 0 & 0 \\ -sin(\theta) & 0 & cos(\theta) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

$$T(t_x, t_y, t_z) = \begin{bmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

where R_y^θ is the rotation matrix for rotating a 3D vector by an angle of θ about Y-axis, and $T(t_x, t_y, t_z)$ is the translation vector for translating the points in 3D.

3.1.1 Rotation invariant

If the signer does not perform sign in parallel to the Kinect sensor, then the torso makes an angle (θ_z) with the Z-axis of the sensor's coordinate system. For calculating θ_z, we have used three specific 3D points of the skeleton, i.e., left shoulder (L), right shoulder (R) and the spine center (C) that constitute the torso-plane (TP) as shown in Fig. 3.

TP has been used as the representative of the 3D skeleton. Next, two vectors $\overrightarrow{CL}$ and $\overrightarrow{CR}$ are computed on TP as shown in Fig. 4a.

Finally, a normal vector ($\hat{n}$) is estimated from TP with the help of $\overrightarrow{CL}$ and $\overrightarrow{CR}$, and it can be computed using (3). In our study, we made a zero-degree (0^o) angle between Z-axis and $\hat{n}$ for all gestures. Thus, while testing, θ_z is calculated using (4) by taking the projection of $\hat{n}$ in the XZ-plane,

$$\hat{n} = \frac{\overrightarrow{CL} \times \overrightarrow{CR}}{|\overrightarrow{CL}||\overrightarrow{CR}|} \tag{3}$$

$$cos(\theta_z) = \frac{\hat{n}.\hat{k}}{|\hat{n}||\hat{k}|} \tag{4}$$

where $\hat{k}$ is an unit vector $< 0, 0, 1 >$ along Z-axis. After estimating the value of θ_z, torso is rotated across Y-axis using (1) to cancel the effect of rotation of the signer as depicted in Fig. 4b.

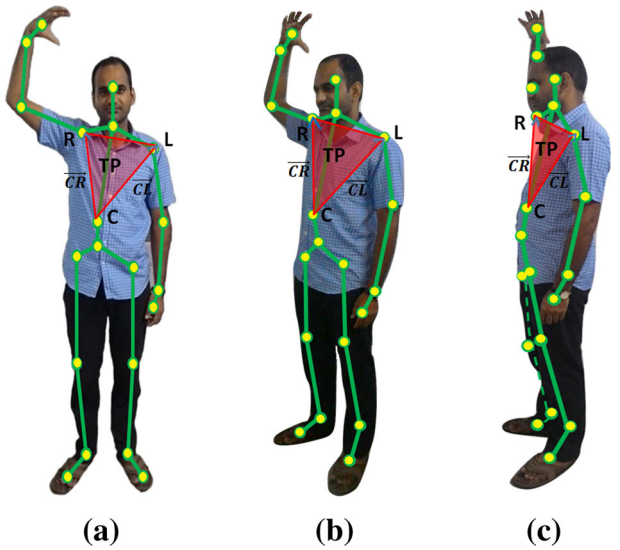


Fig. 3 Computation of the plane TP using three 3D points on skeleton as representative (**a**) with front view (**b**) side view-1 (45 degree approx. with Z-axis) (**c**) side view-2 (90 degree approx. with Z-axis)

3.1.2 Position invariant

After canceling out the rotational effect, we have used another heuristic to make the gesture recognition system independent of the position while the gestures are performed in the XZ-plane. To accomplish this, coordinates of the torso have been transformed from the sensor's frame of reference to a new frame of reference with respect to the signer. This is performed by translating the 3D point (C) of the skeleton to the center and shifting the rest of the data points with respect to this new origin using (2) as illustrated Fig. 5.

An alignment of all gestures improves recognition performance. Also, this makes the system position invariant. Rotation and position invariant steps have been applied on a typical

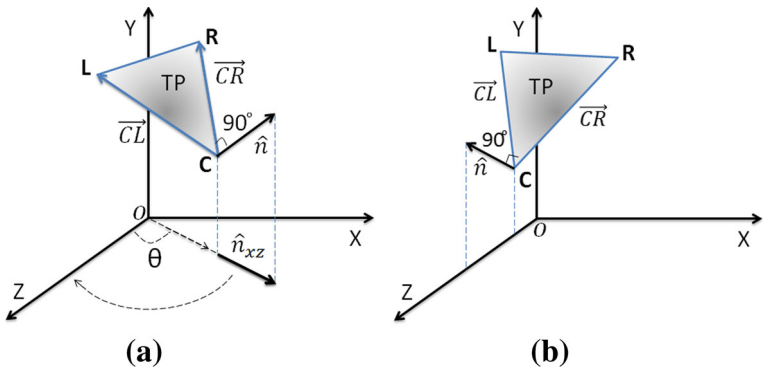


Fig. 4 Computation to cancel out the rotation effect (**a**) computation of θ_z before rotation (**b**) After rotation of θ_z about Y-axis

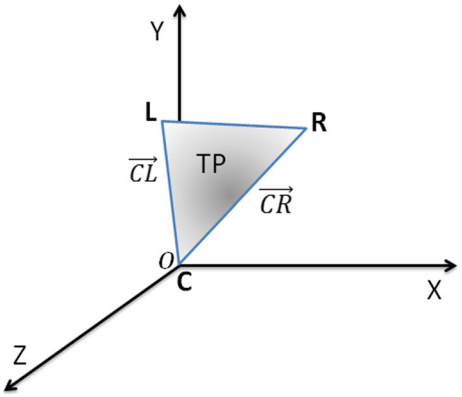


Fig. 5 Computation of position invariant by translating the 3D spine point C to center

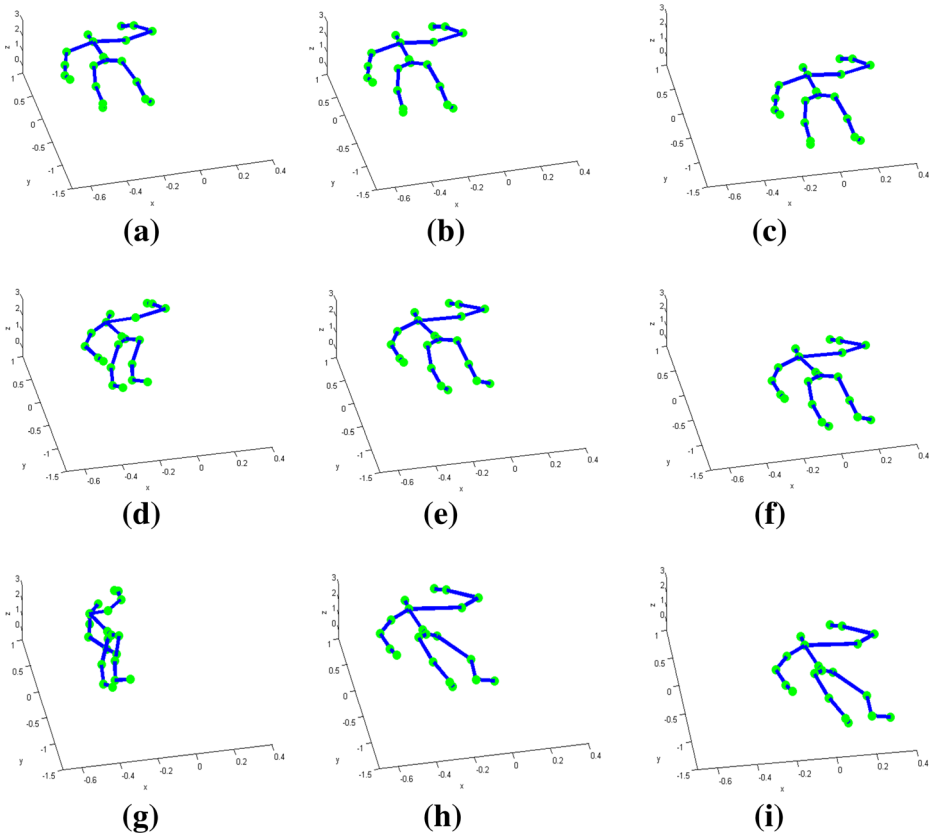


Fig. 6 Applying rotation and position invariant steps to a gesture captured at three different viewing angles: Figures in *first* column show a gesture performed in front, side view-1 and side view-2; Figures in *second* column show the gestures after applying rotation invariant step that rotates the torso about Y-axis; Figures in *third* column shows the outcome of the position invariant step that translates the torso to the origin with respect to spine

gesture performed on three different angles as shown in Fig. 6, where first column shows the raw gesture captured using Kinect, second column shows the outcome of the rotation invariant step, whereas the third column shows the translation of the torso to the center for making the position invariant.

3.1.3 Hand segmentation

Since gestures are performed either using single hand or both hands, both left (H_L) and right (H_R) hands are segmented from the 3D torso. For each hand, two 3D points, namely, wrist and hand-tip have been segmented from the torso that can be obtained using (5) and (6). This makes H_L and H_R of 6 dimensions each by concatenation of two 3D points,

$$H_L = [W_L \mid T_L] \tag{5}$$

$$H_R = [W_R \mid T_R] \tag{6}$$

where $[W_L \mid T_L]$ and $[W_R \mid T_R]$ are the wrist and hand-tips of left and right hands, respectively.

3.2 Feature extraction

Three different features have been extracted from the 3D segmented hands H_L and H_R, namely angular features (A_L and A_R), velocity (V_L and V_R) and curvature features (C_L and C_R). The details are as follows.

3.2.1 Angular direction

Angular features have been considered by various researchers in gesture recognition problems [7, 22, 30]. Angular direction corresponding to a 3D gesture sequence point $M(x, y, z)$ is computed with the help of two neighbor points, i.e., $L(x_1, y_1, z_1)$ and $N(x_2, y_2, z_2)$ as depicted in Fig. 7.

Neighboring points are selected in such a way that all points are non-collinear. In this work, N and L are the third neighboring points that lies on either side of M. By doing this,

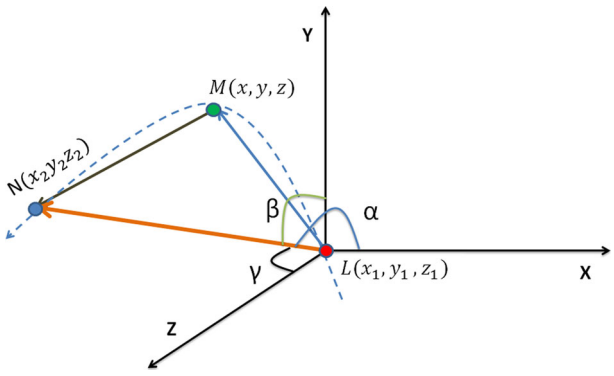


Fig. 7 Computation of the angular features of a 3D gesture sequence

we have ensured that all three points are collinear. A gesture sequence is shown in the Fig. 7 that forms a vector $\overrightarrow{LN}$ making α, β and γ angles with the coordinate axes. These angles can be calculated using (7) and (8). These angles are taken as the three angular direction features of the feature set. Both H_L and H_R consist of two 3D sequences. Therefore, 6 dimensional angular features A_L and A_R have been computed corresponding to H_L and H_R, respectively.

$$\overrightarrow{v} = \overrightarrow{OR} =< v_x, v_y, v_z >, |\overrightarrow{v}| = \sqrt{v_x^2 + v_y^2 + v_z^2} \tag{7}$$

$$cos(\alpha) = \frac{v_x}{|\overrightarrow{v}|}, cos(\beta) = \frac{v_y}{|\overrightarrow{v}|}, cos(\gamma) = \frac{v_z}{|\overrightarrow{v}|} \tag{8}$$

3.2.2 Velocity

Velocity features are based on the fact that, each gesture is performed at different speeds [46]. For example, certain gestures may include simple hand movements, thus, having uniform speed, whereas complex gestures may have varying speeds. Velocity (V) can be computed by measuring the distance between two successive points of the 3D gesture sequence, say (x_t, y_t, z_t) and $(x_{t+1}, y_{t+1}, z_{t+1})$, and it can be computed using (9).

$$V(x, y, z) = (x_{t+1}, y_{t+1}, z_{t+1}) - (x_t, y_t, z_t) \tag{9}$$

In this study, the velocity has been computed for both hands that results into a 6-dimensional feature vector V_L and V_R, each corresponds to one hand.

3.2.3 Curvature

Curvature feature represents a shape's curve and they are used to reflect the structural feature such as concavity and convexity. This has been successfully utilized in various gesture recognition tasks [7]. Curvature of a 3D point $B(x, y, z)$ on the gesture sequence is estimated using its two neighboring points $A(x_1, y_1, z_1)$, $C(x_2, y_2, z_2)$ and a circle is drawn if the points are non-collinear. This is accomplished with the help of two perpendicular bisectors $\overrightarrow{OM}$ and $\overrightarrow{ON}$ as depicted in Fig. 8, where the gesture sequence is marked using dashed line.

The center of the circle is calculated using (10). We have extracted five curvature related features, namely, the center (O), radius (r) of the circle, $\angle AOC$ marked as (θ) in the figure, and two normal vectors $\overrightarrow{OM}$ and $\overrightarrow{ON}$ that comprises of 11 dimensions.

$$\overrightarrow{O} = \frac{sin2A\ \overrightarrow{A} + sin2B\ \overrightarrow{B} + sin2C\ \overrightarrow{C}}{sin2A + sin2B + sin2C} \tag{10}$$

In this study, curvature has been computed for both hands which results into a 22-dimensional feature vector C_L and C_R each hand, respectively. Thus, by applying all three features, a new multi-dimensional feature vector (F_T) of 68-dimension is constructed as given in (11).

$$F_T = [H_L \mid H_R] = [A_L \mid A_R \mid V_L \mid V_R \mid C_L \mid C_R] \tag{11}$$

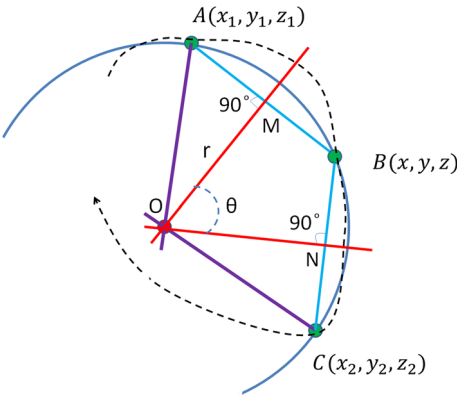


Fig. 8 Computation of the curvature features of a 3D gesture sequence

3.3 HMM guided gesture recognition

HMM is used for modeling the temporal sequences and it has been used by researchers in sign gesture and handwriting recognition systems [21, 22, 45]. HMM can be defined using $\{\pi, A, B\}$, with π as the initial probability distribution, $A = [a_{ij}]$, $i, j = 1, 2, \ldots N$ as the state transition matrix that has transition probability from state i to state j, and B defines the probability of observations with $b_j(O_k)$ as a density function from state j and observing a sequence O_k [19, 37]. For each state of the model, a Gaussian Mixture Model (GMM) is defined. The output probability density of the state j can be computed using (12),

$$b_j(x) = \sum_{k=1}^{M_j} c_{jk} \aleph(x, \mu_{jk}, \Sigma_{jk}) \tag{12}$$

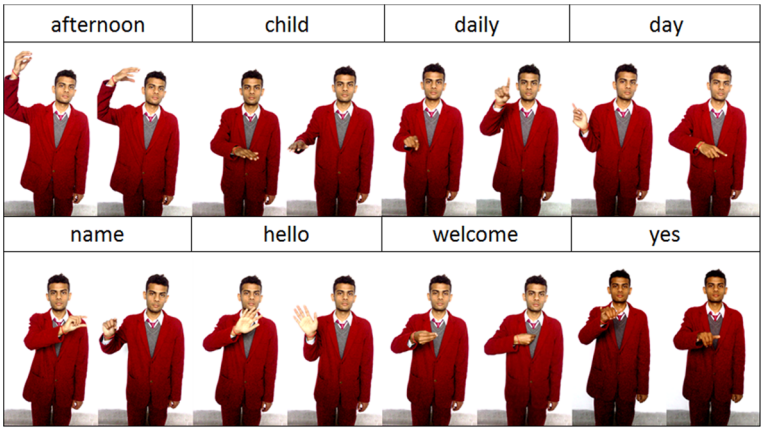
where M_j defines the number of Gaussian components assigned to j, and $\aleph(x, \mu, \Sigma)$ denotes the Gaussian with mean (μ) and co-variance matrix ($\sum$) and a weight coefficient (c_{jk}) of the Gaussian for component k of the state j. The observation probability of the sequence $O = (O_1, O_2, \ldots O_T)$ has been assumed to be generated by a state sequence $Q = Q_1, Q_2, \ldots Q_T$ of length T. This can be computed using (13), where π_{q_1} denotes the initial probability of start state.

$$P(O, Q|\lambda) = \sum_{Q} \pi_{q_1} b_{q_1}(O_1) \prod_{T} a_{q_{T-1}q_T} b_{q_T}(O_T) \tag{13}$$

HMM has been used for recognition of sign gestures and a set of HMMs are trained using the feature vector F_T defined in (11).

3.3.1 Dynamic context-independent feature

For boosting the sign gesture recognition system, we have included contextual information from neighboring windows by adding time derivatives in every feature vector. Such type of contextual and dynamic information in the current window helps to enhance the performance of recognition process [3]. The first and second-order dynamic features are known as



(a)

(b)

Fig. 9 Pictorial representation of sign gestures: (**a**) single-handed (**b**) double-handed. Note: Two instances of each gesture have been shown where one depicts the starting pose and the other is towards ending of gesture

delta and acceleration coefficients, respectively. Computation of delta coefficients is done with the help of first order regression of the feature vector using (14),

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{14}$$

where d_t is a delta coefficient at time t that has been computed in terms of static coefficients $c_{t-\Theta}$ to $c_{t+\Theta}$. Value of Θ is set according to the window size. Likewise, the acceleration coefficients can be obtained using second order regression. A temporal information has been captured by these derivative features at each frame that represents the dynamics of the features around the current window. In this study, the 68-dimensional feature vector (F_T) has been used along with the dynamic features discussed earlier to create a 204-dimensional feature vector for classification.

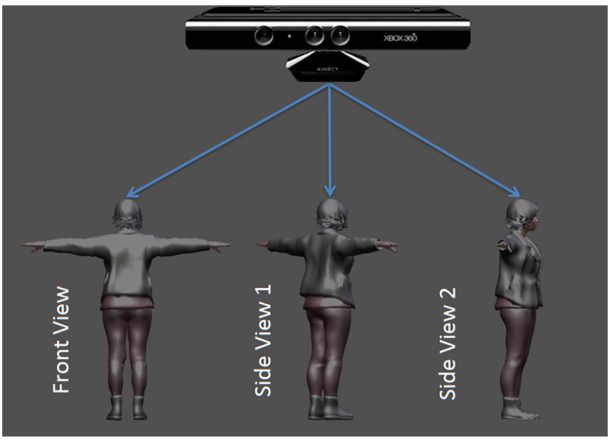


Fig. 10 Figure shows a gesture made by a signer in different view angles. **a** front view with zero-degree **b** side view 1 with 45^o approx. **c** side view 2 with 90^o approx

4 Results

We first present the dataset that has been prepared to test our proposed system. Next, we present gesture recognition results. We have carried out experiments in such a way that the training and test sets include gestures of different users.

4.1 Dataset description

A dataset of 30 isolated sign gestures of Indian Sign Language (ISL) has been prepared. The sign gestures have been performed by 10 different signers, where each sign has been performed 9 times by every signer. Hence, in total 2700 (i.e. $30 \times 9 \times 10$) gestures have been collected. Out of these 30 sign words, 16 words have been performed using single hand (right hand only), whereas remaining 14 words have been performed by both hands. Few examples of single and double-handed gestures are shown in Fig. 9.

In order to show the robustness of the proposed framework, all sign gestures have been performed at three different rotational angles as shown in Fig. 10, where a signer has performed sign gestures in three different directions, that make approximately 0°, 45°, and 90° angles between torso plane (TP) and Z-axis, respectively.

All these gestures from different view-angles were considered in our dataset. Similarly, signers have also changed their positions in the XZ-plane of the sensor's view field when performing different gestures. The 3D visualization of the gesture shows the variations when a gesture is performed by different signers. A 3D plot of the tip points during single-handed sign gesture '*bye*' is shown in Fig. 11 (first row) the gesture has been performed at different angles.

Different colors distinguish amongst various signers. Similarly, a 3D plot for the sign word '*dance*' is shown in Fig. 11 (second row), where large variations in the input sequence can be seen. We have made the dataset public for the research community.[1]

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

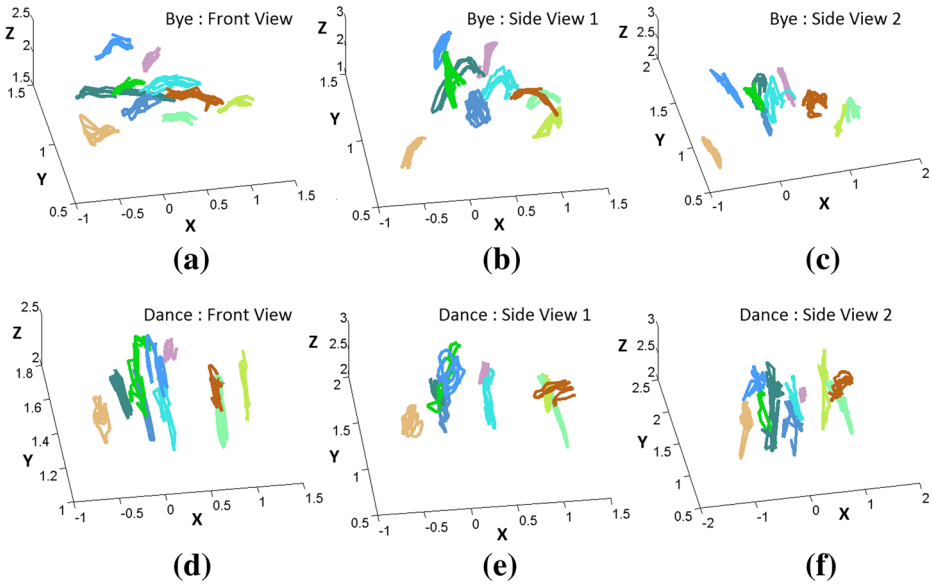


Fig. 11 Figure shows the variation of the same gesture in our dataset. *First row*: 3D plot for the sign word '*bye*' (1-handed) performed at different viewing angles (column 1: front view, column 2: side view 1, column 3: side view-2); *Second row*: 3D plot for sign word '*dance*' (2-handed) performed at different viewing angles by different signers

4.2 Experimental protocol

Experiments have been carried out using user-independent training of the gesture sequences. Our proposed methodology does not require a signer to enroll any gestures in the system for testing the system. HMM models are trained such that they do not depend on users. The results have been recorded using Leave-One-Out Cross-Validation (LOOCV) scheme. According to this scheme, the number of folds are equal to the number of instances. The learning algorithm is applied once for each instance, using all other instances as training set. In our experiments, we have kept gestures of 9 persons in training and test the gestures of 10^{th} person. The process is repeated for every person. Finally, an average of the results is recorded and reported. Recognition of gestures has been carried out in three modes, namely, for single hand, double hand, and using a combination of both.

4.3 HMM based gesture recognition

Gesture recognition has been performed without dynamic features as well as with dynamic features. The experiments have been carries out by varying HMM states, $S_t \in \{3, 4, 5, 6\}$ and by varying number of Gaussian mixture components per state from 1 to 256 with an incremental step of power of 2. Results using the framework by varying GMM components and HMM states are shown in Figs. 12 and 13, respectively using single, double and both hands gestures.

An accuracy of 81.29% has been recorded for single handed gestures at 64 Gaussians components and 3 HMM states, whereas accuracies of 84.81% and 83.77% have been recorded on double handed gestures and combined gestures with 4 HMM and 5 HMM states

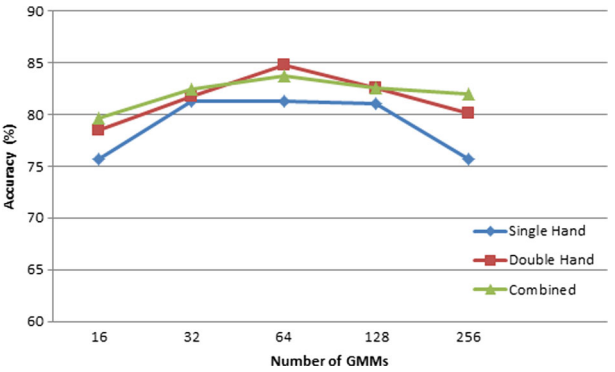


Fig. 12 Gesture recognition rate by varying Gaussian mixture components

and 128 Gaussians, respectively, when tested with dynamic features. Similarly, recognition of gestures have also been tested without using the dynamic features, where recognition rate of 75.29%, 78.56% and 73.54% have been recorded for single, double and combination modes, respectively. A comparison between the recognition rate of gesture with and without using dynamic features, is shown in Table 2.

It can be observed that the dynamic feature based gesture recognition outperforms non-dynamic set. The confusion matrix, using dynamic features in the form of heat map is shown in Fig. 14.

4.4 Rotation-wise results

In this section, we present the results obtained using different rotations as shown in Fig. 10. HMM classifier has been trained with the gestures that have been performed in the front view of the signer, whereas the gestures from side-views are kept for testing. Recognition has been carried out jointly on single as well as double-handed gestures. Performance has been compared with raw data and the results are presented in Fig. 15.

An accuracy of 86.67% has been obtained on front-view setup. We have obtained accuracies of 78.45% and 64.39% respectively on side view data. In all views, the proposed feature outperforms recognition using raw data.

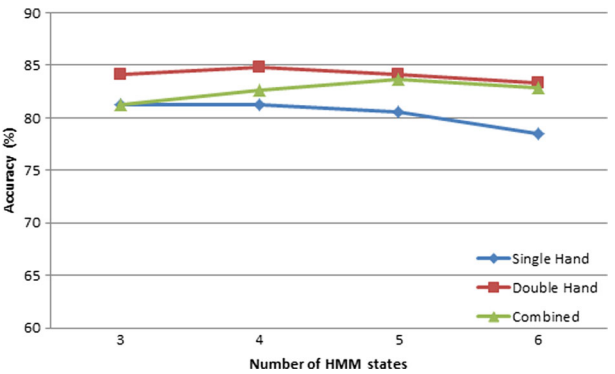


Fig. 13 Gesture recognition rate by varying HMM states

Table 2 Gesture recognition rate on HMM with and without dynamic features

Gesture-type	HMM with dynamic features	HMM without dynamic features
Single-handed	81.29 %	75.29 %
Double-handed	84.81 %	78.56 %
Combined (Single + Double)	83.77 %	73.54 %

In addition, the rotation-wise performance has also been computed by training the system with the complete preprocessed gestures including front and side views. Recognition results are depicted in Fig. 16, where the performance of the system has been increased in comparison to the accuracies obtained using front gestures based training.

4.5 Scalability test

A scalability test has also been performed by varying the training data, e.g. by varying number of signers (2,4,6,8) and by keeping the test data fixed during the experiments. These experiments have been carried out to test user-independence on the combined data for single as well as double-handed gestures and testing them on gestures of two signers while varying the training data. The recognition results are shown in Fig. 17, where an accuracy of 83% was recorded with 8 number of signers participated in training of the HMM classifier.

4.6 Comparative analysis

A comparative analysis of the proposed framework has been performed using SVM guided sign gesture recognition system. For this purpose, two different features, i.e., Mean and

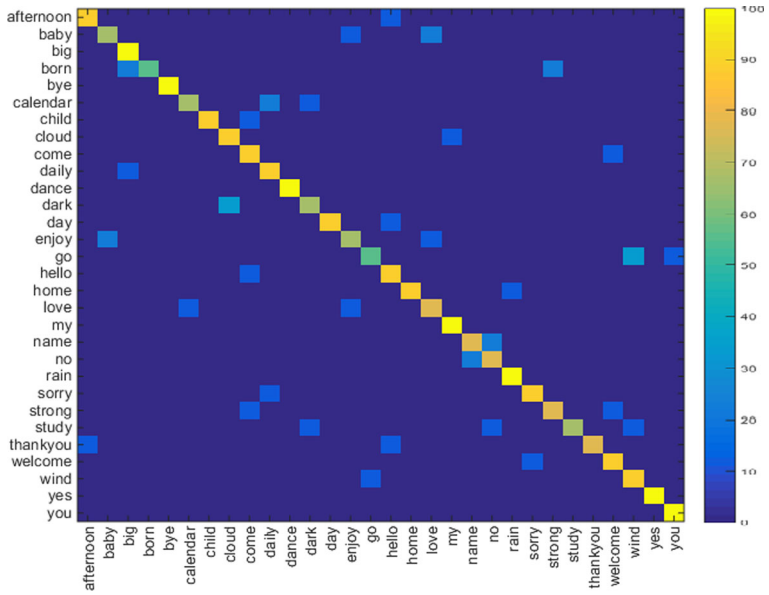


Fig. 14 Gesture recognition performance in the form of confusion matrix

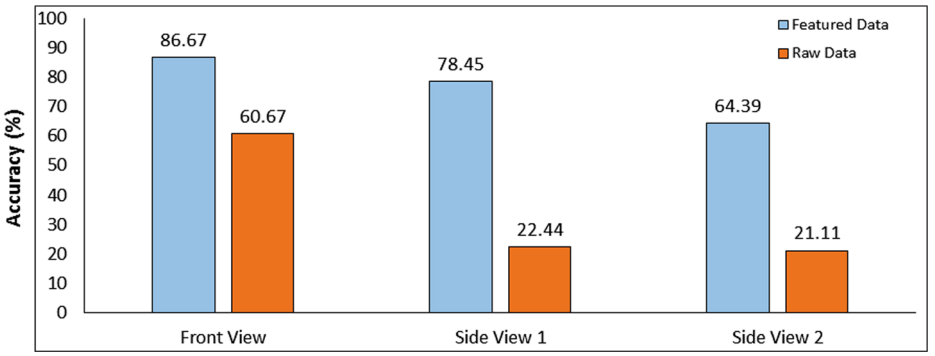


Fig. 15 Rotation wise gesture recognition in comparison to raw data when trained with only front view gestures

Standard Deviation have been extracted from the feature vector F_T. SVM classifier directly uses a hypothesis space for estimating the decision surface instead of modeling probability distribution of the training samples [17, 23]. The basic idea is to search an optimal hyperplane such that it maximizes the margins of the decision boundaries such that the worst-case generalization errors are minimized. For a set of M labeled training samples (x_i, y_i), where $x_i \in R^d$ and $y_i \in \{+1, -1\}$, the SVM classifier maps it into higher dimensional feature space using a non-linear operator $\phi(x)$. The optimal hyperplane is computed by the decision surface defined in (15),

$$f(x) = sign\left(\sum_i y_i \alpha_i K(x_i, x) + b\right) \tag{15}$$

where $K(x_i, x)$ is the kernel function. In this study, Radial Basis Function (RBF) kernel has been used to train the SVM model. Performance has been evaluated using the complete setup, i.e., preprocessed gestures in training and then applying the user-independent test mode. Finally, average results are reported. The value of the γ is kept fixed at 0.0049, whereas the regularization parameter C has been varied from 1 to 99 as shown in Fig. 18.

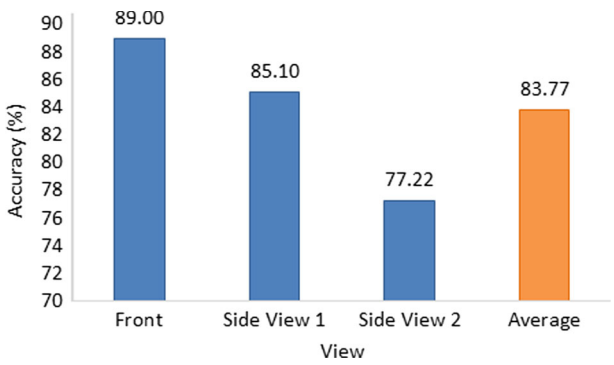


Fig. 16 Rotation wise gesture recognition when trained with all gestures (including front, side view 1 and side view 2)

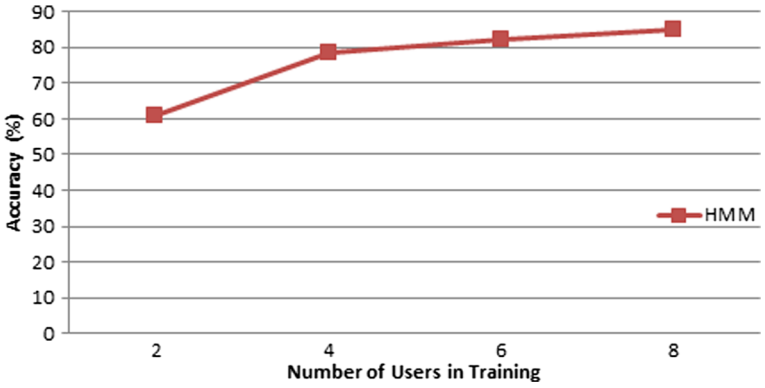


Fig. 17 Gesture recognition performance by varying training signers

Accuracies of 71.75%, 77.77% and 70.91% have been recorded on three different values of C, i.e., 98, 87, and 91 for single-handed, double-handed and combined gestures, respectively.

Rotation-wise results have been computed by training the system with front gestures as well as with complete dataset. Recognition results are depicted in Fig. 19.

In addition, the accuracies of all views have also been computed in user-dependent training using 9-fold cross validation scheme. The dataset has been divided into 9 equal parts and 8 parts of them have been kept in training and test the remaining part. Similarly, all the parts have been tested and average results are computed. Recognition accuracies of all views are shown in Fig. 20, where an average performance of 90.26% is recorded.

To the best of our knowledge, no other method exists with which we can compare our method directly. However, viable comparisons of the front-view gestures are performed with two publicly available datasets, namely, GSL20 [34] and CHALEARN [10]. The dataset GSL20 consist 20 sign gestures of Greek Sign Language (GSL), whereas the CHALEARN

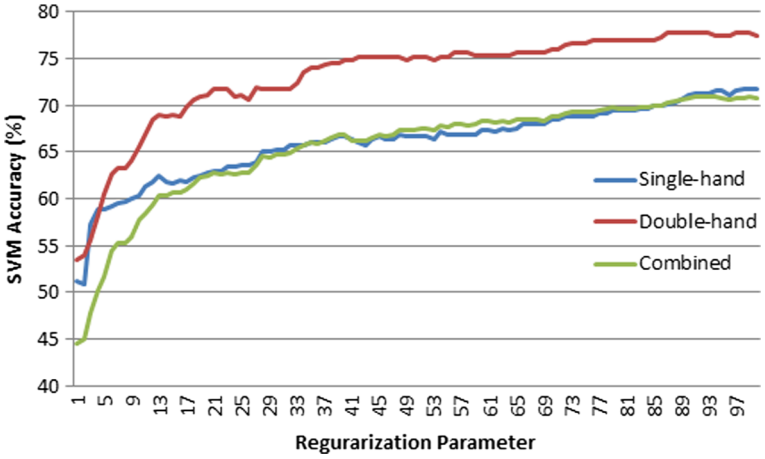


Fig. 18 Gesture recognition performance using SVM by varying regularization parameter

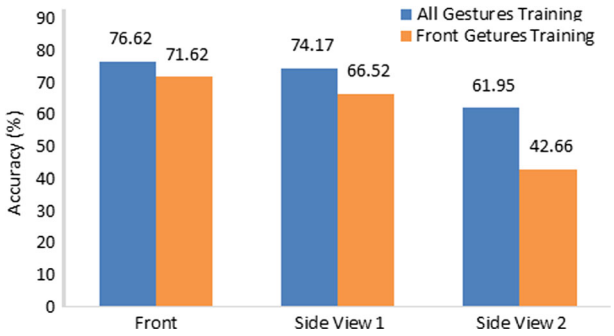


Fig. 19 Rotation-wise results by training with front view and all gestures (including front and side views) using SVM

dataset consist of 20 Italian sign gestures recorded with Kinect sensor. In CHALEARN dataset, the accuracy is reported on the validation set due to non-availability of labels in test data [12]. The authors in [11] have considered 7 joints of the 3D skeleton, namely, shoulder center (SC), elbow right (ER), elbow left (EL), hand right (HR), hand left (HL), wrist right (WR) and wrist left (WL) whereas in our methodology we have considered only 4 joints, i.e. hand right (HR), hand left (HL), wrist right (WR) and wrist left (WL). Therefore, achieved lower accuracy in comparison to [11]. The comparative performance is presented in Table 3.

4.7 Error analysis

This section presents an analysis on failure cases. We show a confusion matrix in Fig. 14 for such results. Some gestures have not been recognized because of the presence of distortions in the data even after the affine transformation. Moreover, some gestures share similar movements, shape and hand orientation that creates some confusion within the set. Hence, they have been recognized falsely. For examples, the single-handed sign gesture representing '*name*' and '*no*' have similar hand movements except the speed and position of hand. Similarly, two-handed sign gestures for the word '*wind*' and '*go*' also share similar characteristics in terms of movements and positions of the hands.

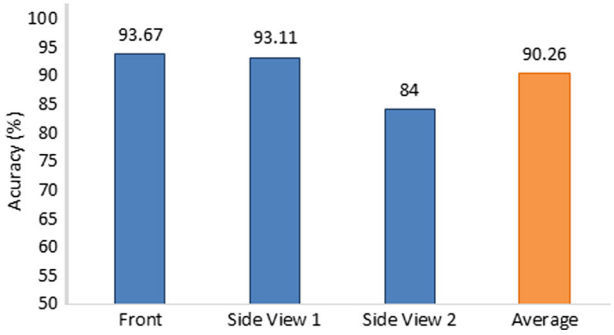


Fig. 20 Recognition results of all views in user-dependent training

Table 3 Comparative analysis of proposed SLR system with existing methodologies

Dataset	Number of Signers	Number of Gestures	Accuracy [9]	Accuracy [11]	Accuracy [16]	Proposed Methodology
GSL20	6	20	76%	n.a.	n.a.	75.97%
CHALEARN	27	20	n.a.	76% (Local Features, 7 joints)	59.91% (only skeleton)	63.34% (4 joints)
Proposed Dataset	10	30	n.a.	n.a.	n.a.	83.77 %

5 Conclusion and future work

In this paper, we have proposed a rotation and position invariant framework for SLR that provides an effective solutions to recognize self-occluded gestures. Our system does not require a signer to perform sign gestures in front of the sensor. Hence, provides a natural way of interaction. The proposed framework has been tested on a large dataset of 2700 sign words of ISL that have been collected with varying rotations and positions of the signer in the field of view of the sensor. Recognition has been carried out using HMM classifier in three modes using single-handed, double-handed and combined setup. Results show the robustness of the proposed framework with an overall accuracy of 83.77% using the combined setup. In future, the work can be extended to the recognition of interaction between multiple persons. Vision based approaches in combination with depth sequences and 3D skeleton could also help in boosting recognition performance. Additionally, more robust features and classifiers such as Recurrent Neural Network (RNN) could also be explored to improve the performance further.

Acknowledgments The authors would like to thank the anonymous reviewers for their constructive comments and suggestions to improve the quality of the paper. We are also thankful to our signers who are the students of an intermediate school ‘Anushruti’ at IIT Roorkee, India.

References

1. Almeida SGM, Guimarães FG, Ramírez JA (2014) Feature extraction in Brazilian sign language recognition based on phonological structure and using rgb-d sensors. Expert Syst Appl 41(16):7259–7271
2. Athitsos V, Sclaroff S (2003) Estimating 3d hand pose from a cluttered image. In: Computer Vision and Pattern Recognition, volume 2, pp II–432
3. Bianne-Bernard A-L, Menasri F, Mohamad RA-H, Mokbel C, Kermorvant C, Likforman-Sulem L (2011) Dynamic and contextual information in hmm modeling for handwritten word recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(10):2066–2080
4. Bleiweiss A, Eshar D, Kutliroff G, Lerner A, Oshrat Y, Yanai Y (2010) Enhanced interactive gaming by blending full-body tracking and gesture animation. In: ACM SIGGRAPH ASIA Sketches, p 34
5. Chai X, Li G, Lin Y, Xu Z, Tang Y, Chen X, Zhou M (2013) Sign language recognition and translation with kinect. In: Conference on Automatic Face and Gesture Recognition
6. de Campos TE, Murray DW (2006) Regression-based hand pose estimation from multiple cameras. In: International Conference on Computer Vision and Pattern Recognition, vol 1, pp 782–789
7. Dominio F, Donadeo M, Zanuttigh P (2014) Combining multiple depth-based descriptors for hand gesture recognition. Pattern Recogn Lett 50:101–111
8. Dong C, Leu MC, Yin Z (2015) American sign language alphabet recognition using microsoft kinect. In: Conference on Computer Vision and Pattern Recognition Workshops, pp 44–52

9. Elliott R, Cooper H, Ong E-J, Glauert J, Bowden R, Lefebvre-Albaret F (2011) Search-by-example in multilingual sign language databases. In: Sign Language Translation and Avatar Technologies Workshops
10. Escalera S, Gonzàlez J, Baró X, Reyes M, Lopes O, Guyon I, Athitsos V, Escalante H (2013) Multi-modal gesture recognition challenge 2013: Dataset and results. In: 15th International conference on multimodal interaction, pp 445–452
11. Escobedo-Cardenas E, Camara-Chavez G (2015) A robust gesture recognition using hand local data and skeleton trajectory. In: International Conference on Image Processing, pp 1240–1244
12. Fernando B, Efstratios G, Oramas J, Ghodrati A, Tuytelaars T (2016) Rank pooling for action recognition. IEEE transactions on pattern analysis and machine intelligence
13. García Incertis I, Gomez Garcia-Bermejo J, Zalama Casanova E (2006) Hand gesture recognition for deaf people interfacing. In: 18th International Conference on Pattern Recognition, vol 2, pp 100–103
14. González-Ortega D, Díaz-pernas FJ, Martínez-Zarzuela M, Antón-Rodríguez M (2014) A kinect-based system for cognitive rehabilitation exercises monitoring. Comput Methods Prog Biomed 113(2):620–631
15. Huang J, Zhou W, Li H, Li W (2015) Sign language recognition using 3d convolutional neural networks. In: International Conference on Multimedia and Expo, pp 1–6
16. Jiaxiang WU, Cheng J, Zhao C, Hanqing LU (2013) Fusing multi-modal features for gesture recognition. In: 15th International conference on multimodal interaction, pp 453–460
17. Kaur B, Singh D, Roy PP A novel framework of eeg-based user identification by analyzing music-listening behavior. Multimedia Tools and Applications
18. Keskin C, Kıraç F, Kara YE, Akarun L (2013) Real time hand pose estimation using depth sensors. In: Consumer Depth Cameras for Computer Vision. Springer, pp 119–137
19. Kumar P, Gauba H, Roy PP, Dogra DP (2016) A multimodal framework for sensor based sign language recognition. Neurocomputing
20. Kumar P, Gauba H, Roy PP, Dogra DP (2016) Coupled hmm-based multi-sensor data fusion for sign language recognition. Pattern Recognition Letters
21. Kumar P, Saini R, Roy P, Dogra D (2016) Study of text segmentation and recognition using leap motion sensor. IEEE Sensors Journal
22. Kumar P, Saini R, Roy PP, Dogra DP (2016) 3d text segmentation and recognition using leap motion. Multimedia Tools and Applications
23. Kumar P, Saini R, Roy PP, Dogra DP (2017) A bio-signal based framework to secure mobile devices. Journal of Network and Computer Applications
24. Kuznetsova A, Leal-Taixé L, Rosenhahn B (2013) Real-time sign language recognition using a consumer depth camera. In: International Conference on Computer Vision Workshops, pp 83–90
25. Lang S, Block M, Rojas R (2012) Sign language recognition using kinect. In: International Conference on Artificial Intelligence and Soft Computing, pp 394–402
26. Li Y (2012) Hand gesture recognition using kinect. In: International Conference on Computer Science and Automation Engineering, pp 196–199
27. Lim KM, Tan AWC, Tan SC (2016) A feature covariance matrix with serial particle filter for isolated sign language recognition. Expert Syst Appl 54:208–218
28. Liu X, Fujimura K (2004) Hand gesture recognition using depth data. In: Automatic Face and Gesture Recognition, pp 529–534
29. Machida E, Cao M, Murao T, Hashimoto H (2012) Human motion tracking of mobile robot with kinect 3d sensor. In: Annual Conference of The Society of Instrument and Control Enginners, pp 2207–2211
30. Marin G, Dominio F, Zanuttigh P (2015) Hand gesture recognition with jointly calibrated leap motion and depth sensor. Multimedia Tools and Applications
31. Martínez-Camarena M, Oramas MJ, Tuytelaars T (2015) Towards sign language recognition based on body parts relations. In: International Conference on Image Processing, pp 2454–2458
32. Miranda L, Vieira T, Martinez D, Lewiner T, Vieira AW, Campos MFM (2012) Real-time gesture recognition from depth data through key poses learning and decision forests. In: Conference on graphics, Patterns and Images, vol 25, pp 268–275
33. Monir S, Rubya S, Ferdous HS (2012) Rotation and scale invariant posture recognition using microsoft kinect skeletal tracking feature. In: International Conference on Intelligent Systems Design and Applications, vol 12, pp 404–409
34. Ong E-J, Cooper H, Pugeault N, Bowden R (2012) Sign language recognition using sequential pattern trees. In: Conference on Computer Vision and Pattern Recognition, pp 2200–2207
35. Patsadu O, Nukoolkit C, Watanapa B (2012) Human gesture recognition using kinect camera. In: International Joint Conference on Computer Science and Software Engineering, pp 28–32
36. Potter LE, Araullo J, Carter L (2013) The leap motion controller: a view on sign language. In: 25th Australian computer-human interaction conference: augmentation, application, innovation, collaboration, pp 175–178

37. Rabiner L (1989) A tutorial on hidden markov models and selected applications in speech recognition. Readings in Speech Recognition 77(2):257–286
38. Ren Z, Meng J, Yuan J, Zhang Z (2011) Robust hand gesture recognition with kinect sensor. In: 19th International Conference on Multimedia, pp 759–760
39. Starner T, Weaver J, Pentland A (1998) Real-time american sign language recognition using desk and wearable computer based video. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(12):1371–1375
40. Suarez J, Murphy RR (2012) Hand gesture recognition with depth images: A review. In: International Symposium on Robot and Human Interactive Communication, vol 21, pp 411–417
41. Sun C, Zhang T, Bao B-K, Changsheng XU, Mei T (2013) Discriminative exemplar coding for sign language recognition with kinect. IEEE Transactions on Cybernetics 43(5):1418–1428
42. Uebersax D, Gall J, Van den Bergh M, Gool LV (2011) Real-time sign language letter and word recognition from depth data. In: International Conference on Computer Vision Workshops, pp 383–390
43. Yang H-D (2014) Sign language recognition with the kinect sensor based on conditional random fields. Sensors 15(1):135–147
44. Yao Y, Yun FU (2014) Contour model-based hand-gesture recognition using the kinect sensor. IEEE Transactions on Circuits and Systems for Video Technology 24(11):1935–1944
45. Zafrulla Z, Brashear H, Starner T, Hamilton H, Presti P (2011) American sign language recognition with the kinect. In: International conference on multimodal interfaces, vol 13, pp 279–286
46. Zhang XU, Chen X, Li Y, Lantz V, Wang K, Yang J (2011) A framework for hand gesture recognition based on accelerometer and emg sensors. Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 41(6):1064–1076

Pradeep Kumar is pursuing Ph D in Department of Computer Science and Engineering at IIT Roorkee, India. His research interest includes Human Computer Interaction (HCI) and Brain Computer Interface (BCI).

Rajkumar Saini is pursuing his PhD in Department of Computer Science at IIT Roorkee, India. His research interest includes Pattern Recognition, Machine Learning.

Partha Pratim Roy received his Ph.D. degree in computer science in 2010 from Universitat Autònoma 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.