

# An efficient gesture based humanoid learning using wavelet descriptor and MFCC techniques

Neha Baranwal<sup>1</sup> · G. C. Nandi<sup>1</sup>

Received: 23 May 2015 / Accepted: 5 February 2016 / Published online: 17 February 2016  
© Springer-Verlag Berlin Heidelberg 2016

**Abstract** Recognizing any gesture, pre-processing and feature extraction are the two major issues which we have solved by proposing a novel concept of Indian Sign Language (ISL) gesture recognition in which a combination of wavelet descriptor (WD) and Mel Sec Frequency Cepstral Coefficients (MFCC) feature extraction technique have been used. This combination is very effective against noise reduction and extraction of invariant features. Here we used WD for reducing dimensionality of the data and moment invariant point extraction of hand gestures. After that MFCC is used for finding the spectral envelope of an image frame. This spectral envelope quality is useful for recognizing hand gestures in complex environment by eliminating darkness present in each gesture. These feature vectors are then used for classifying a probe gestures using support vector machine (SVM) and K nearest neighbour classifiers. Performance of our proposed methodology has been tested on in house ISL datasets as well as on Sheffield Kinect gesture dataset. From experimental results we observed that WD with MFCC method provides high recognition rate as compare to other existing techniques [MFCC, orientation histogram (OH)]. Subsequently, ISL gestures have been transferred to a Humanoid HOAP-2 (humanoid open architecture platform) robot in Webots simulation platform. Then these gestures are imitated by HOAP-2 robot exactly in a same manner.

**Keywords** Computer vision · DWT · Indian sign language · MFCC · SVM

## 1 Introduction

Nowadays robot can assist in various fields like doctors in his/her surgery during the critical operations, war field, household applications, etc. For such type of activity, interaction of human with robot is necessary. Many algorithms and many methodologies have already been evolved and many research works are currently running to make the robot/machine as intelligent as human beings. Both gesture and speech are the good ways of establishing a communication between human and robot [1, 2]. Gestures can be formed by hand and head movements or sometimes by full body movements which is mostly used by hearing impaired society. This society uses this sign language for establishing a communication with each other. In this work hand gestures have been taken into account. Previously gestures are recognized using data glove based gesture capturing techniques where sensors are used for capturing the joint angle values of hand. Using these angle values movements of hand gestures are identified. This method is not substantial for recognizing hand gestures therefore vision based gesture recognition technique have been evolved. This technique has the good capability of capturing gestures because of the advancement in technology of capturing devices like camera and processing of the high quality images and also it is easy to handle. In today's scenario sign language recognition attracts the system which can make the communication straightforward and easy for impaired society. Every country has its own sign language like American has American Sign Language, British has British sign language, etc. in the same way

---

✉ Neha Baranwal  
baranwal.neha24@gmail.com

G. C. Nandi  
gcnandi@gmail.com

<sup>1</sup> Robotics and AI Lab, Indian Institute of Information Technology, Allahabad, India

Indian has its own sign language which is called as Indian sign language (ISL). ISL is a best way of establishing a communication with hearing impaired society present in India. In ISL dynamic and static hand movements are performed by single as well as both the hands. After giving much time to an end ever to understand this language, we found that the beauty of establishing communication through sign gestures depends upon the way of making it. Sometimes hand gestures contain other body parts too. ISL Library helped us a lot to understand ISL signs.

This paper is an extension of our previous work [25] where we apply MFCC technique to extract features of hand. Here data set have been taken in very structured environment like homogeneous light condition, black background and also we concentrate on the hands portion from full human body. Due to these constraints the data set are linear in nature which does not require any pre-processing. The MFCC technique provides good recognition accuracy when applied on these structured dataset. But its performance decreases as the complexity increases in real time like variation in background, light, shadow, illumination conditions and minor change in shape colour of hands which are difficult to discriminate and also it will depend on the speed of the gesture. The major challenges we have solved in this paper are:

- (a) Hand segmentation from upper half of the body image.
- (b) During gesture segmentation process some of the gestures are deformed or slide change in their shape.
- (c) Boundaries of each gesture may vary from one person to another.

To minimize all such challenges we have proposed a novel framework which works well in real time scenarios. Here we first extract hand from upper half of the body image. After solving this, next step is to find appropriate features for gesture recognition like shape, orientation, spatial temporal motion, etc. In this framework we have applied a combination of WD and MFCC feature extraction technique for recognizing an unknown ISL gesture with the help of SVM and KNN as a classifier. The combinations of these two feature extraction techniques are never been applied before for ISL gesture recognition purpose and also it is very effective against translation, scaling, orientation, background variation and light variation when gestures are performed in real time environment. After WD, 12 MFCC coefficients of each frame are taken as a feature vector. All the experiments are performed in various background conditions with different illuminations like red, yellow, etc. and we get 98 percent classification results. After classification process humanoid learning is performed by HOAP-2 robot using Webots robotic simulation software.

This paper organizes into six sections in which second section describes about analysis of previous research where we describes that what are the research work has already been done and what are the flaws in existing techniques. In third section, detail of the proposed framework has been described. Experimental results and analysis are explained in fourth section. Fifth section tells about performance comparison between different gesture recognition techniques and explains how humanoid learning is performed. Section sixth shows the findings and future scope in this area. Finally end of the paper includes acknowledgement and references.

## 2 Analysis of previous research work

Many vision based sign language gesture recognition techniques have already been evolved. Here we will discuss few of them. A vision-based method for hand gesture data acquisition requires many important aspects for consideration like placement of camera and number of cameras. In general, one camera is used by Starner [5] and Martin [6] and they suggested that one camera can provide appropriate information. More than one camera can be used for getting depth or stereo information. Here gestures are recognized in real time environment using hidden Markov model (HMM) which is very effective tool in speech recognition field. The major drawback of this paper is the features here only mean and variance has been used as an input for HMM. These feature loses information regarding hand.

In case of feature extraction Sturman [7, 8] and Watson [9] have discussed template matching based technique which is done by creating templates for collecting data values of each frame of each gesture. Firstly the sensor reading is averaged and saved as templates. Test template is classified by closest matching criteria. It is easy to implement but not well suited for hand gestures because of overlapping templates in case of large posture set. Histograms of local orientation used as feature vectors by William T. Freeman and Michal Roth [10]. They added advantages of using orientation as robustness towards light variations and histogram as translational invariance. Nandy et al. [11] proposed an ISL gesture recognition technique where features are extracted using histogram equalization and atan2 function and classified using Euclidean distance metric. In histogram equalization cumulative distribution function (CDF) is used for normalizing image database. Here database is created with fixed background (black) and different illumination conditions. They also mentioned the flaws in this method in case of complex background conditions [12]. In addition to this, only fixed background dynamic gestures are well classified and feature vector for

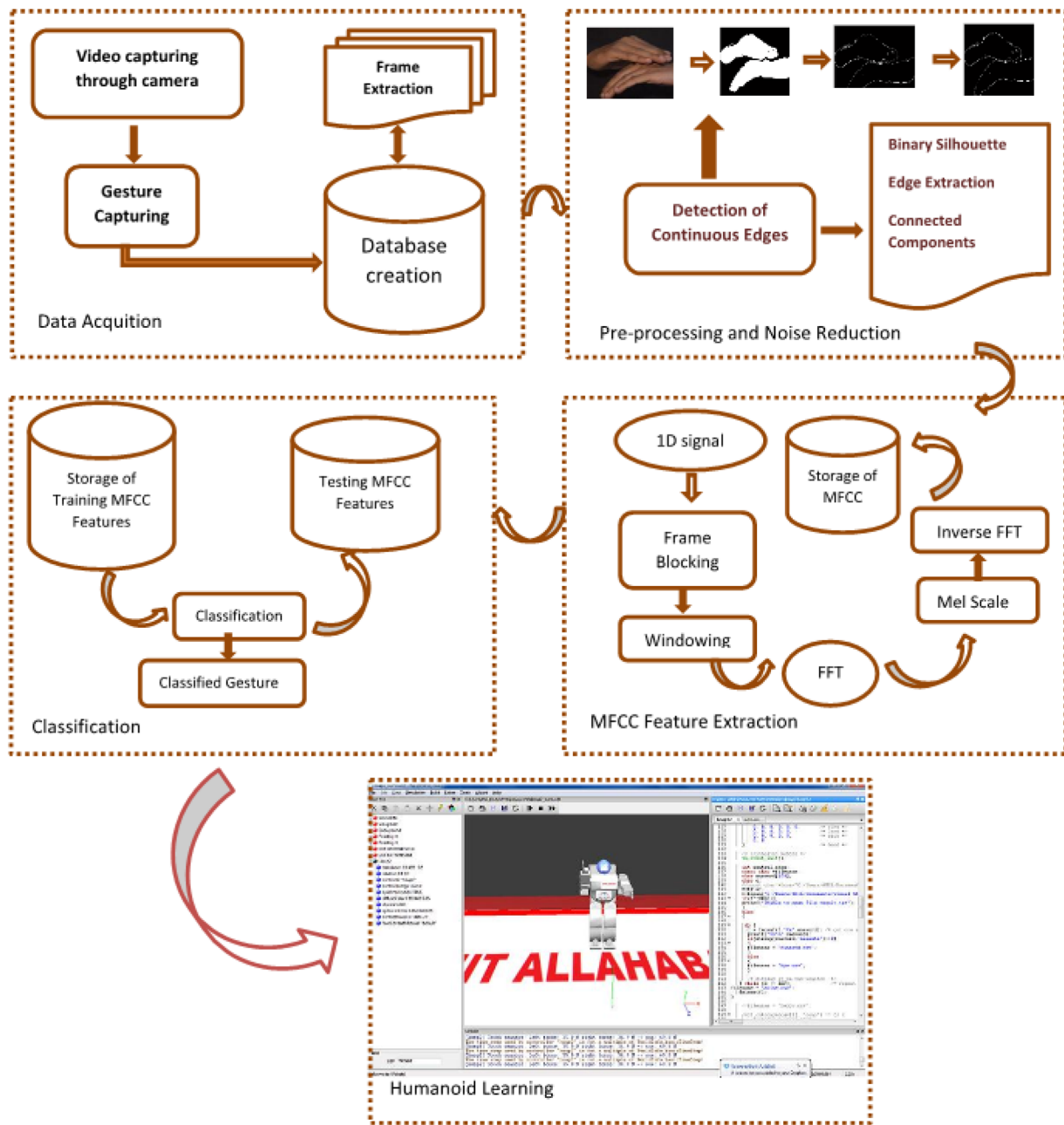














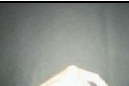



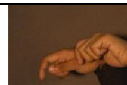


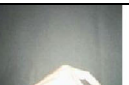

Fig. 1 Proposed framework

every degree of hand angle needed for correct classification that makes it expensive as time aspects. Jafreezal and Fatimah [13] jointly applied a MFCC technique for recognition of ASL database. They evaluate this technique with 10 gestures and calculate recognition rate and get 95 % accuracy. They tested the results with very small amount of dataset which is not sufficient for experimental purpose. This technique is computationally extensive in nature.

From above literature we found that gesture segmentation and appropriate feature extraction are the major issues

in any gesture recognition technique. In segmentation, hand extraction is the biggest challenge. After hand extraction feature extraction is the next biggest challenge for gesture recognition. Which we have solved using a combination of wavelet descriptor and MFCC based ISL gesture recognition technique. The combinations of these two methodologies are mainly used in speech processing [14, 15] and it would be very prominent in the field of speech processing because of its inherent properties makes it unique. In palm print recognition [16] a similar type of methodology has been applied. Here features are extracted

**Table 1** Database of static and dynamic gestures in different illumination condition

Gesture Name	Above (Dynamic)	Across (Dynamic)	Afraid (Dynamic)	Alone (Dynamic)	Bag (Static)	Yes (Dynamic)	All Gone (Static)
Start Frame							
Middle Frame							
End Frame							

from DWT and features extracted from MFCC are fused. Than this combined features are processed for classification and it will give satisfactory results. Therefore we have inspired to apply this technique for ISL gesture recognition and this technique will also perform well on ISL gesture recognition.

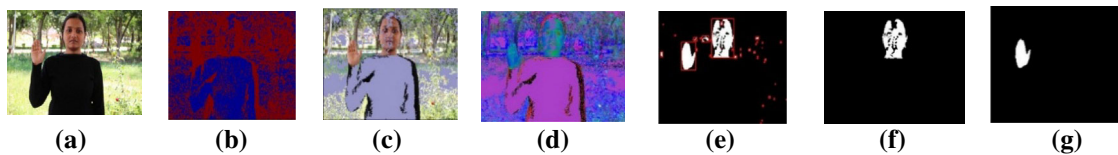
### 3 Proposed framework

In this paper we have proposed a gesture recognition method where MFCC features have been extracted by processing of transformed images by wavelet descriptors. The proposed framework for ISL gesture recognition is shown in Fig. 1. In Fig. 1 we have seen that the overall framework is divided into five modules: data acquisition where data collection has been done through webcam and Sony handycam then each video is divided into sequence of frames. These frames are further go to the pre-processing phase in which hand extraction, silhouette image formation, and boundary point extraction is performed. In this module silhouette images are converted into 1D signal and then wavelet descriptors have been applied for finding boundary points which are moment invariant. These features are further processed for finding MFCC coefficients which is done in feature extraction module. MFCC feature extraction technique is widely used in speech processing here we explored MFCC in hand gesture recognition in combination with WD and see the performance on image frames. Due to spectral envelop property of MFCC, the proposed method gives high recognition rate with less processing time. Further probes are classified using SVM and KNN in the classification phase. In the last module these classified gestures are performed by HOAP-2 robot

using Webots platform. All the modules of proposed framework are described in detailed in subsequent sections.

- (A) Database creation: Sir William Tomkins mentioned 100 signs posture and guaranteed about them of being true Indian Signs in his book Universal Indian Sign Language [17]. We have created database of 42 ISL static as well as dynamic gestures among 100 signs by using simple Logitech HD 720p camera shown in Table 1. The database contains both dynamic gestures as well as static gestures recorded in different light conditions (yellow light, red light, white light, etc.) and different background like white paper, paper with different symbols, red background, etc. All the database is created with black full sleeves dress. We have captured full body dataset where portion of both the hands come. In our database we have considered both types of gestures containing single hand as well as both hands.
- (B) Pre-processing: Gesture videos are first divided into sequence of frames of size  $(m, n)$  represented as  $I_i(m, n)$ , where  $i$  is the number of frames in each video. Hand region are extracted from whole image frame shown in Fig. 2.

These images are further converted into binary images using threshold  $T$ . Binarisation of an image is necessary for extracting good edges which are continuous in nature. Contour obtained are distorted at some points therefore binarisation of an image is necessary. This binarisation of an image is performed using Otsu's thresholding method where threshold is chosen in such a manner that intra-class variance of black and white pixel is minimum. Then after WD is applied for reducing dimension of an



**Fig. 2** Pre-processing steps of each RGB frame. **a** RGB image, **b** Background, **c** Forground, **d** HSV, **e** Bounding box, **f** Face extraction, **g** hand image

image and extracting moment invariant features which does not contain noise.

*Wavelet descriptor* In WD first DWT [18, 19] is applied up to the third level of decomposition of a binary image frames. The decomposition of image matrix in each level is done by row wise multiplication of images with wavelet filter followed by column wise multiplication and getting the four components of single image [LL, LH, HL and HH] where LL is the approximate coefficients which discards all high frequency coefficients and LH, HL, HH is the high frequency sub bands where LH contains horizontal components, HL contains vertical components and HH has a diagonal components of original image. The decomposition is performed as:

$$LL = D_{w_0}(j, p, q) = \frac{1}{\sqrt{PQ}} \sum_{x=0}^{P-1} \sum_{y=0}^{Q-1} f(x, y) \phi_{j,p,q}(x, y) \tag{1}$$

$$i = LH, HL, HH = D_{w_i}(j, p, q) = \frac{1}{\sqrt{PQ}} \sum_{x=0}^{P-1} \sum_{y=0}^{Q-1} f(x, y) \Omega_{j,p,q}^i(x, y) \tag{2}$$

where  $\phi$  and  $\Omega$  is a scaling and wavelet coefficients, P and Q is the dimension of image matrix, j is the level of decomposition, f(x,y) is the image matrix obtained after binarisation and p, q is the row and column of image matrix.

Suppose Image matrix of size (4, 4)

$$I = \begin{bmatrix} c11 & c12 & c13 & c14 \\ c21 & c22 & c23 & c24 \\ c31 & c32 & c33 & c34 \\ c41 & c42 & c43 & c44 \end{bmatrix}$$

Coefficients of Daubachies wavelet filter H is defined as: Low pass:  $\frac{1}{4\sqrt{2}}(3 - \sqrt{3}, 3 + \sqrt{3}, 1 + \sqrt{3}, 1 - \sqrt{3})$  High pass:  $\frac{1}{4\sqrt{2}}(1 - \sqrt{3}, -1 - \sqrt{3}, 3 + \sqrt{3}, -3 + \sqrt{3})$

$$H = \begin{bmatrix} W11 & W12 & W13 & W14 \\ W21 & W22 & W23 & W24 \\ W31 & W32 & W33 & W34 \\ W41 & W42 & W43 & W44 \end{bmatrix}$$

where W11, W12 ... are wavelet filter bank coefficients.

$$I_r = I \times H \tag{3}$$

$$a_c^1 | d_c^1 \approx \frac{1}{4\sqrt{2}} \begin{bmatrix} r11, r12, r13, r14 \\ r21, r22, r23, r24 \\ r31, r32, r33, r34 \\ r41, r42, r43, r44 \end{bmatrix} \tag{4}$$

where  $1/4\sqrt{2}$  is a normalizing factor. It normalizes the approximate  $a_c$  and detail  $d_c$  coefficients such that

$$\|a_c\| = \|d_c\| = 1.$$

$$I_c = H' \times I_r. \tag{5}$$

$$a_c^1 | d_c^1 = \frac{1}{4\sqrt{2}} \begin{bmatrix} \left( \begin{matrix} D11, D12 \\ D21, D22 \end{matrix} \right) | \left( \begin{matrix} D13, D14 \\ D23, D24 \end{matrix} \right) \\ \left( \begin{matrix} D31, D32 \\ D41, D42 \end{matrix} \right) | \left( \begin{matrix} D33, D34 \\ D43, D44 \end{matrix} \right) \end{bmatrix} \tag{6}$$

where  $I_c$  is an orthogonal matrix which preserves magnitude and angle of any vector is belongs to  $\mathbb{R}^n$ . Prove of these properties are given below:

Suppose  $z$  is a one dimensional vector of  $\mathbb{R}^n$  and  $Ic$  is a orthogonal matrix obtained after decomposition of an image  $I$  then proof that

$$\begin{aligned} \|zIc\| &= \|z\|, \\ \|Icz\|^2 &= (Icz)^T(Icz) \\ &= z^T Ic^T Ic z \\ &= z^T I z \\ &= z^T z \\ &= \|z\|^2 \end{aligned}$$

This shows that  $Ic$  preserves magnitude of any vector  $z \in \mathbb{R}^n$  i.e. it is scale invariant.

Suppose  $z_1$  and  $z_2$  is two vector of  $\mathbb{R}^n$  and  $Ic$  is a decomposed matrix obtained after DWT decomposition of an image  $I$  then proof that angle between  $z_1$  and  $z_2$  is always same.

Let us assume that angle between  $z_1$  and  $z_2$  is  $\phi$

$$\cos \phi = \frac{z_1 \cdot z_2}{\|z_1\| \|z_2\|}$$

If  $Ic$  is an orthogonal matrix then angle between  $Icz_1$  and  $Icz_2$  is

$$\begin{aligned} \cos \delta &= \frac{Icz_1 \cdot Icz_2}{\|Icz_1\| \|Icz_2\|} \\ &= \frac{(Icz_1)^T (Icz_2)}{\|z_1\| \|z_2\|} \\ &= \frac{z_1^T Ic^T Ic z_2}{\|z_1\| \|z_2\|} \\ &= \frac{z_1^T z_2}{\|z_1\| \|z_2\|} \\ &= \frac{z_1 \cdot z_2}{\|z_1\| \|z_2\|} \\ &= \cos \phi \end{aligned}$$

This proves that in WD angle between two vectors are always same i.e. phase invariant.

These two properties of WD shows that the angle and amplitude of original image does not change when it converted into the transformed image means image will be less distorted after transformation.

The Eq. 4 shows the final matrix obtained after 1st level decomposition which has 4 components termed as LL, LH, HL and HH. Each component of the matrix is of size  $(m/2, n/2)$ . In second level decomposition LL component is further decomposed into four parts because it contains maximum information with minimum noise similar to original image. Whereas LH, HL and HH are high frequency coefficients having low signal to noise ratio. This process continuous up to third level of decomposition and finally we get four coefficients FF, FV, VF and VV. After decomposition process contour of an image is calculated by any known contour detection method. Then moment invariant features are deliberated by converting 2-D contour image  $[G(x, y)]$  into 1 dimension. For this conversion, 2-D image  $G(x, y)$  in  $x$ - $y$  plane is converted into  $r$ - $\theta$  plane  $G(r, \theta)$  described as:  $x = r \cos \theta$  and  $y = r \sin \theta$ .

$$G_{ab} = \iint G(r, \theta) g_a(r) e^{ib\theta} r dr d\theta \quad (7)$$

where  $r$  is the radius of the circle,  $\theta$  is the orientation angle,  $G_{ab}$  is the moment of hand,  $g_a(r)$  is a radial basis function and  $a, b$  are constants. In case of wavelet descriptor  $g_a(r)$  has been treated as a wavelet basis function and replaced with  $\vartheta^{p,q}(r) = \frac{1}{\sqrt{p}} \vartheta\left(\frac{r-q}{p}\right)$   $p$  and  $q$  are the dilation and shifting parameter.

Now convert 2D image into 1D form for reducing feature extraction problem and increasing performance quality. We choose cubic B-spline (Gaussian approximation) function as a mother wavelet define as:

$$\begin{aligned} \vartheta(r) &= \frac{4p^{n+1}}{\sqrt{2\pi(n+1)}} \sigma_y \cos(2\pi g_0(2r-1)) \\ &\quad \times \exp\left(-\frac{(2r-1)^2}{2\sigma_y^2(n+1)}\right) \end{aligned}$$

Analyzing a moments in shape of an image the values of dilation and shifting parameter  $p$  and  $q$  are chosen to be discrete expressed as:

$p = p_0^m, \quad m \text{ is an integer}$   
 and  $q = nq_0p_0^m, \quad n \text{ is an integer}$

$p_0 > 1$  or  $p_0 < 1$  and  $q_0 > 0$ . These constraints has been considered so that  $\vartheta\left(\frac{r-q}{p}\right)$  covers the complete shape of gesture. Here we considered circle for representing shape of an image whereas  $(r \leq 1)$ . The values we choose is  $(p_0 \text{ and } q_0 = 0.5)$ . Then the wavelet basis function  $\vartheta^{p,q}(r)$  has been modified as:

$$\vartheta_{m,n}(r) = 2^{\frac{m}{2}}\vartheta(2^m r - 0.5n)$$

$\vartheta_{m,n}(r)$  defines for any orientation along radial axis  $r$ . It is used for finding the local as well as global features of hand by varying the values of  $m, n$ . After that we define moment invariant wavelet feature vector as:

$$\left\| G_{m,n,b}^{wavelet} \right\| = \left\| \int f_b(r) \times \vartheta_{m,n}(r) r dr \right\| \tag{8}$$

Comparing Eqs. 7 and 8 we get  $g_a(r) = \vartheta_{m,n}(r)$  and  $f_b(r) = \int G(r, \theta) e^{jb\theta} d\theta$  shows the  $b$ th frequency feature of image  $G(r, \theta)$  in  $r-\theta$  plane where  $0 \leq \theta \leq 2\pi$ .

$\left\| G_{m,n,b}^{wavelet} \right\|$  is the wavelet transform of  $f_b(r)r$ . It analyses the signal in both time domain as well as frequency domain and extracts features which are locally descriptive in nature. Features shown in Eq. 8 are moment invariant for each gesture with feature vector  $\left\| G_{m,n,b}^{wavelet} \right\|$ . Where  $m = 0, 1, 2, 3$  and  $n = 0, 1 \dots 2m + 1$ .

$\left\| G_{m,n,b}^{wavelet} \right\|$  is the generalization of moment  $f_b(r)$  at  $m$ th scale level and  $n$ th sift position.

In WD  $\left\| G_{m,n,b}^{wavelet} \right\|$  represents the moment invariant property of image  $I$ , if this image is rotated by an angle  $\alpha$  then moment invariant property

$\left\| G_{m,n,b}^{waveletrotated} \right\|$  defined as:

$$G_{m,n,b}^{waveletrotated} = G_{m,n,b}^{wavelet} e^{jbx}$$

Since

$$\begin{aligned} \left\| G_{m,n,b}^{waveletrotated} \right\| &= \sqrt{\left(\left\| G_{m,n,b}^{waveletrotated} \right\|\right) \left(\left\| G_{m,n,b}^{waveletrotated} \right\|\right)^*} \\ &= \sqrt{\left(\left\| G_{m,n,b}^{wavelet} e^{jbx} \right\|\right) \left(\left\| G_{m,n,b}^{wavelet} e^{-jbx} \right\|\right)} \\ &= \left\| G_{m,n,b}^{wavelet} \right\| \end{aligned}$$

shows the moment invariant properties of WD.

(C) MFCC feature extraction: After that MFCC coefficients of each image has been calculated. Each image

behaves like a signal. On this signal, MFCC feature extraction technique is applied for calculating the spectral envelope of each frame. The spectral envelop properties of MFCC feature [20] provides robustness towards background noise as here in case of getting false edges of hands. Spectral envelope is a property of an image which gives the knowledge about image intensity over frequency by providing the smooth curve or regular curve having no discontinuity in frequency domain. It is a curve in the frequency amplitude plane, which tightly connects all the points of the magnitude spectrum, linking the peaks. These peaks represents the highest intensity value of the image, carries maximum information. It derived from a Fourier magnitude spectrum.

With the help of spectral envelop graph shown in Fig. 5 we see the nature of the signal how tightly the curve will cover the signal. The steps of MFCC features and spectral envelop calculation are:

- (a) In noisy background condition there is a large variation in consecutive pixel values. Thus by assuming it statistically stationary in short period of time, the column vector block broken up into small sections called frames which is having  $N$  samples/frame. These frames are separated by  $M$  samples ( $N > M$ ) therefore overlapping is done by  $N-M$  samples.
- (b) For removing the side ripples (spectral distortion) and maintaining continuity in the signal windowing is performed using hamming window function. It provides lower side lobe and narrower main lobe in the image frame. Which is expressed as:

$$\begin{aligned} hm(n) &= 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \\ 0 \leq n \leq N-1 \end{aligned} \tag{9}$$

The resultant signal  $y(n)$  is

$$y(n) = x(n) \times hm(n). \tag{10}$$

where  $hm(n)$  is the window signal,  $x(n)$  is the wavelet coefficients.

- (c) After that Fast Fourier Transform (FFT) is applied on resultant signal  $y(n)$  which converts time domain image frame into frequency domain image frame represented as  $S_k$ .

$$S_k = \sum_{n=0}^{N-1} y(n) \times e^{-\frac{j2\pi kn}{N}} \tag{11}$$

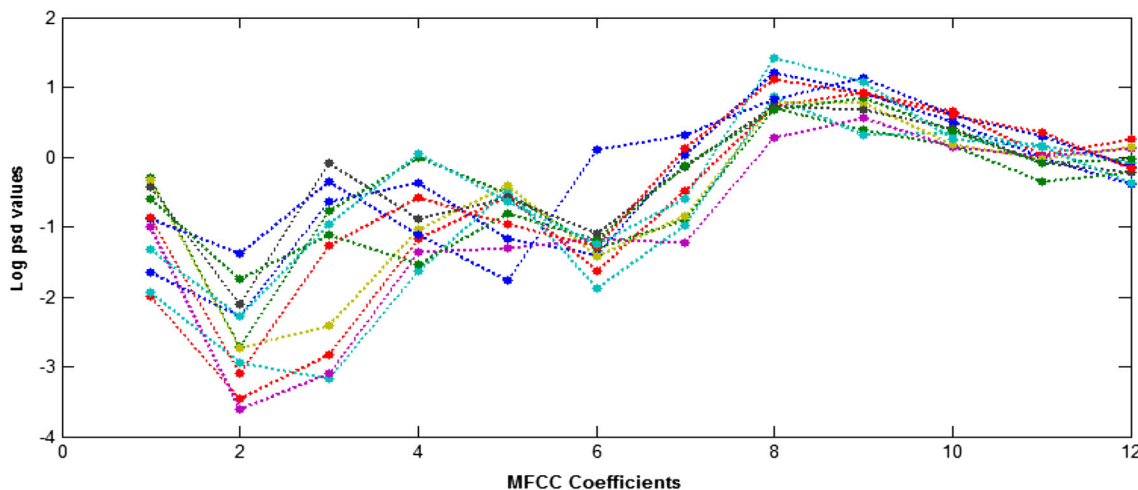


Fig. 3 MFCC with DWT plot for gesture above

- where  $k = 0, 1, 2 \dots N - 1$ .
- (d) Periodogram based power spectral density ( $P_i(k)$ ) is calculated by taking absolute value of complex Fourier transform and square the result as.

$$P_i(k) = \frac{1}{N} |S_i(k)|^2 \tag{12}$$

where positive frequencies  $0 \leq f < fs/2$  correspond to values  $0 \leq n \leq N/2 - 1$ , while negative frequencies  $-fs/2 < f < 0$  correspond to values  $N/2 + 1 \leq n \leq N - 1$ .

- (e) The irrelevant frequencies are mixed up with very closely spaced relevant frequencies and this effect is more effective with frequency increase. Thus to get a clear idea about exact energy amplitude at various frequencies, PSD coefficients are binned and correlated with each filter from a Mel filter bank. Mel is a nonlinear scale represented as-

$$\text{Mel Frequency} = 2595 \times \log\left(1 + \frac{fs}{700}\right) \tag{13}$$

where  $fs$  is the frequency range used for generating filter bank is designed in next step.

- (f) This Mel scale is used by filter bank as their Centre frequency follows this Mel frequency and thus the filters near the lower frequencies are having the narrow bandwidth and as the frequency increases the filters width increases.

$$H(k,b) = \begin{cases} 0, & \text{if } f(k) < f_c(b-1) \\ f(k) - \frac{f_c(b-1)}{f_c(b) - f_c(b-1)}, & \text{if } f_c(b-1) \leq f(k) < f_c(b) \\ f(k) - \frac{f_c(b+1) - f_c(b)}{f_c(b) - f_c(b-1)}, & \text{if } f_c(b) \leq f(k) < f_c(b+1) \\ 0, & \text{if } f(k) > f_c(b+1) \end{cases} \tag{14}$$

where  $f_c(b)$  is the central frequency of the filter,  $b$  is the number of filter used in filter bank and  $f(k)$  is the Mel frequency.

- (g) Find the log energy output of each of the Mel frequencies.

$$S(b) = \sum_{k=0}^{N-1} H(k,b) \times P \tag{15}$$

where  $b = 1, 2, \dots, m$  and  $m$  is the number of filter and  $P$  is PSD matrix

- (h) Coefficients  $me_1, me_2, \dots$  are generated by applying Discrete Cosine Transform (DCT) on these Mel frequencies.

$$me = \text{DCT}(s(b)) \tag{16}$$

where  $me$  is the number of Cepstral coefficients. These coefficients are saved as feature vector shown in Figs. 3, 4.

In Figs. 3 and 4 X axis represents the number of MFCC coefficients and Y axis represents log magnitude value corresponding to that coefficients. The graph shows MFCC plot for each blocked frames of one image corresponding to their mentioned gesture. 12 MFCC coefficients are taken for each image frame. Here the single image of one gesture is blocked into 11 frames and graph shows MFCC plot for those 11 frames.



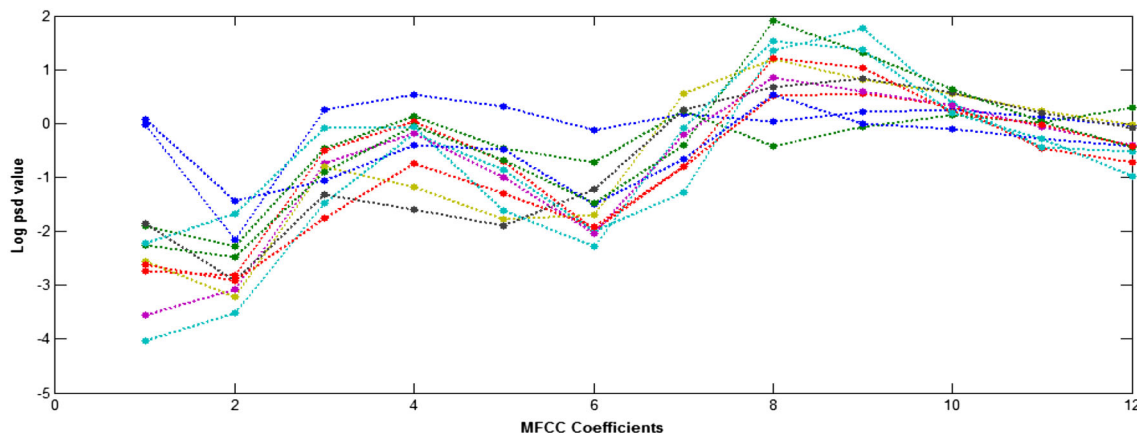


Fig. 4 MFCC with DWT plot for gesture across

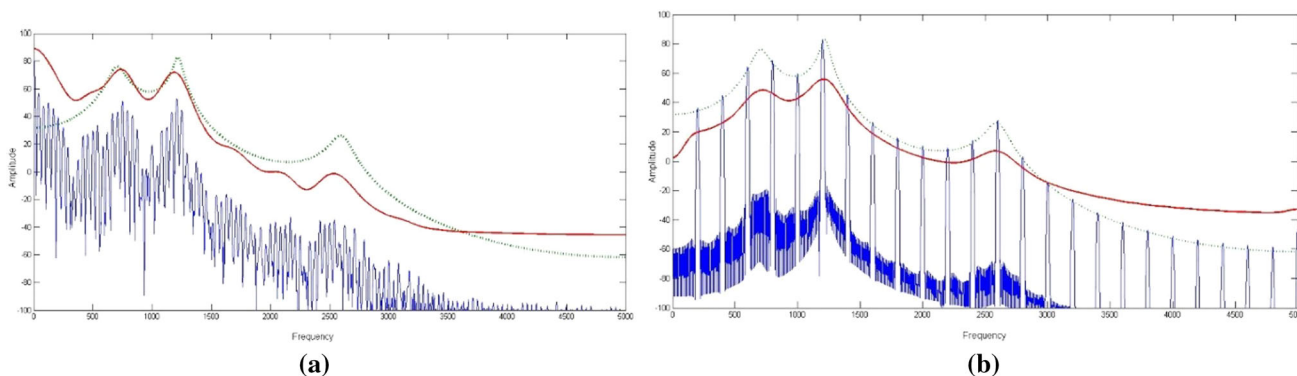


Fig. 5 Spectral envelop of a MFCC features, b WD with MFCC features

The spectral envelop in MFCC features are calculated as:

Spectral envelop of any signal has been calculated by applying FFT function on cepstral coefficients obtained in the last step of MFCC feature extraction technique.

1. Number of bins present in any spectral envelop of the signal has been taken by dividing the range of frequency  $f_t$  up to Nyquist frequency  $f_s/2$ . Where  $f_s$  is the sampling rate. Here we divide frequency into equal parts up to  $f_s/2$ .

$$f_t = t \frac{f_s/2}{n}, \quad t = 1 \dots n \tag{17}$$

2. Calculate the angular frequency  $w_t = f_t \frac{2\pi}{f_s}$  where  $w_t$  is the angular frequency.
3. Finally we calculate the spectral envelop  $\vartheta_t$  of frequency  $f_t$  as:

$$\vartheta_t = \exp\left(\sum_{i=1}^n m_{e_i} \times \cos iw_t\right) \tag{18}$$

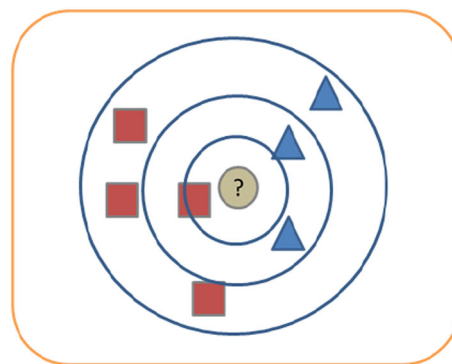


Fig. 6 KNN classification example for  $k = 1, 3, 5$  or  $7$

where  $m_{e_i}$  is the cepstral coefficient. Spectral envelop of above gesture is shown in Fig. 5. From this graph we see that number of bins of frequency in which the spectral envelop lies is 10 and the red solid line shows the spectral envelop of the signal which is curved generated by combining the peaks of the signal. From Fig. 5a, b we see the nature of the

curve of spectral envelop and found that in case of simple MFCC the curve is not tightly bound. When signal becomes invisible it becomes straight line but in case of WD with MFCC the spectral envelop curve is more tightly bound containing peaks of the signal which will properly discriminate the changes in the signal.

(D) Classification: MFCC coefficients of various gestures are classified using K nearest neighbour (KNN) [21] and support vector machine (SVM) [22].

(a) K nearest neighbour (KNN):

It is a simple classifier used for classification of an unknown gesture. It is based on nearest neighbour method presents nearest to the unknown gesture shown in Fig. 6. In KNN initially assume the value of k then measured distance between all testing and training vectors of gestures using Euclidean distance. Then select k closest vectors whose Euclidean distance is minimum. Calculate the vote and assign the labels according to the majority vote the class gets. Here proposed algorithm is tested on k = 1, 3, 5 and 7. In Fig. 6 we have seen that for k = 1, unlabeled data assigned to square class, for k = 3, unlabeled data assigned to triangle class and k = 7, unlabeled data assigned to square class.

(b) Support vector machine (SVM): It is a classifier, handles nonlinearity present in features by transforming it into higher dimension. Here we use linear kernel function for discriminating different classes.

$$y_i(w \cdot tr_i - k) \geq 0 \tag{19}$$

$tr_i$  is the training class where i represents the class number,  $\cdot$  is the dot product, w is the normal vector  $\|w\| = 1$ ,  $y_i$  is the class label for two class problem the value of  $y_i$  is 1 and  $-1$ . k is the constant. Constant  $\frac{k}{\|w\|}$  sets the width of the hyper plane between different classes. If  $k = 2$  then there will be two classes. For multiclass problem SVM uses multiple binary classifiers. Let we have a t classes then generate t binary classifiers  $f_1, f_2, \dots, f_t$ . Each classifier is trained with one class from rest of the classes. Combine all the binary classifier to get a classification for multiclass SVM expressed as:

$$\operatorname{argmax}_{i=1, \dots, t} F^i(x)$$

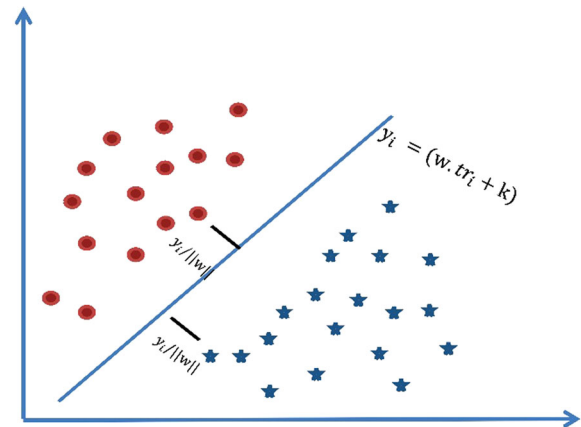


Fig. 7 SVM classification

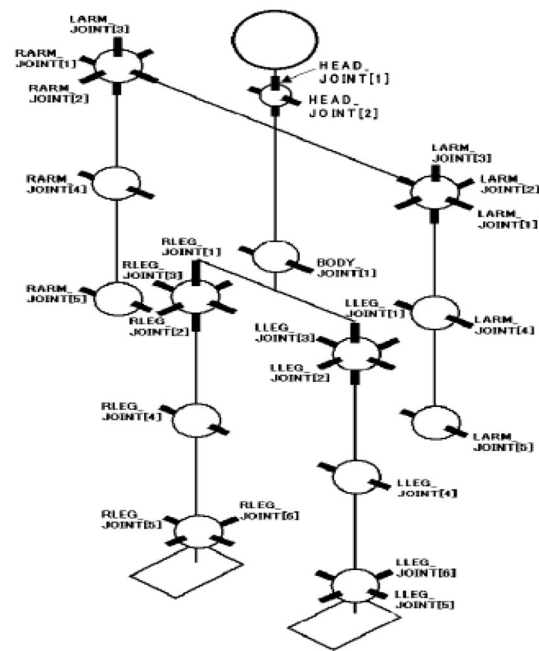


Fig. 8 Architecture of HOAP-2 Robot [3]

where

$$F^i(x) = \sum_{n=1}^m y_n \alpha_n^i K(te, tr_n) + k^n \tag{20}$$

$\alpha_n^i$  is the lagrangian multiplier, K is the linear kernel function defined in Eq. 20. For classifying an unknown test data  $te$  within t classes then  $\frac{t-(t-1)}{2}$  binary classifiers have been applied and count the number of times the test data  $te$  belongs to a particular class. Those classes having maximum number of class labels, the test data  $te$  belong to that class. The binary SVM classification is shown in Fig. 7. Here

**Table 2** Allowable range of joint angle for humanoid robot HOAP-2

Joint name	Min angle (degree)	Min counter (decimal)	Max angle (degree)	Max counter (decimal)	CSV columns
RLEG_Joint [1]	−91	−19,019	31	6479	C
RLEG_Joint [2]	−31	−6479	21	4389	D
RLEG_Joint [3]	−82	17,138	71	−14,839	E
RLEG_Joint [4]	−1	−209	130	27170	F
RLEG_Joint [5]	−61	−12,749	61	12,749	G
RLEG_Joint [6]	−25	5225	25	−5225	H
RARM_Joint [1]	−91	−19,019	151	31,559	I
RARM_Joint [2]	−96	−20,064	1	209	J
RARM_Joint [3]	−91	19019	91	−19,019	K
RARM_Joint [4]	−115	24,035	1	−209	L
LLEG_Joint [1]	−31	−6479	91	19,019	M
LLEG_Joint [2]	−21	−4389	31	6479	N
LLEG_Joint [3]	−82	−17,138	71	14,839	O
LLEG_Joint [4]	−1	209	130	−27,170	P
LLEG_Joint [5]	−61	12,749	61	−12,749	Q
LLEG_Joint [6]	−25	5225	25	−5225	R
LARM_Joint [1]	−91	19,019	151	−31,559	S
LARM_Joint [2]	−1	−209	96	20064	T
LARM_Joint [3]	−91	19019	91	−19,019	U
LARM_Joint [4]	−115	−24,035	1	209	V
Body_Joint [1]	−1	209	90	−18,810	W
RARM_Joint [5]	−60	−	60	−	X
LARM_Joint [5]	−60	−	60	−	Y
Head_Joint [1]	−60	−	6	−	Z
Head_Joint [2]	−15	−	60	−	AA

we use “multisvm” function of Matlab for classifying a test vector of an unknown gesture w.r.t. known gesture (one verses all strategy).

- (E) Humanoid learning: After classification process humanoid learning is performed by HOAP-2 robot in Webots robotic simulation software. It has various types of robots like social robots, industrial robots, etc. It is a software platform for establishing an interaction between human and robot. In this paper gestures are used for establishing interaction which a human performs and then all these gestures are performed by the humanoid robot HOAP-2. HOAP-2 is a humanoid open architecture platform [3] having 25 degrees of freedom (DOF) (6 DOF on foot  $\times$  2, 4 DOF on arm  $\times$  2, 1 DOF on waist, 1 DOF on hand  $\times$  2, 2 DOF on neck). The architecture of HOAP-2 robot is shown in Fig. 8. It is a social robot used for human robot interaction either by using speech or by using gestures (head motions, hand motions) [4]. To implement on HOAP-2 we need the joint angle data of each joint of the human hand. When the user performs the gesture, hand of a person moves horizontally as well as vertically, or ups or down. At each instant of time the joint angle value

changes with respect to initial coordinate frame. These joint angle values are captured either by using data glove sensor or using atan2 function than these joint angle values are interpolated for smooth motion of HOAP joint-2 angles. Finally these values are passed to the.csv file which are uploaded into robot controller programme. Here we linked Webots software to the MATLAB where gestures are classified. The range of each joint of HOAP-2 is shown in Table 2.

From all the joints shown in Table 2 we used only 10 joints (RARM Joint 1, 2, 3, 4, 5 and LARM Joint 1, 2, 3, 4, 5) and corresponding to their column values present in.csv file for performing any gestures, otherwise all other joint values keep it constant. Position of joint of humanoid robot at a particular position is defined using joint angle values expressed as:

$$Po = A \times \varphi$$

where Po is the position value,  $\varphi$  is the joint angle in degree and A is the change of coefficients in pulses/deg. In HOAP-2 robot architecture the moment of joint angle values are performed in pulses because each joint having its own motor. Performing 1 degree of motion the motor of

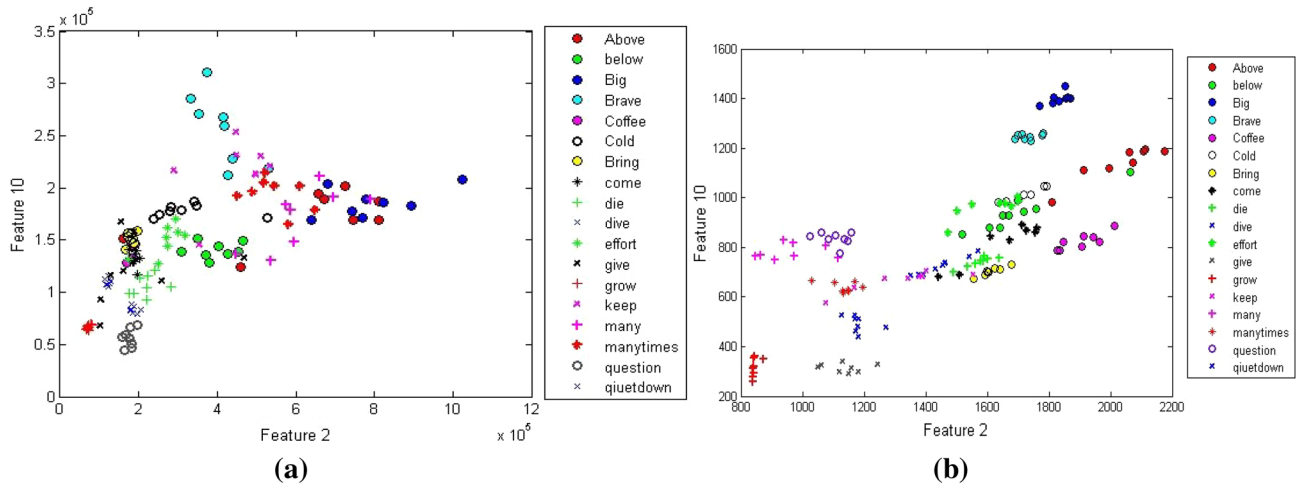


Fig. 9 Feature space plot between feature 2 vs feature 10 a MFCC features, b WD with MFCC features

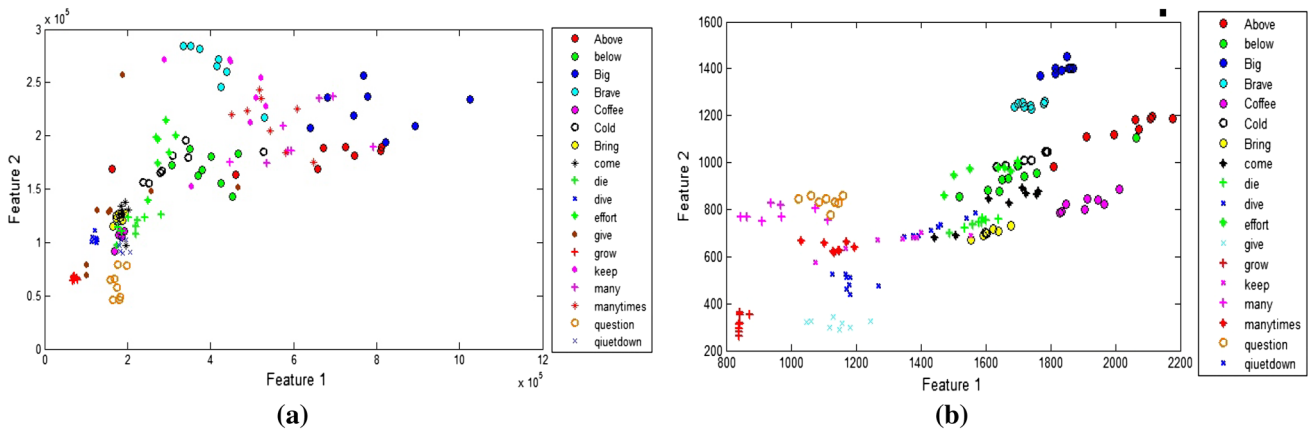


Fig. 10 Feature space plot between feature 1 and feature 2 a MFCC features, b WD with MFCC features

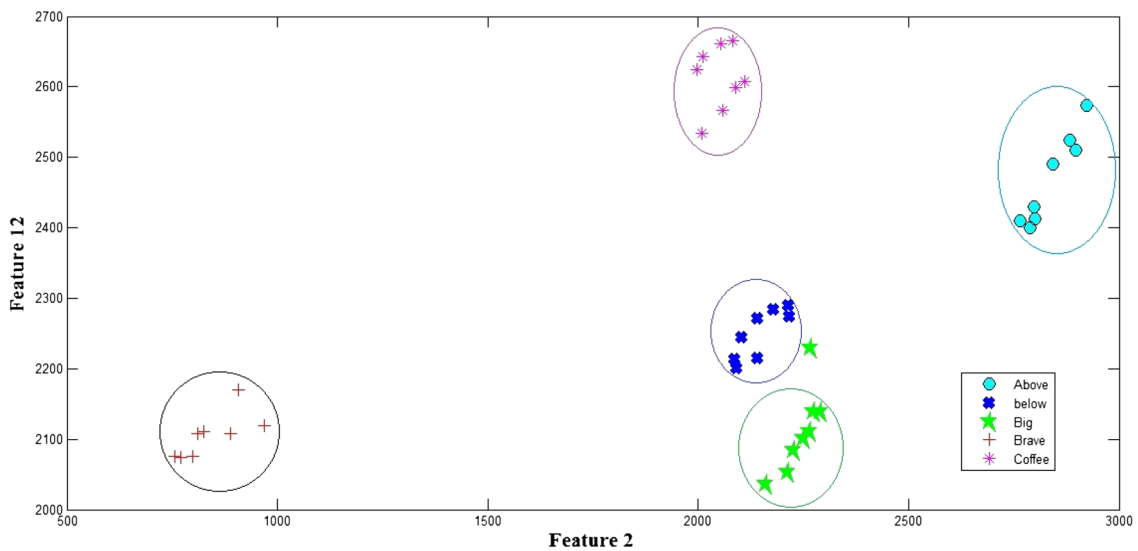


Fig. 11 Feature space plot between feature 2 and feature 12 (WD with MFCC features)

each joint requires 209 pulses. In csv file every column has the position value of each joint with some negative value and some positive value i.e. joints has been move sometimes in positive direction and sometimes in negative direction which is calculated by multiplying number of degree of motion to the pulse value.

### 4 Experimental results and analysis

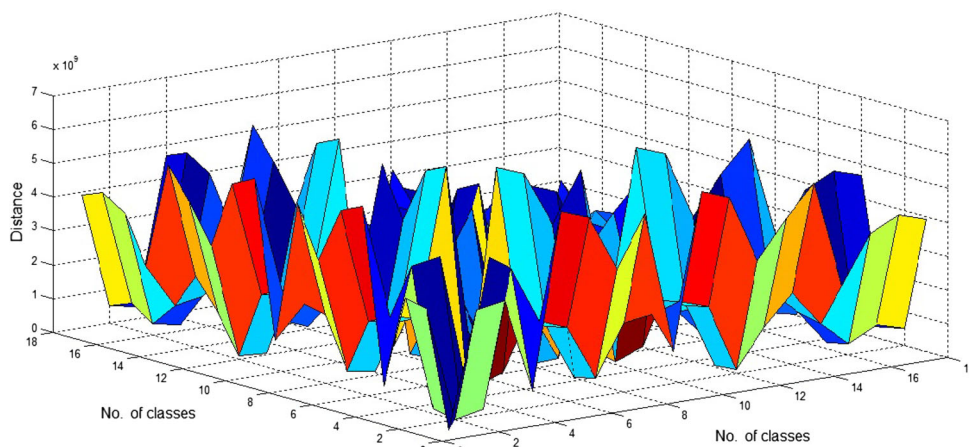
Data set of 42 ISL gestures like big, bring, below etc. (23 static and 19 dynamic) are created in three light conditions and different background conditions. In this paper five videos of each gesture are taken as a training gesture and three videos are taken as a testing set where each video having 150 frames. For training we take five persons dataset and for testing seven person’s dataset where six are males and six are females (all are Indian Institute of Information Technology, Allahabad students). Matlab Image Acquisition Toolbox has been used for capturing

ISL video data. The toolbox is configured as: Colour space: RGB (320\*240), Triggering mode: immediate, Number of triggers: 10, Frame rate: 30 frames per second.

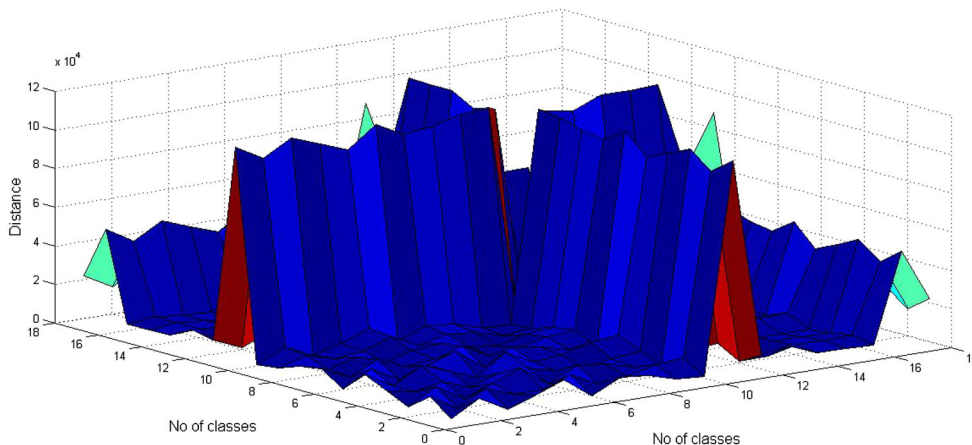
After data acquisition and preprocessing the proposed WD with MFCC features have been extracted and demonstrated through feature space plot [27, 28] which is shown in Figs. 9, 10 and 11.

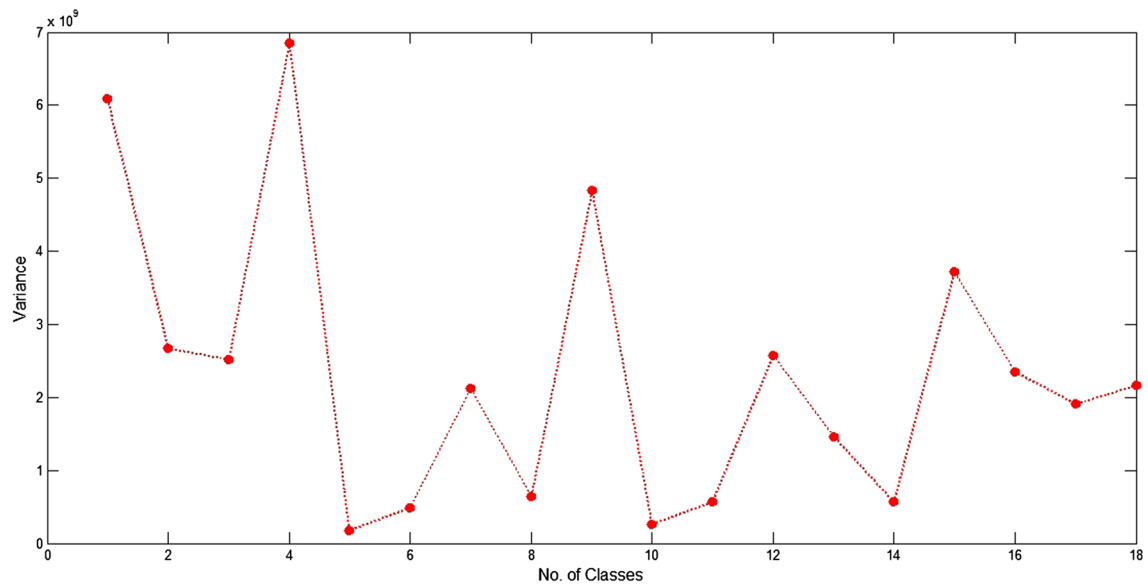
We have compared two feature extraction techniques one is MFCC [25] and second one is WD with MFCC feature extraction technique using feature space plot [27] shown in Figs. 9 and 10 where graph is created between feature 1 vs feature 2 and feature 2 vs feature 10. Here I have shown best three feature space plots. From all five graphs we found that the proposed technique has more discriminative capability to separate all 18 classes’ then simple MFCC feature extraction technique which means that the overlapping between the MFCC features are more than the WD with MFCC features. This overlapping shows that the nature of data, if overlapping is more than the dataset having maximum nonlinearity and vice versa.

**Fig. 12** Distance plot for MFCC features

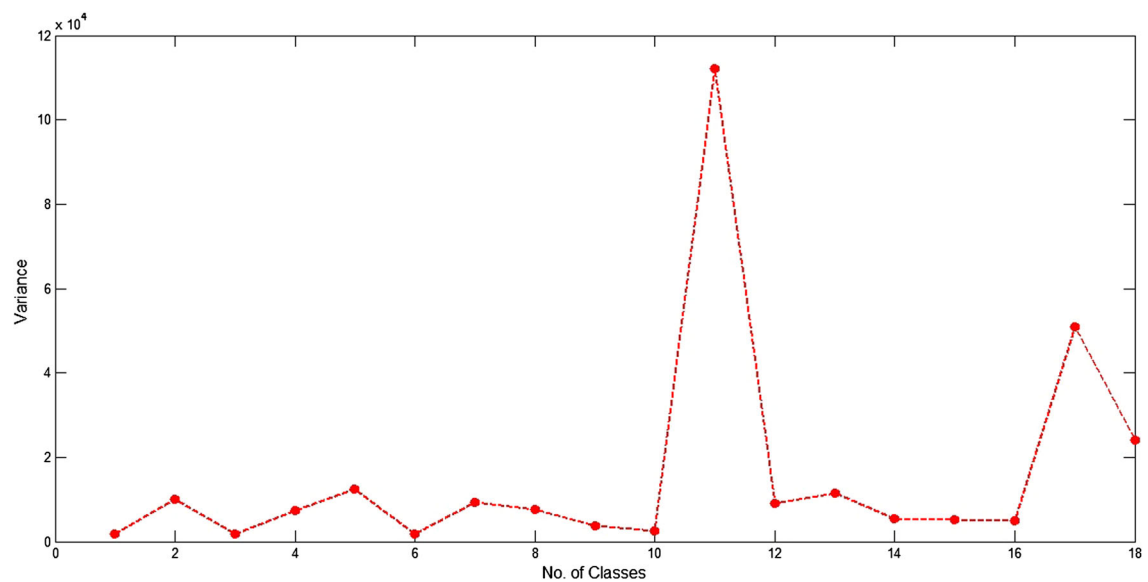


**Fig. 13** Distance plot for WD with MFCC features





**Fig. 14** Variance plot for MFCC features



**Fig. 15** Variance plot for WD with MFCC features

Figure 11 shows the feature space plot of proposed methodology with five gestures. Figure 11 perfectly shows the discriminative ability of the proposed technique rather than the other existing technique. These MFCC features are statistical in nature which has been shown by calculating between class distance matrix and within class matrix shown in Figs. 12, 13, 14 and 15.

Graph shown in Figs. 12, 13, 14 and 15 represents the variance of within class matrix and distances of between class matrixes. In Figs. 14 and 15 we see that mean of the variance of each class lies below a threshold value 2db for maximum

classes in case of WD with MFCC features in comparison to simple MFCC features where variance of the classes varies frequently which shows non linearity in dataset. From graph shown in Figs. 12 and 13 we see that the distance between classes is high in WD with MFCC then MFCC. In between class matrix the Euclidean distance between same class is zero and other classes have some value but in Fig. 12 we see that in some cases the distance between same class is also not zero where as in WD with MFCC the distance between same class is always zero which proves that the proposed method has more discriminative power then other methods (MFCC [25]).

**Table 3** Frame based classification results for dynamic gestures using SVM and KNN

Class	For KNN (K = 3)		SVM (linear kernel)	
	Frame misclassified	Accuracy (%)	Frame misclassified	Accuracy (%)
Above	60	86	12	97
Add	56	88	2	99.55
Big	102	77	10	97.77
Below	60	86	5	98.88
Bring	50	89	3	99
Coffee	115	77	14	97
Colour	45	90	12	97
Code	49	89	1	99.98
Come	50	89	1	99.98
Die	68	85	1	99.98
Effort	75	83	1	99.98
Grow	54	88	15	96.66
Drive	27	94	2	99.78
Many	130	71.45	33	92.66
Many times	50	89	–	100
Give	50	89	6	98.72
Keep	55	87	12	97.2
Quite down	45	90	1	99.98
Question	69	84.34	15	96.66

**Table 4** MFCC with DWT based classification for dynamic gestures (SVM as Classifier)

Gesture	Above	Brave	Big	Below	Bring	Coffee	Color	Cold	Come	die	Effort	grow	Dive	Many	Many times	Give	keep	Quiet down	Question
Above	67	0	0	0	0	0	0	33	0	0	0	0	0	0	0	0	0	0	0
Brave	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Big	0	33	67	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Below	0	0	0	67	0	0	0	0	0	33	0	0	0	0	0	0	0	0	0
Bring	0	0	0	0	67	0	33	0	0	0	0	0	0	0	0	0	0	0	0
Coffee	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
Color	0	0	0	0	0	67	33	0	0	0	0	0	0	0	0	0	0	0	0
Cold	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
Come	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
Die	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
Effort	0	0	0	0	0	0	0	0	0	0	67	0	33	0	0	0	0	0	0
Grow	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
Dive	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
Many	0	0	0	0	0	0	0	0	0	0	0	0	0	67	33	0	0	0	0
Many times	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0
Give	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
Keep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0
Quiet down	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0
Question	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

**Table 5** KNN k = 3 based classification for static gestures in different illumination condition

Static gesture	Cross	Hide	Hold	Stable	Straight	Together	Warm	With
Cross	100	0	0	0	0	0	0	0
Hide	0	100	0	0	0	0	0	0
Hold	0	0	100	0	0	0	0	0
Stable	0	0	0	100	0	0	0	0
Straight	0	0	0	0	100	0	0	0
Together	0	0	0	0	0	100	0	0
Warm	0	0	0	0	0	0	100	0
With	0	0	0	0	0	0	0	100

Performance of proposed method has been analysed by percentage of classification rate and confusion matrix where KNN and SVM are used as a classifier. Classification rate is calculated as:

$$\text{Classification rate} = \frac{(1 - \text{No. of frames misclassified})}{\text{total no. of frames}} \times 100 \quad (21)$$

In KNN experiments are performed at various values of k like k = 1, 3, 5, 7, etc. becomes the simple nearest neighbour classification. Here Euclidean distance is used for finding the nearest neighbour of a particular class. We get approximately similar results for k = 3, 5 and 7 may leads to the classification of ISL gesture and similarly in SVM we use linear kernel function for classifying an unknown gesture for a multi class SVM.

From Table 3 we have seen that SVM gives better classification accuracy then KNN because SVM handles non linearity present in the dataset by using kernel function. But KNN is a simple nearest neighbourhood based classifier which is not appropriate for classifying non-linear data and also it is very sensitive against noise. We have also observe that the classification accuracy decreases when shapes of two gestures are very similar.

Performance of our proposed algorithm is also tested by calculating confusion matrix shown in Table 4. Here three videos per gesture for training and five videos for testing have been taken for classification. Each video contains 150 frames. Confusion matrix is created by calculating True positive, true negative, false positive and false negative for each gesture case.

True positive = 50, True negative = 1018, False positive = 8, False negative = 7, Total positive = 58,

Total negative = 1025, Total population = 1083. Accuracy =  $(50 + 1018)/1083 = 1068/1083 = 98.61\%$ .

From Tables 3 and 4 we see that our proposed method gives satisfactory results for all the 19 dynamic gestures. We have also tested our experiments on static gestures. For static gestures, training is performed with one video and all other videos in different light condition are used for testing. 150 frames are there in each video. Accuracy based on confusion matrix shown in Table 5.

True positive = 7, False negative = 1, False positive = 1, True negative = 55, Total positive = 8,

Total negative = 144, Total population = 64; Accuracy =  $62/64 = 96.875\%$ .

Table 5 shows that the proposed method gives 100 percent classification result for all the static gestures because of no changes between the frames and also no change in the orientation therefore all the environmental noises and moment variation problems has been easily eliminated using WD and MFCC method. We have also

**Table 6** Classification results for SKIG kinect gestures dataset

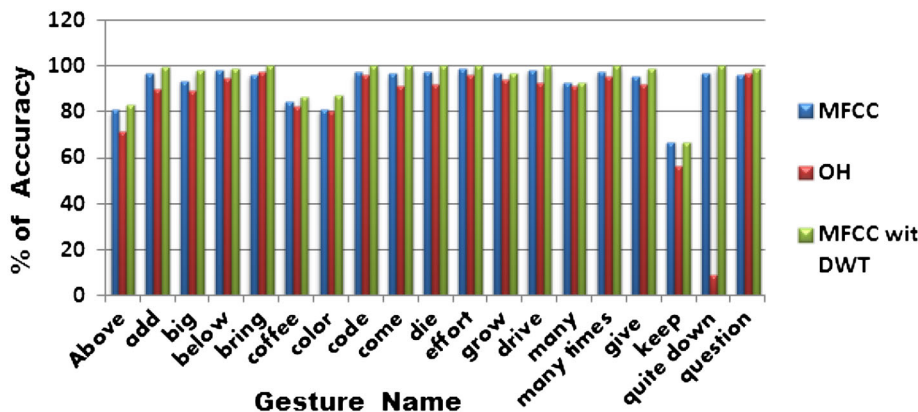
Class (action)	Wooden board		White plain paper		Paper with characters	
	Strong light		Strong light		Strong light	
	Frame misclassified	Acc. %	Frame misclassified	Acc. %	Frame misclassified	Acc. %
Circle	5	92	9	82	5	92
Triangle	22	56	10	80	7	90
Up-down	6	90	8	84	5	92
Right-left	13	73	9	82	4	93
Wave	5	92	11	80	3	94
Z	4	94	8	84	3	94
			20	60	7	86



**Table 7** Performance parameters

Method	Training dataset (150 Frames)	Testing dataset (150 Frames)	Parameters	KNN	SVM
OH [26]	5 video	3 video	9, 18, 27, 36	K = 1, 3, 5, 7 (Euclidean distance)	Linear kernel
MFCC [25]	5 video	3 video	10, 12, 13, 14, 15, 17	K = 1, 3, 5, 7 (Euclidean)	Linear kernel
WD + MFCC (proposed method)	5 video	3 video	HAAR, Daubachies wavelet/10, 12, 13, 14, 15, 17	K = 1, 3, 5, 7 euclidean	Linear kernel

**Fig. 16** Comparative results of MFCC, OH, WD with MFCC based methods



observe that all the static gestures are constant w.r.t. time whereas dynamic gestures vary with respect to time.

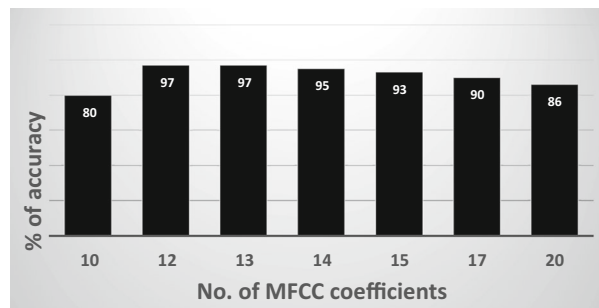
We have also tested our algorithm on Sheffield Kinect Gesture data set (SKIG) created by L. Liu at university of Sheffield [23, 24]. These dataset are dynamic in nature containing ten types of actions like circle, triangle, up-down, etc. Here data set are captured with 3 backgrounds (white wooden board, white plain paper, white paper with symbols), two light conditions (dark light, poor light) and two poses (clockwise and anti-clock wise). Therefore each gesture has  $3 \times 2 \times 2 = 12$  sample. We have tested our method on total 360 SKIG gesture dataset shown in Table 6.

From Table 6 we see the classification results on SKIG dataset with different light conditions and different backgrounds and observe that when light conditions will change drastically (white light to red light) and also the background is very much similar to skin colour, the performance of proposed method degrades.

### 4.1 Comparative results and analysis

We have performed experiments on various parameters of various methods explain in Table 7.

On the basis of these parameters comparative analysis of various feature extraction methods like [MFCC [25], Orientation histogram (OH) [26] and WD with MFCC

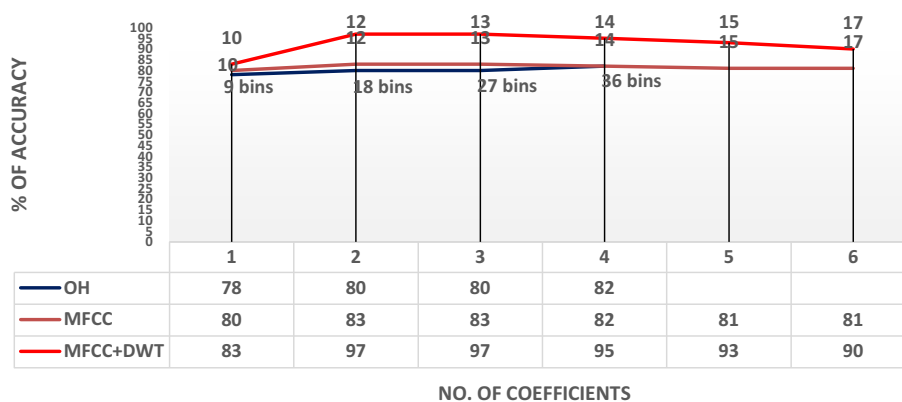


**Fig. 17** Comparative analysis of % of accuracy at different no. of MFCC coefficients

(proposed method)] have been performed on the basis of classification rate which is shown in Fig. 15. In OH [26] experiments are performed of 9, 18, 27, 36 bins (no. of features), in MFCC [25] we take 10,12,13,14, 15 and 17 cepstral coefficients for analysing the recognition accuracy and finally we tested our proposed algorithm on HAAR wavelet, Daubachies wavelet descriptor with 10, 12, 13, 14, 15 and 17 MFCC coefficients. Here number of decomposition levels in WD is 3.

Graph shown in Fig. 16 represents that WD with MFCC method provides high classification rate than other methods because of the multi-resolution and moment invariant properties of WD and spectral envelop property of MFCC. In this proposed framework WD reduces the 3/4 of the

**Fig. 18** Comparative analysis of % of accuracy at different no. of MFCC coefficients



**Table 8** Avg. processing time of 19 dynamic gestures

Total avg. time of OH	Total avg. time of MFCC	Total avg. time of DWT with MFCC
18.08 s	22.69 s	18.65 s

**Table 9** Average classification accuracy for static gestures

Features	% of classification rate static gestures
OH	97
MFCC	96
DWT with MFCC	100

**Table 10** Comparative results of SKIG dataset for different feature extraction technique

Features	% of classification rate SKIG dataset
OH	84
MFCC	85
DWT with MFCC	93

original image which is helpful for the reduction of time complexity as well as it requires very less space to store features. Contour of each image are created for finding moment invariant features of each gesture image by converting 2D image matrix into 1 D signal. MFCC extracts the spectral envelop of the features extracted from WD which minimizes the illumination variation present in the image.

We have also compared the classification accuracy of proposed gesture recognition algorithm at various numbers of MFCC coefficients like 10, 12, 13...20 which is shown in Fig. 17. From this graph we have found that the recognition accuracy is maximum at 12 and 13 MFCC

coefficients in comparison to other MFCC coefficients (10, 15, 20) when it applied with WD. We have also compare the results of proposed method with the existing methods OH and MFCC. The results are shown in Fig. 17.

In Fig. 18 we found that as the complexity (background variation, illumination condition, etc.) of dataset increases the recognition accuracy decreases in both the existing techniques. But proposed method provides approximately 97 % average accuracy with 12 WD with MFCC coefficients but this technique also fails when two of the gestures are approximately similar in shape. In this paper the time complexity of different methods is also compared shown in Table 8 where configuration of the system are core i5 processor with 4 GB RAM and MATLAB 2013 software.

Table 8 compares the average processing time (CPU time) of OH, MFCC and WD with MFCC method and found approximately similar processing time for all the methods but classification rate of WD with MFCC is much higher than other two methods.

Comparative analysis of all the three features have also been done on static gestures which is shown in Table 9. This analysis shows that all the three methods will give good recognition accuracy for static gestures because of variation between intermediate frames are none with respect to time. We have also perform the comparative analysis of all three features of SKIG dataset.

From Table 10 we have seen that the performance of our proposed method on SKIG dataset is good as compare to other feature extraction technique like OH and MFCC.

Finally learning is performed on HOAP-2 robot using Webots software. After classifying an unknown gesture generate.csv file to perform a particular gesture on HOAP-2 robot shown in Figs. 19 and 20.

In Figs. 19 and 20 the Webots simulation robotic software has been shown and hoap-2 is performing gestures ‘ARISE’, ‘ADD’, etc. Right hand side of Fig. 18 the Webots controller is shown where HOPA-2 programming is done. This controller has been interfaced with Matlab classification result where.csv file have been uploaded.

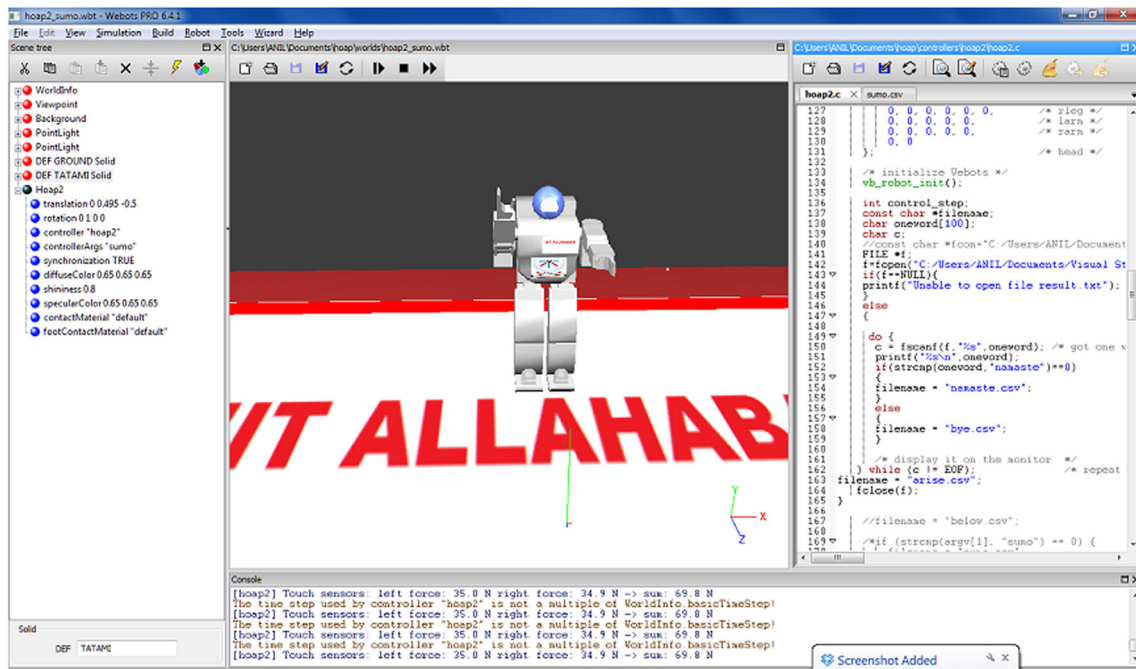


Fig. 19 Webot simulation of ‘ARISE’ gesture

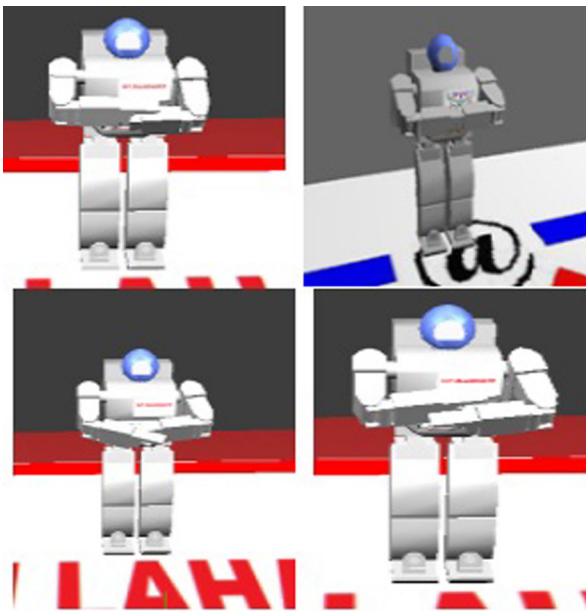


Fig. 20 Hoap-2 performing gesture, ‘Below’, ‘Add’, ‘Across’, ‘Above’

### 5 Conclusion and future work

In ISL dynamic gesture recognition the major problems are hand segmentation and feature extraction, which is helpful for classifying gestures in real time environment. In this work, we proposed a novel WD with MFCC based ISL gesture recognition method in which both the hands has

been used for performing any gestures. WD provides a time and frequency resolution as well as a moment invariant properties about any gesture. These properties make it invariant against scale, orientation, moment, phase, etc. It also reduces the 1/4 feature space of data set in first level of decomposition which is a solution for reduction of time complexity as well as space complexity. After reduction and noise elimination of images moment invariant features are extracted by converting 2d contour image into 1d plane. From these points spectral envelop of MFCC features are calculated using MFCC feature extraction technique. This technique is helpful against different background and various illumination conditions then OH and simple MFCC technique and also it is very effective for discriminating one gesture to another gesture. MFCC has been generally used for speech/voice recognition we tried to use it for gesture image analysis and it can be said that it is effective technique for gesture recognition also. When we used it with WD it gives 97 % recognition accuracy. Gestures are also classified using different types of classifiers like (KNN, SVM) with various parameters like  $K = 1, 3, 5, 7$  and different cepstral coefficients, etc. Analysis is also done by calculating FP, FN, TP and TN. By using these values a confusion matrix has been created. We also observed from these confusion matrix that the proposed technique gives better accuracy then other existing methods. Proposed technique has also been tested on SKIG data set which was published by University of Sheffield and found 93 percent accuracy towards this data set but other techniques like OH and MFCC provides only 84 and 85

percent accuracy which is very less in comparison to proposed method. After classifying an unknown gesture, HOAP-2 learning is performed through.csv file. Learning is done on Webots simulation software. Through this simulation software HOAP-2 performs similar gestures which has been classified by our proposed technique.

In advancement of ISL gesture recognition the full sentence recognition has to perform. The ISL full sentence based recognition with combination of voice recognition will be grate contribution in ISL gesture recognition field. The humanoid robot interaction with deaf people has to make based on real time voice with gesture recognition learning.

**Acknowledgments** We would like to thank our robita lab scholar's Avinash Kumar Singh and Anup Nandy. We would also thank Ms. Neha Singh M.Tech student and as well as thank all the research scholars of our robita lab of Indian Institute of Information Technology, Allahabad, for their comments and suggestions. We also thank our technical staff of robita lab for their help in data collection.

## References

- Mavridis N (2014) A review of verbal and non-verbal human-robot interactive communication. *Robot Auton Syst*. Available online 13 October 2014, ISSN 0921-8890. doi:10.1016/j.robot.2014.09.031. <http://www.sciencedirect.com/science/article/pii/S0921889014002164>
- Böhme H-J, Wilhelm T, Key J, Schauer C, Schröter C, Groß H-M, Hempel T (2003) An approach to multi-modal human-machine interaction for intelligent service robots. *Robot Auton Syst* 44(1):83–96, ISSN 0921-8890, doi:10.1016/S0921-8890(03)00012-5
- Ltd Fujitsu Automation Co. HOAP-2 instruction manual (2004) <http://robota.iitit.ac.in/hoap2instruction03e.pdf>
- Wagner P, Malisz Z, Kopp S (2014) Gesture and speech in interaction: an overview. *Speech Comm* 57:209–232, ISSN 0167-6393
- Starner T, Pentland A (1995) Real-time American sign language recognition from video using hidden markov models. In: International symposium on computer vision, pp 265–270
- Martin J, Crowley JL (1997) An appearance-based approach to gesture recognition. In: Proceedings of the 9th international conference on image analysis and processing, pp 340–347
- Kadous W (1995) Recognition of Australian sign language using instrumented gloves. 'Bachelor's thesis', University of New South Wales
- Sturman DJ (1992) Whole-hand input. Ph. D dissertation, Massachusetts Institute of Technology, 1992
- Watson R (1993) A survey of gesture recognition techniques. Technical Report TCD-CS-93-11, Department of Computer Science, Trinity College Dublin
- Freeman WT, Roth M (1995) Orientation histograms for hand gesture recognition. IEEE international workshop on automatic face and gesture recognition
- Nandy A, Prasad JS, Chakraborty P, Nandi GC (2010) Recognizing and interpreting Indian Sign Language gesture for human robot interaction. In: International conference on computer and communication technology (ICCCT 2010), and 17–19 Sept. 2010
- Singha J, Das K (2013) Hand gesture recognition based on Karhunen-Loeve transform. In: International conference on mobile and embedded technology, pp 365–371
- Gupta S, Jaafar J, Ahmad WFW, Bansal A (2013) Feature extraction using MFCC. *Signal Image Process* 4(4):101
- Burges CJC (1998) A tutorial on support vector machines for pattern recognition. *Int J Data Mining Knowl Discov* (Springer) 2(2):121–167
- Böhme H-J, Wilhelm T, Key J, Schauer C, Schröter C, Groß H-M, Hempel T (2003) An approach to multi-modal human-machine interaction for intelligent service robots. *Robot Auton Syst* 44(1):83–96, ISSN 0921-8890. doi:10.1016/S0921-8890(03)00012-5
- Fahmy MMM (2010) Palmprint recognition based on Mel frequency Cepstral coefficients feature extraction. *Ain Shams Eng J* 1(1):39–47, ISSN 2090-4479. <http://www.sciencedirect.com/science/article/pii/S209044791000067>
- Tomkins W (1931) Indian Sign language. Vol. 92. Courier Corporation
- Lin C, Liu A (2010) A tutorial of wavelet transform. NTUEE, Taiwan
- Kristian S (1998) A tutorial on Daubechies wavelet transforms. February 19, 1998
- Han W, Chan CF, Choy CS, Pun KP (2006) An efficient MFCC extraction method in speech recognition. In: Proceedings of IEEE international symposium on circuits and systems (ISCAS), September 2006, pp 4
- Cunningham P, Delany SJ (2007) K-nearest neighbour classifiers. *Mult Classif Syst* 1–17
- Burges CJC (1998) A tutorial on support vector machines for pattern recognition. *Int J Data Mining Knowl Discov* (Springer) 2(2):121–167
- <http://lshao.staff.shef.ac.uk/data/SheffieldKinectGesture.htm>
- Liu L, Shao L (2013) Learning discriminative representations from RGB-D video data. In: Proceedings of international joint conference on artificial intelligence (IJCAI), Beijing, China, 2013
- Baranwal N, Singh N, Nandi GC (2014) Implementation of MFCC based hand gesture recognition on HOAP-2 using Webots platform. In: 2014 international conference on advances in computing, communications and informatics, ICACCI, pp 1897–1902. IEEE
- Nandy A, Mondal S, Prasad JS, Chakraborty P, Nandi GC (2010) Recognizing and interpreting indian sign language gesture for human robot interaction. In: 2010 international conference on computer and communication technology (ICCCT), pp 712–717, IEEE
- Bhalke DG, Rama Rao CB, Bormane DS (2015). Automatic musical instrument classification using fractional Fourier Transform based-MFCC features and counter propagation neural network. *J Intell Inf Syst* 1–22
- Bhalke DG, Ram Rao CB, Bormane DS (2014) Stringed instrument recognition using fractional fourier transform and linear discriminant analysis. In: International conference in issues and challenges in intelligent computing techniques, ICICT-2014, 7–8 Feb 2014