Int. J. Mach. Learn. & Cyber.
DOI 10.1007/s13042-017-0705-5

ORIGINAL ARTICLE

A review of hand gesture and sign language recognition techniques

Ming Jin Cheok[1] · Zaid Omar[1] · Mohamed Hisham Jaward[2]

Received: 23 June 2016 / Accepted: 31 July 2017

Abstract Hand gesture recognition serves as a key for overcoming many difficulties and providing convenience for human life. The ability of machines to understand human activities and their meaning can be utilized in a vast array of applications. One specific field of interest is sign language recognition. This paper provides a thorough review of state-of-the-art techniques used in recent hand gesture and sign language recognition research. The techniques reviewed are suitably categorized into different stages: data acquisition, pre-processing, segmentation, feature extraction and classification, where the various algorithms at each stage are elaborated and their merits compared. Further, we also discuss the challenges and limitations faced by gesture recognition research in general, as well as those exclusive to sign language recognition. Overall, it is hoped that the study may provide readers with a comprehensive introduction into the field of automated gesture and sign language recognition, and further facilitate future research efforts in this area.

Keywords Computer vision · Gesture recognition · Image processing · Machine learning · Sign language

1 Introduction

Hand gestures are used as a way for people to express thoughts and feelings, it serves to reinforce information delivered in our daily conversation. Sign language is a structured form of hand gestures involving visual motions and signs, which are used as a communication system. For the deaf and speech-impaired community, sign language serves as useful tools for daily interaction. Sign language involves the use of different parts of body namely fingers, hand, arm, head, body and facial expression to deliver information. However, sign language is not common among the hearing community, and fewer are able to understand it. This poses a genuine communication barrier between the deaf community and the rest of the society, as a problem yet to be fully solved until this day.

Majority of sign language involves only upper part of the body from waist level upwards [46]. Besides, the same sign can have considerably large changes in shapes when it is in different location in the sentence [44]. Hand gestures can be categorized into several types such as conversational gestures, controlling gestures, manipulative gestures, and communicative gestures [62]. Sign language is a type of communicative gestures. Since sign language is highly structural, it is suitable to be used as a test-bed for computer vision algorithm [61].

The focus of this paper is on sign language recognition. However, research in sign language recognition is highly influenced by hand gesture recognition research, as sign language is a form of communicative gestures. Therefore, when reviewing literature in sign language recognition, it is also pertinent to study literature on gesture recognition.

Gestures and sign language recognition includes the whole process of tracking and identifying the signs performed and converting into semantically meaningful words

✉ Ming Jin Cheok
mingjin_91@hotmail.com

Zaid Omar
zaid@fke.utm.my

Mohamed Hisham Jaward
Mohamed.hisham@monash.edu

[1] Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Skudai, Malaysia

[2] School of Engineering, Monash University Malaysia, Subang Jaya, Malaysia

Published online: 08 August 2017

and expression. Some early efforts on gesture recognition can be dated back to 1993, where gesture recognition techniques are adapted from speech and handwriting recognition techniques. Darrell and Pentland [52] adapted Dynamic Time Warping (DTW) that had been successfully implemented in speech recognition to recognize dynamic gestures.

Later, Starner et al. [1] proposes using Hidden Markov Models (HMMs) to classify orientation, trajectory information and resultant shape of the sign language. HMMs is adapted from speech recognition, and its intrinsic properties make it suitable to be applied in gesture recognition. In [48], a total of 262 signs were collected from two different signers, and the average accuracy using HMMs classifier reaches accuracy of 94%. It is found out that the accuracy is greatly reduced when the database trained by the signs of one person is used to test by signs of another person, dropping to accuracy as low as 47.6%. Training the database with both signers improve accuracy to 91.3% [48].

Vogler and Metaxas [68] stated that the use of HMMs alone has several limitations especially in training context-dependent models. In [71], the authors employed Ascension Technologies Flock of Birds devices to collect the three-dimensional translation and rotation data of the sign. By using a bigram and epenthesis modeling, the average accuracy achieved is 95.83%. Research [68] used similar experiment setup, and by using a context-dependent HMMs and a method of coupling three-dimensional techniques, the system classifies 53 ASL and attained highest accuracy of 89.91%.

From the literature review, the most common sign languages recognition researches are based on American Sign Language (ASL), Indian Sign Language (ISL) and Arabic Sign Language (ArSL). Several other sign languages which are reviewed in this paper includes Tamil sign language (TSL), Dutch sign language (DSL), Korean sign language (KSL), Malaysian sign language (MSL), Persian sign language (PSL), English sign language (ESL), New Zealand sign language (NZSL), Chinese sign language (CSL), Japanese sign language (JPL), Vietnamese sign language (VSL), Brazilian sign language (Libras), Bangla sign language and Indonesian sign language.

This paper intends to focus on the reviewing of the state-of-the-art methods. Facial expression is used as part of sign language, it is however not discussed in this paper. The rest of the paper are organized as follows: Sect. 1 discusses the challenges, types of approaches and application domain of gesture recognition. Section 2 discusses the state-of-the-art techniques used in vision-based gesture and sign language recognition. Techniques used for pre-processing, segmentation, feature extraction, and classification are discussed separately. Section 3 discusses the techniques and technologies used in sensor-based gesture recognition. In Sect. 4, the techniques and finding by previous works are discussed and summarized. Lastly, thoughts about future works and conclusion are stated in Sect. 5.

1.1 Challenges in gesture recognition

Gestures recognition involves complex processes such as motion modeling, motion analysis, pattern recognition and machine learning [61]. It consists of methods with manual and non-manual parameters [48]. The structure of environment such as background illumination and speed of movement affects the predictive ability. The difference in viewpoints causes the gesture to appear different in 2D space. In some research, signer wears wrist band or colored glove to aid the hand segmentation process, such as in [3, 30, 48]. The use of colored gloves reduces the complexity of segmentation process. Several anticipated problems in a dynamic gesture recognition, includes temporal variance, spatial complexity, movement epenthesis, repeatability and connectivity as well as multiple attributes such as change of orientation and region of gesture carried out [53]. There are several evaluation criteria to measure the performance of a gesture recognition system in overcoming the challenges. These criteria are scalability, robustness, real-time performance and user-independent [57].

1.2 Type of approaches

Recognition of hand gestures can be achieved by using either a vision-based or sensor-based approaches.

1.2.1 Vision-based

Vision-based approaches require the acquisition of images or video of the hand gestures through video camera.

1. Single camera—Webcam, video camera and smartphone camera.
2. Stereo-camera—Using multiple monocular cameras to provides depth information.
3. Active techniques—Uses the projection of structured light. Such devices include Kinect and Leap Motion Controller (LMC).
4. Invasive techniques—Body markers such as colored gloves, wrist bands, and LED lights.

1.2.2 Sensor-based

This approach requires the use of sensors, instruments to capture the motion, position, and velocity of the hand.

1. Inertial measurement unit (IMU)—Measure the acceleration, position, degree of freedom and acceleration

of the fingers. This includes the use of gyroscope and accelerometer.
2. Electromyography (EMG)—Measures human muscle's electrical pulses and harness the bio-signal to detect fingers movements.
3. WiFi and Radar—Uses radio waves, broad beam radar or spectrogram to detect in-air signal strength changes.
4. Others—Utilizes flex sensors, ultrasonic, mechanical, electromagnetics and haptic technologies.

1.3 Hand gesture representation

The following are the type of gesture representation namely 3D model based and appearance based.

1. Model-based—This method describes the shape of the hand gesture in 2D or 3D space. It can be categorized into volumetric models and skeletal models. Volumetric model represents the hand gestures with high accuracy. Skeletal model reduces the hand gestures into set of equivalent joint angle parameters with segment length.
2. Appearance-based—Features are directly derived from the images or video using a template database. Image sequences is used a gesture templates which can be used as hand-tracking or simple gesture classification.

1.4 Hand gesture recognition application domain

The ability of a computer or machine to understand the hand gestures is the key to unlock numerous potential application. Potential application domains of gesture recognition system are as follows:

1. Sign language recognition—Communication medium for the deaf. It consists of several categories namely fingerspelling, isolated words, lexicon of words, and continuous signs.
2. Robotics and Tele-robotic—Actuators and motions of the robotic arms, legs and other parts can be moved by simulating a human's action.
3. Games and virtual reality—Virtual reality enable realistic interaction between user and the virtual environment. It simulates movement of users and translate the movement in 3D world.
4. Human–computer interaction (HCI)—Includes application of gesture control in military, medical field, manipulating graphics, design tools, annotating or editing documents.

2 Literature review on vision-based gesture recognition

The process of gesture recognition can be categorized into few stages in general, namely data acquisition, pre-processing, segmentation, feature extraction and classification as shown in Fig. 1. The input of static gesture recognition is single frames of images, while dynamic sign languages takes video, which is continuous frames of images as input. Vison-based approaches differs from sensor-based approaches mainly by the data-acquisition method. The focus of this section are the methodologies and techniques used by vision-based gesture recognition researches.

2.1 Data acquisition

In vision-based gesture recognition, the data acquired is frame of images. The input of such system is collected using images capturing devices such as standard video camera, webcam, stereo camera, thermal camera or more advanced active techniques such as Kinect and LMC. Stereo cameras, Kinect and LMC are 3D cameras which can collect depth information. In this paper, sensor-based recognition involves all techniques of data acquisition which does not uses cameras.

2.2 Image pre-processing

Image pre-processing stage are performed to modify the image or video inputs to improves the overall performance

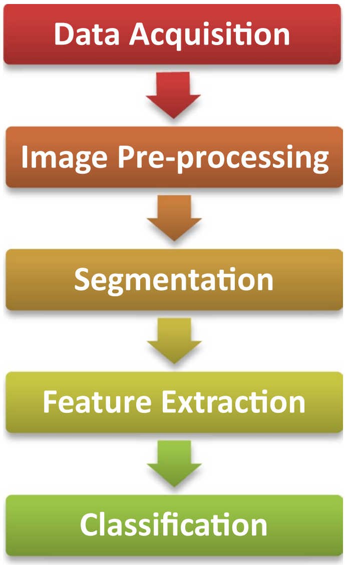

Fig. 1 Vision-based gesture recognition stages

of the system. Median filter and Gaussian filter are some of the commonly used techniques to reduce noises in images or video acquired. In research [79, 112], only median filtering is applied in this stage. Next, morphological operation is also widely used to remove unwanted information. For instance, Pansare et al. [19] first threshold the input image into binary image, then median and Gaussian filters is used to remove noises followed by using morphological operations as the pre-processing stage. In some researches, the images captured are downsized into a smaller resolution prior to subsequent stages. This technique is used in [12, 18, 19, 26, 66] has shown that reducing the resolution of the input image is able to improve the computational efficiency. Research in [120] tabulated the processing time associated with different downsizing factor of image resolution. In this research, division by 64 is the optimum scale as it reduced processing time by 43.8% without affecting the overall accuracy. Histogram equalization is used in [91] to enhance the contrast of the input images taken under different environment to uniform the brightness and illumination of the images.

2.3 Segmentation

Segmentation is the process of partitioning images into multiple distinct parts. It is a stage whereby the Region of Interest (ROI), is segmented from the remaining of the image. Segmentation method can be contextual or non-contextual. Contextual segmentation takes the spatial relationship between features into account, such as edge detection techniques. Whereas a non-contextual segmentation does not consider spatial relationship but group pixels based on global attributes.

2.3.1 Skin color segmentation

Skin color segmentation are mostly performed in RGB, YCbCr, HSV and HSI color spaces [6]. Several challenges toward achieving a robust skin color segmentation is sensitivity to illumination, camera characteristic and skin color [136]. HSV color space is popular as the Hue of palm and arm differs greatly, hence palm can be segmented from the arm easily [25]. Research [15] segments the face and hand in HSV color space. Chen et al. [33] performed skin color segmentation in RGB color space, using the rule of $R > G > B$ and matching with pre-stored sample skin color to find the skin color. Research [115] found that YCbCr is more robust for skin color segmentation compared to HSV in different illumination condition. Researches in [116, 119] found that CIE Lab color space is more robust as compared to YCbCr under different light variation. A normalized RG space was introduced in [117] to overcome the weakness of RGB which suffers from non-uniformity. Research in [118] proposed using K-means clustering method on the chrominance channels in YCbCr color space to separate the foreground which is the skin pixel from the rest of the background.

Skin color distribution and skin-color model classification can overcome the shortfall of applying constant skin- color threshold. Elmezain et al. [47] performed skin color segmentation in YCbCr color space. In [51], a single Gaussian Model based on YCbCr are used, and the classifier detects skin pixels from the background effectively [48]. Yang et al. [44] implemented the methodology in [48], however, Gaussian model is built instead of histogram model. Authors in [120] proposed a dynamic skin color modeling method by introducing weighting factors to locally trained skin model and globally trained skin model to obtain an adaptive skin color model.

2.3.2 Other segmentation method

Zhang et al. [9] introduced a segmentation based on difference background image in the presence of complex background. Otsu thresholding is first applied to the images, the proposed method of maximal between-class variance ‘3 s—principal’ is then used. Ghotkar and Kharate introduced a Hand Tracking and Segmentation (HTS) framework in [17]. The method involves applying Continuously Adaptive Mean-Shift (CAMShift) in HSV color space to create a histogram of skin pixels to find the suitable segmentation threshold value. Canny edge detection is then applied followed by dilation and erosion. Edge traversal algorithm is used lastly to segment the hand gesture from the background.

Lionnie et al. [18] compared the performance of ten variant including Sobel edge detection, low pass filtering, histogram equalization, skin color segmentation in HSI color space and desaturation, and found that desaturation provides highest accuracy. Desaturation process includes first converting into grayscale image by removing the chromatic channel while preserving only the intensity channel in HSI color space.

Entropy is measured by subtracting adjacent image frame to obtain hand motion information. Lee et al. [4] subtract one image from another successive image. The process includes measuring the entropy, separating hand region from images, tracking the hand region and recognizing hand gestures. A method of combining both entropy and skin color information named Entropy Analysis and PIM is proposed in [6] to segment hand gestures in a static and complex background.

2.3.3 Tracking

Tracking is considered as part of segmentation in this paper, as both tracking and segmentation is to extract the hand from the background. Tracking of a hands is usually difficult as the movement of hand can be very fast and their appearance can change vastly within a few frames. CAMShift method is used in several researches to track the position of the hand gestures such as application in [17, 27] to detect and track hand gestures. The CAMShift method detects the location of hand gestures by continuously adjusting the search window size.

Adaboost consist of linear combination of several weak classifiers with the aim to computes the sign of a weighted combination of weak classifiers to output a strong classifier. The authors in [29] detect hand movement using Adaboost with Histogram of gradient (HOG).

Particle filtering is normally used with other techniques for gesture tracking. In research [134, 139], combination of particle filtering and mean shift algorithm has shown to be able to recognize hand accurately. In research [133, 135, 138], tracking is performed using color features, and particle filtering has shown to be able to accurately track the movement of the gesture. Research in [137] introduced Kalman Particle Filter (KPF) as improvement to particle filtering in gesture tracking.

2.4 Feature extraction

Feature extraction is the transformation of interesting parts of input data into sets of compact feature vector [83]. In gesture recognition context, the features extracted should contain relevant information from the hand gestures input and represented in a compact version which serves as an identity of the gesture to be classified apart from other gestures.

2.4.1 Shift-invariant feature transform (SIFT)

SIFT is a scale and rotation invariant feature extraction technique introduced by Lowe [40]. SIFT describe an image by its interest points whereby detection requires multi-scale approach. At each level of the pyramid, the image is rescaled and smoothed by Gaussian function. The scale-space is defined by function, $L(x, y, \sigma)$ in Eq. 1.

$$L(x, y, \sigma) = G(x,\ y, \sigma) \times I(x, y) \tag{1}$$

The key-points extracted are the maxima and minima, which are calculated using difference-of-Gaussian (DoG) function, $D(x, y, \sigma)$. The Gaussian function convolved with the images, $D(x, y, \sigma)$ which is computed by subtracting two subsequent scales which is separated by a constant scale factor k with $k = \sqrt{2}$ as the optimum value as in Eq. 2.

$$\begin{aligned} D(x, y, \sigma) &= (G(x,\ y, k\sigma) - \ G(x,\ y, \sigma)) \times I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \tag{2}$$

At each point, $D(x, y, \sigma)$ is compared with eight neighbors of its scale, and nine neighbors up and down one scale. If the $D(x, y, \sigma)$ value is the maximum or minimum among the points, then it is extrema. In key-point localization stage, key-points with low contrast or are poorly localized are removed. The location of extremum, $\hat{x}$ is in Eq. 3.

$$\hat{x} = -\frac{\partial^2 D^{-1}}{\partial x^2}\frac{\partial D}{\partial x} \tag{3}$$

In orientation assignment, each key-points are assigned a consistent orientation based on local image properties. Finally, the SIFT descriptors is created in this stage by first lining up the key-points by offsetting the orientation. The matching of SIFT descriptors can then be performed by calculating the nearest neighbor and the ratio of closest-distance to second-closest distance. SIFT is invariant to a certain range of affine transformation, illumination variation, and changes in 3D viewpoint. In several gesture classification applications like in [15], the SIFT feature extracted from images are later quantized using K-means clustering before mapped into Bag-of-Feature (BoF). The above steps are taken to address the issue of different dimensionality of each SIFT features extracted as most classification technique requires inputs of equal dimensionality [11]. Using a similar method as [15], recognition of four gestures are performed with an average accuracy of 90% [66]. The authors claimed that although SURF has a faster processing speed, it is however is not as rotation invariant as SIFT [66]. Principal component analysis (PCA)-SIFT on the other hand has better illumination invariant, but are not scale invariant. SIFT features is extracted from ArSL in [23], and authors has shown the system to be robust against occlusion and rotation.

2.4.2 Speeded up robust feature (SURF)

SURF is developed based on SIFT. SIFT constructs scale pyramid, convolving the upper and lower scales of the image with DoG operator and searching the local extreme in scale space. Meanwhile, SURF scales filter up instead of iteratively reducing the image size. In SIFT, Laplacian of Gaussian (LoG) is approximated with DoG for finding scale-space. SURF approximates LoG with Box Filter. The convolution of box filter can be calculated easily using integral images, which is a fast and effective method in calculating the sum of pixels value.

In detection of key-points or descriptors, SURF uses an integer approximation of the determinant of Hessian blob detector. Integral image is the sum of intensity value for points in the image with location less than or equal to (x, y) as shown in Eq. 4.

$$S(x, y) = \sum_{i=1}^{x}\sum_{j=1}^{y} I(i,j) \tag{4}$$

SURF employs hessian blob detector to obtain interest points. The determinant of Hessian matrix describes the extent of the response. Hessian matrix with point x and scale σ is defined as in Eq. 5.

$$H(x,\sigma) = \begin{bmatrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma) \end{bmatrix} \tag{5}$$

where $L_{xx}(x,\sigma)$ is the convolution of the image with the second-order derivative of the Gaussian as described by Bay et al. [7]. To make the system scale-invariant, the scale space is realized as an image pyramid. With the use of integral image and box filter, the scale space can be realized by up-scaling. Finally, non-maximum suppression is applied in a $3 \times\ 3\ \times 3$ neighborhood to localize interest point in the image. Key-points between two images are matched nearest neighbors.

In research [144], using 500 test images, Support Vector Machine (SVM) classifier is built to classify both SIFT and SURF features, achieving accuracy of 81.2 and 82.8% respectively. In [145], the authors extracted SURF features from 12 images of each 24 classes of sign language, the overall accuracy is 63%. The author stated that SURF features are invariant to rotation if rotation is within 15°. In research [34], the authors extract SURF features to obtain the dominant movement direction of matched SURF feature points in adjacent frames, accuracy of 84.6% is achieved.

2.4.3 Principal component analysis (PCA)

PCA is a mathematical operation which utilizes orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components [83]. Given a training set of M images with an S-dimensional vector, PCA finds a t-dimensional subspace which its basis vectors correspond to the maximum variance direction in the original image space. The dimension of the new subspace is usually lower, where $t \ll s$. The mean, μ of all images in the training set given in Eq. 6, with x_i as the ith image with its columns concatenated in a vector.

$$\mu = \frac{1}{M}\sum_{i=1}^{M} x_i \tag{6}$$

PCA basis vectors are defined as eigenvectors of the Scatter matrix, S_T is computed as in Eq. 7.

$$S_T = \sum_{i=1}^{M} \left(x_i - \mu\right) \cdot \left(x_i - \mu\right)^2 \tag{7}$$

The eigenvectors and corresponding eigenvalues are calculated and the eigenvectors are stored by decreasing eigenvalues order. The eigenvectors with lower eigenvalues contains less information on the distribution of data, and these are filtered to reduce the dimensionality of data.

PCA has been widely used as a dimensionality reduction technique. PCA transforms possibly correlated variables into smaller number of principal components which are the uncorrelated variables [70]. PCA is used in [27] to extract features of 24 MSL. Locality Preserving Projection (LPP) is a modified PCA which utilizes known similarity between input features to adjust feature vector distances. Performance of PCA is compared LPP whereby the former achieved 92.8% and latter achieved 96.5% accuracies. PCA features are also used in [70] as measures of hand configuration and orientation. The authors however combined PCA with kurtosis position and chain code to improve the overall accuracy. PCA is used for dimensionality reduction in [77]. By calculating the eigenvalues, the authors omitted the components after the 12th and hence it reduces the computational complexity. In research [141], by classifying PCA features from 25 classes of VSL using Mahalonobis distance, it achieves accuracy of 91.5%.

2.4.4 Linear discriminant analysis (LDA)

Both LDA and PCA approaches finds the linear combination of features which best describe the data. For all samples of all classes, the between-class scatter matrix S_B and within-class scatter matrix S_W are given in Eq. 8.

$$\begin{aligned} S_B &= \sum_{i=1}^{M} M_i\left(x_i - \mu\right) \cdot \left(x_i - \mu\right)^T \\ S_W &= \sum_{i=1}^{M} \sum_{x_k \in X_i} \left(x_k - \mu_i\right) \cdot \left(x_k - \mu_i\right)^T \end{aligned} \tag{8}$$

M_i is the number of training samples in class i, c represents the number of distinct classes, μ_i is the mean vector of samples respective to class i with x_k being the kth images of the class. The aim of LDA is to determine matrix $W = \max \frac{S_B}{S_W}$ that maximizes S_B and minimizing S_W. Transformation matrix, W which projects the samples into reduced dimension space is as in Eq. 9.

$$W = \ W^T{}_{LDA} W^T{}_{PCA} \tag{9}$$

LDA maximizes class separability by finding linear combination of features which best discriminate among classes of objects [43]. PCA finds only the direction of maximal variance among features and does not consider the difference in classes [132]. LDA can be applied as a linear classifier and dimensionality reduction method. Author in [131] extracted PCA and LDA features from five classes of gesture. The accuracy of PCA is merely 26% while LDA is

100%, the poor performance of PCA could be due to overfitting. Research in [132] also compared the accuracy between PCA and LDA using five classes with 50 input images. The accuracy achieved for PCA and LDA is 60 and 62% respectively. It is stated that the noise rate can be reduced by reducing the dimensionality using both PCA and LDA. LDA is used by Tharwat et al. [23] to perform sign language recognition on Arabic sign language using a similar method as in [11, 15, 16]. SIFT features are first extracted from the images; however, LDA is applied in this research to widen the separation between classes of sign languages.

2.4.5 *Convexity defects and K-curvature*

Convexity defects and K-curvature method involves finding convex hull, convex and defects, center of the palm, angle between fingertips and palm center. This method was used in research [101, 102, 104, 109–114]. Several research uses global features with convexity defects to identify the gestures, research in [104] which uses solidity for identifying the fingers.

Shukla and Dwivedi utilized convexity defects and contour area as features and able to recognize five gestures with 100% accuracy [101]. Maisto et al. [102] uses Douglass Pecker method to approximate the hand gesture segmentation result with simpler contour. Research in [103] utilizes K-curvature in addition to convexity defects to improve the accuracy of recognizing fingertips. K-curvature are useful in finding the maximum and minimum points of the hand edges to identify the fingertips. Research in [105] classifies the fingers using angle between fingertips and palm center, and assumed the potential position of fingers which are unable to be detected. Research [106] uses Randomized Decision Forest (RDF) and estimation of joint position to classify gesture based on fingertips. Author in [107] further improve the accuracy of convexity defects by improving rule-based method suggested by [108] to identify fingers whether they are upright, bent, looped, joint or separated.

2.4.6 *Features extraction in frequency domain*

Feature extraction in frequency domain involves transformation of time domain input data into frequency domain. This includes Cosine Transform, Fourier Transform and Wavelet Transform. In research [33], the authors stated the advantage of Fourier Descriptors (FD) is its size-invariant properties. FD is also rotation invariant as rotation in hand gestures only causes a phase change. Also, noises can be reduced by removing the high frequency, as noises and quantization errors only cause local variation of high frequency.

The authors in [27] claimed that contour based features including FD, Wavelet Descriptors (WD) and B-spline prone to suffers from poor performance when the fingers are curled inward and lose its contour properties. Region-based features such as Principal Curvature Based Region detector (PCBR) utilizes semi-local structural information for instance the curvilinear shapes and edges which are robust to intensity, color and shape variation. 2-D Wavelet Packet Decomposition (WPD-2) uses Haar basis function up to level two which utilizes the high frequency channels with significant information. A hybrid feature extraction method of PCBR, WPD-2 and Convexity defect are performed in [51] to recognize 23 static ISL. The hybrid of three features are shown to outperform the hybrid of only any two of the features using k-NN classifier. Similar findings are obtained when classified using SVM. Discrete Wavelet Transform (DWT) features is extracted for classification of 23 static PSL in [69]. DWT can be realized by iteration of filters with rescaling. The resolution of the signal, is determined by the filtering operations [69].

2.4.7 *Others feature extraction method*

Some features have advantages over the others, yet suffers from others drawback. For instance, SURF is much computational efficient as compared with SIFT [7]. However, SURF is not as rotational and illumination invariant [26]. Hybrid features extraction has been used in several researches to overcome the limitations in single features. Hu moment invariant geometric features is extracted from hand gestures and combined with SURF in [26]. Using hybrid of SVM and *k*-NN as classifier, the authors compared their proposed method with SIFT, SURF, Hu-moment. It is shown that hybrid of SURF with Hu moment has the highest accuracy.

Liu et al. [25] proposed a hybrid features fusion of Hu moment invariant, finger angle count, skin color angle, and non-skin color angle. Accuracy of 90% is achieved in matching ten gestures. Local Binary Pattern (LBP) is a computational efficient texture operator which labels pixels in an image by thresholding the neighborhood of each pixels and the results is considered as binary number. LBP is used as feature extraction [142] on both Chinese and Bangladeshi numeral gesture dataset and able to achieve accuracy of 87.13 and 85.10% respectively [142]. In [154], the authors extract both HOG and Zernike invariant moment (ZIM) shape descriptors to classify 40 classes of Libras. The magnitude of ZIM are rotational invariant, and hence the magnitude is used as features. The overall accuracy achieved is 96.77%. Chakraborty et al. [8] compared four methods of gesture recognition techniques namely Subtraction, Gradient, PCA, and Rotation invariant. Rotation Invariant which is based on LBP provides the highest accuracy. Pansare [20] compared performance of different feature extraction method namely Discrete Cosine Transform (DCT), edge oriented histogram, centroid and Fourier Transform and shown that DCT has better result. In research [149], combination of SIFT, Hu Moments are FD features are extracted from input images. PCA and LDA is applied to these features to reduce the dimensionality. Using SVM, classification of 26 classes of CSL achieves accuracy of 99.8%.

2.5 Classification

Classification can be categorized into supervised and unsupervised machine learning techniques. Supervised machine learning is a technique that teaches the system to recognize certain pattern of input data, which are then used to predict future data. Supervised machine learning takes in a set of known training data and it is used to infer a function from labeled training data. An unsupervised machine learning is used to draw inferences from datasets with input data with no labeled response. Since no labeled response is fed into the classifier, there is no reward or penalty weightage to which classes the data is supposed to belong.

2.5.1 Static gesture classification

Static gestures are single images which involves no time frame. Static gestures recognition is mostly used to recognize finger-spelled signs.

2.5.1.1 Support vector machine (SVM) SVM is a supervised machine learning technique. It finds the optimal hyperplane to separate the data points. SVM maximize the margin around the separating hyperplane. Optimization techniques are employed in finding the optimal hyper plane. Two hyperplanes are found which best represent the data. w is the weight vector for $\vec{w}$, for training data $\left(\vec{x}_1, \vec{y}_1\right), \ldots \left(\vec{x}_n, \vec{y}_n\right)$, where y_i are either 1 or -1, indicating to which class the data $\vec{x}_i$ belong. The weight vector decides the orientation of decision boundary, whereas bias point, b decides its location. The hyperplane can be represented by Eq. 10.

$$\vec{w} \cdot \vec{x}_i + b = 0 \tag{10}$$

The points above the hyperplane will have positive y_i, and points below will have negative y_i. The distance between the support vector and plane is $distance = \frac{1}{\|\vec{w}\|}$. The Margin, M is twice the distance to support vector, hence margin is defined as $M = \frac{2}{\|\vec{w}\|}$. w need to be minimized as in Eq. 11 to maximize the margin, M.

$$\min\ L = \frac{1}{2}\left\|\vec{w}\right\|^2 \quad where\ \ y_i(\vec{w} \cdot \vec{x}_i + b) \geqslant 1 \tag{11}$$

The performance of SVM has been compared with k-NN [23], Naive Bayes [16], and shown that SVM has better performance over the other methods. SVM with linear kernel perform better than non-linear Gaussian kernel [76]. The authors experimented with two size of gesture database. The accuracy of classification using linear SVM with 12 ESL dropped from 99.2 to 82.3% when the number of gestures increased to 25 ESL. The method of using SIFT to extract features from images followed by quantization using K-means clustering before mapped into BoF classification using SVM has shown promising results in [11, 15, 16, 66]. Proximal SVM (PSVM) employs an equality constraint instead of inequality constraint in SVM. PSVM is used in [21], seven features are extracted which are group into a matrix with each row representing single feature vector. PSVM handle multiple classes more efficiently and classification of 20 TSL achieved 91% accuracy. Multi-dimensional classification using non-linear SVM has higher accuracy as compared to using linear SVM [16]. In [23], SIFT features is extracted from 30 ArSL. With 7 train images each, accuracy of 99% is obtained.

2.5.1.2 Artificial neural network (ANN) ANN is an information-processing system with several performance characteristics in common with that of biological neural networks [69]. ANN is generally defined by three parameters, namely the interconnection pattern between different layers of neurons, the weight of interconnections, and the activation function. A neuron has inputs $x_1,\ x_2 \ldots x_n$, which each are labelled with a weight $w_1,\ w_2 \ldots w_n$ that measures the permeability. The neuron function can be represented as nonlinear weighted sum in Eq. 12, where K is the activation function.

$$y = K\sum_{i=1}^{n} w_i x_i \tag{12}$$

Akmeliawati et al. [27] applied ANN with 7392 gestures signals to train a system to recognize 13 gestures. Using a single ANN with 45 inputs and 14 outputs with two hidden layers, an average accuracy of 96.02% is achieved. Gesture Recognition Fuzzy Neural Network (GRFNN) was introduced in [5] to adapt fuzzy control for learning parameters. The advantage of eliminating the needs of preselecting training pattern improves the accuracy. In recognition of 36 ASL, GRFNN achieved 92.19% accuracy [5]. Time Delay NN (TDNN) focus on working with continuous data. Multi-Layered Perceptron NN (MLPNN) is a feedforward neural network with one or more layers between input and output layer. It is devised from linear perceptron to distinguish data which are not linearly separable.

Karami et al. [69] employed MLPNN to classify 32 classes of PSL. Using an MLPNN with 92 input nodes, one hidden layer with 21 neurons, and five linear output neurons, the accuracy achieved is 94.06%. A recurrent NN is when the connections between neural forms a directed cycle.

Elman RNN is a partial RNN, whereby the feedforward connections are flexible while the recurrent connections are fixed. The connections have a set of feedback connection which allows the network to remember cues from recent past while the rest is feedforward network. By performing the back-propagation with Simulated Annealing training method, results are promising for dynamic sequences training in both [79, 82].

2.5.1.3 *K-nearest neighbor (k-NN)*

K-NN is a non-parametric statistical method whereby input data is classified by a majority vote of its neighbor. The data will be assigned to the class most common among its k nearest neighbors. Euclidean distance as in Eq. 13 is a commonly used similarity measures.

$$distance = \sum_{i=1}^{N} (a_i - b_i)^2. \tag{13}$$

The Euclidean distance between each testing data point to the training data points are calculated. The testing data are then labelled according to the majority classes in the k th nearest training data. *K*-NN is used in comparison with the parametric Bayes classifier in [35] and shown that the former has better performance. In research [146], *k*-NN is used to classify 30 test images from each 26 gestures, the highest overall accuracy achieved is 90%. However, several researches on comparing the accuracy of *k*-NN against SVM in equal test and train data size, has shown that the overall accuracy of *k*-NN is comparatively lower [23, 51, 76, 89]. Nevertheless, *k*-NN has the advantage of being computational efficient and easy to be implemented.

2.5.1.4 *Unsupervised static classification method*

Unsupervised classification is often referred to as clustering. It differs from supervised classification where the input data are not labelled. K-means clustering is one of the commonly used unsupervised classification in gesture recognition. It is a vector quantization technique which partition n observations into k clusters in which each observation belongs to the cluster of the nearest mean. In researches [16, 66], K-means clustering is used to cluster the training features vectors into classes of sign languages. The centroids are then used as inputs to BoF model to simplify the classification problems. In literature [81], the authors employed K-means clustering to calculate the code vector coordinates in four dimensions.

Self-organizing maps (SOM) is a variant of ANN which is an unsupervised learning method. SOM differs from other supervised ANN method as it uses competitive learning in contrast with error-correction learning such as backpropagation with gradient descent. Self-Growing and SOM (SGONG) proposed in [84] combines the advantages of Growing Neural Gas (GNG) while adapting a reduce parameter and more biologically plausible design. It retains the ability to insert nodes and neurons where needed in SOM without the need to introduce new nodes [85]. In literature [84], the construction of SGONG on the hand gestures input, allows the position of each fingers to be identified. Classification of 31 static gestures achieved average accuracy of 90.45%. Euclidean distance is the real distance between two points in the m-dimensional space [25]. In some researches, classification is performed through template matching by calculating Euclidean distance between feature vectors of input gestures and a template. The nearest distance is the matching result. Examples of gesture classification by calculating Euclidean distance can be found in [8, 19, 25].

2.5.2 *Dynamic gesture classification*

In dynamic gestures recognition, two different signs cannot be compared using Euclidean space due to the misalignment in time. DTW and HMM are widely applied due to the ability to align frames of signs and compute the likelihood of similarity [49]. Other notable classification techniques in dynamic environment include Finite State Machine (FSM), Kalman filtering, advanced particle filtering, and condensation algorithm.

2.5.2.1 *Dynamic time warping (DTW)*

DTW is useful in measuring the similarity between two temporal sequences which may be different in length and speed. DTW finds the best mapping with the minimum distance using ‘time warping’ which allows compress of expand in time to obtain the best match. The goal of DTW is to find the mapping path mapping path mapping path $p = (p_1, \ldots, p_L)$ with $p_L = \left(n_l, m_l\right) \in [1{:}N] \times [1{:}M]$ for $l \in [1{:}L]$ satisfying the following constraints:

1. Boundary condition:
   $p_1 = (1, 1)$ and $p_L = (N, M)$.
2. Step size condition:
   $p_{l+1} - p_l \in \{(1, 0), (0, 1), (1, 1)\}$
   $for\ l \in [1{:}L - 1]$.
3. Monotonicity condition:
   $n_1 \le n_2 \le \cdots \le n_L$ and
   $m_1 \le m_2 \le \cdots \le m_L$.

Given two sequences x_{nl} , y_{mi}, the local distance can be compared. The total cost $c_p(X, Y)$ of a warping path p between X and Y with respect to the local cost measure c is given as in Eq. 14.

$$c_p(X, Y) = \sum_{l=1}^{L} c(x_{nl}\, , y_{mi}) \tag{14}$$

DTW was implemented in [52], reaching an accuracy of 96% in recognizing dynamic sign ‘hello’ among a continuous sentence consisting of four other signs. In [50], the author introduced Statistical DTW which use DTW to train a statistical model, and shown to outperform HMMs in handwriting recognition. Lichtenauer et al. [49] introduced a hybrid approach by using Statistical DTW (SDTW) only for time warping and a separate classifier on the warped features. Two statistical classifiers for warped features are proposed by the authors, namely the Combined Discriminative Feature Detectors (CDFDs) and Quadratic Classification on DF Fisher Mapping (Q-DFFM). Both proposed method of SDTW with CDFDs and SDTW with Q-DFFM are shown to have better accuracy than SDTW alone and HMMs. Both methods uses a selective-based discriminative features (DFs) which is able to reduce the dimensionality and noises by removing non-DFs. DTW has also been successfully applied in the classification of dynamic gestures in [51] on features vectors of PCBR, WPD-2D, and convexity defects.

2.5.2.2 Hidden Markov models (HMMs) HMMs is a stochastic method of analyzing time-varying data with spatio-temporal variability [63]. A first-order HMM has two assumptions, namely the probability of a state depends only on the previous state, and the probability of an output observation k depends only on the state that results in the observation q_i and not any other observations. A HMM is defined by three fundamental problem, namely finding the likelihood of observation, decoding the best hidden state sequence, and training the HMM parameters.

The likelihood computation can be achieved using Forward algorithm. Viterbi algorithm is used to decode the sequence of state which results in the observation sequence. The parameter learning or training stage can be achieved by using Baum–Welch algorithm or Forward–Backward algorithm.

Nianjun and Lovell [81] experimented HMMs with different model structure namely the Left–right and full connection topologies, and found that it has no significant effect on the accuracy. HMMs is used in [33] to classify 20 gestures with 1200 test and train sequences respectively, and accuracy achieved is 98.5%. Another application of HMM to classify ten dynamic gestures using 200 train sequences and 98 test sequences achieved 94.29% [47]. Elmezain et al. [10] applied Gaussian Mixture Model (GMM) in segmentation and Baum–Welch algorithm with Forward algorithm in gesture classification stage.

Parametric HMMs (PHMMs) is introduced in [31] to improve the parameter-dependent nature of a standard HMMs. PHMMs is parameterizes the underlying output probabilities of the states in HMMs. There are several researches in improving scalability of HMMs. Performance of HMMs, Linked HMMs (LHMMs) and CHMMs are compared for three gestures and found out that accuracy of CHMMs is least sensitive to the initial values of the parameters [64]. Parallel HMMs (PaHMMs) is proposed in [28] as improvement to factorial HMMs (FHMMs) in [65] and coupled HMMs (CHMMs) in [64]. Both FMMs and CHMMs require the interactions of the processes to be modelled and hence every combination of actions must be trained [28]. The authors used PaHMMs to classify 22 ASL with 400 training sentences and 99 test sentences. An average accuracy of 94.23 and 84.85% for sign and sentence accuracy respectively are achieved [32]. Other application of HMMs can also be found in [2, 45].

2.5.2.3 Other dynamic classification methods There are several other supervised classification techniques used in classification of static gestures. Wong and Cipolla [77] employed Sparse Bayesian Classifier and Relevance Vector Machine (RVM) in classification of ten gestures. The authors stated that the benefit of using Bayesian classifier with probabilistic nature enable the system to be applied in complex motion analysis that must maintain multiple hypotheses [77]. In this research, the authors used RVM classifier, which is a simple binary classifier over SVM classifier as the output of RVM is a probabilistic value instead of a binary true-or-false value. In addition, the dispersity of the model stored by the RVM classifier enables RVM to be less computational heavy.

Hong et al. [87] used FSMs for classification of dynamic gestures. The advantage of FSMs over the commonly used HMMs is that in HMMs, the states and structure must be predefined. In FSMs, the alignment of training data can be done simultaneously with the construction of gesture model [87].

2.6 Active techniques

LMC is a portable USB peripheral device with two monochromatic cameras and three infrared Light-Emitting Diode (LED). It models the 3D position of both hands and fingers and provides 28 information features including fingertips, palm center, hand orientation and so on. LMC has been used to aid the recognition of sign languages in [97]. Classification performed using Naive Bayes classifier and MLP-NN and achieve average accuracy of 98.3 and 99.1% respectively. Chuan et al. [96] utilizes seven features obtained from LMC and using SVM to classify 26 ASL with 79.83% accuracy.

Kinect is a device with a color sensor, an Infrared Emitter, and a depth camera, which collects color and depth information. Chai et al. [22] utilized Kinect to obtain color and depth information which are used to create a 3D motion trajectories database. With database of 239 Chinese Sign languages and four samples per language, recognition rate achieved is 96.32%.

Marin et al. [100] utilized LMC and Kinect to obtain position of fingertips, palm center and hand orientation features obtained from LMC together with color and depth information from Kinect and form a histogram of features. Multi-classes SVM with Gaussian Radial Basis Function (RBF) kernel are then used to classify ten different sign languages with 140 samples each and shown an average accuracy of 91.3% in real-time recognition [33].

LMC however is unable to detect fingers when they are touching with each other or when fingers are occluded [99]. LMC is also limited when the hand is not perpendicular to the camera or when signer is wearing bracelet and long sleeves [100]. The tracking ability of LMC is tested in [74] by using 1500 samples by performing the known gestures and actual outcome of tracking. The average accuracy experimented is 96.34%.

3 Literature review on sensor-based gesture recognition

This section discusses the techniques used in sensor-based gesture recognition research. Sensor-based approaches generally relies on the use of sensors which are physically attached to users to collect position, motion and trajectories of fingers and hand data. These approaches reduce the need of pre-processing and segmentation stage, which are essential to vision-based gesture recognition. Features such as flex angle of fingers, orientation and the absolute position of hand are often in 3D space, and hence it contains the depth information which is useful in telling distance of gesture away from source of sensors. Sensor-based approaches often requires users to wear a glove with sensors or with probes attached to the arm of users. These instruments are required to be set up prior to the recognition, and these often limit the approaches to a laboratory setup.

3.1 Data glove

Data gloves used in gesture and sign language recognition utilizes IMU sensors such as gyroscope and accelerometer to obtain the orientation, angular, acceleration information. Flex sensors are present in some data gloves to obtain finger bending information. VLP-Data glove is a pair of flex-sensor gloves that consist of fiber optic transducer, which measures the flex angles, position, and orientation data. Kim et al. [41] used 16 raw data generated from VPL-Data glove and categorized the motion of both hands into ten basic motion which are used as input to Fuzzy Min–max Neural Network (FMNN). With 25 KSL words, the authors achieved an accuracy of 85%. In [42] recognizes 250 Taiwanese Sign Language words. The features extracted from Data Glove include flexion of fingers, position, angles and motion trajectory data. The features are used as input to HMM to recognize 51 types of posture, six types of orientation and eight types of motion and achieved 100% accuracy for all three categories. The authors also tested isolated gestures, short sentence and long sentences with 250 vocabularies and achieve 89.5, 70.4, and 81.6% respectively. In [53], ten flex angle and 3D absolute position generated by VPL Dataglove is used, HMMs are applied to recognize ten dynamic gestures and achieve accuracy of 99%.

3.2 Electromyography (EMG)

Electromyography is the recording of the electrical activities of the muscle tissues using electrodes attached to the skin or inserted into the muscles. Zhang et al. [73] uses a fusion of 3-axis input from accelerometer and 5-channel of EMG signals attached on the hand of the user. Using Fuzzy K-means clustering as classifier, 72 dynamic CSL is recognized with 93.1% accuracy. Kim et al. [35] used EMG sensors attached on the arm of users to obtain finger movement input. Using a linear combination of both *k*-NN and Bayes classifier to classify 20 classes of gestures, the approach achieved 94% accuracy. Ahsan et al. [24] extracted EMG pattern signatures from the signals for each movement and then ANN utilized to classify the EMG signals based on features. Myo armband is an arm wearable with both IMU and EMG sensors. Research in [144] uses Myo armband in recognition of 20 classes of Libras. Using SVM classifier, the average accuracy is 98.6%.

A hybrid method of combining vision input from LMC and surface EMG (SEMG) is done in [74]. Using SEMG alone, an accuracy of 86% is achieved. Together with LMC depth camera input, the accuracy is increased to 95%. Research in [140] utilizes both SEMG and Cyberglove to classify the flexion and extension of all five fingers. PCA is used before Independent Component Analysis (ICA) as PCA reduce computational complexity. Classification using LDA reaches accuracy of 90%.

3.3 WiFi and Radar

Another type of technology used for gesture recognition is WiFi oriented gesture control [75]. The authors claimed that this method is much simple to be applied as compared to Kinect technology. It uses WiSee technology that consists of multiple antennas to focus on one user to detect the user's gesture. Signals used in Wifi do not require line of sight and can traverse through walls. It utilizes the properties of Doppler shift, which is the change in frequency of a wave as its sources move relative to the observer. A similar research is done in literature [67].

Abdelnasser et al. [92] proposed a gesture recognition system using WiFi named WiGest. WiGest system leverages changes in WiFi signal strength to detect in-air hand gestures nearby the user's mobile device. Using single access point (AP), the recognition rate is 87.5%. The accuracy increases to 96% when the three overheard APs are used. In research [93], the authors used smart radar sensor that operates in the 2.4 GHz Industrial, Scientific and medical (ISM) band. The features are extracted based on magnitude differences and Doppler shifts of the gesture performed. *K*-NN is used for classification of four gestures, and achieved accuracy of 98%. Unlike vision-based gesture recognition, WiFi and radar offers the flexibility of position and orientation, without having to face the source of camera.

4 Discussion

This section provides an overview of previous surveys done on gesture and sign language recognition works as well as the techniques applied in different researches.

4.1 Previous survey on gesture and sign language recognition works

Reviews and surveys had been conducted on researches on gesture and sign language recognition, these papers may provide a comprehensive overview of methods used in gesture recognition. Table 1 lists the previous works on the analysis of hand gesture recognition and their focus.

4.2 Summary of techniques and algorithm reviewed

Information including techniques applied, database size, performance, and scope of previous work are presented and tabulated in this section. Tables 2 and 3 includes the techniques used and summary of vision-based gestures and sign language recognition researches reviewed. Table 2 listed research in static gesture recognition, whereas Table 3 listed research in dynamic gesture recognition. Table 4 highlights technologies used in vision-based active techniques and sensor-based gesture recognition. The techniques used are categorized by the classification, feature extraction, and segmentation. Pre-processing method are however not included in this section as it is found that many papers lack detailed information of this stage.

The *accuracy/sample sizes* column stated the highest accuracy achieved by the proposed method as well as the sample sizes of the dataset. The samples size are the total samples used, including both train and test samples. Sample size 15 × 80 for instance, translate to 15 classes of gesture with 80 sample each.

In Table 3, the information of the numbers of sentences or sign used for training and testing of dynamic gestures which are stated explicitly by the authors are included. Most literature reviewed in this paper focus on recognition of only one hand. Research which involves recognition of more than a single hand is stated explicitly in the *Scope* column. Most vision-based research reviewed uses a standard camera or a webcam. For research involving stereo camera or invasive

Table 1 List of gesture recognition reviews

References	Authors	Year	Focus
[72]	Gavrila	1998	Analyzed human movement based on 2-D approaches with, and without explicit models as well as 3-D approaches
[61]	Wu and Huang	1999	Discussed the application of HMMs in gesture recognition as well as other static and dynamic recognition approaches
[58]	Moeslund and Granum	2001	Discussed papers on approaches used in initialization, tracking, pose estimation and recognition from 1980 to 2000
[63]	Wang and Liang	2003	Discussed hand gesture techniques until 2003
[54]	Thomas	2005	Discuss segmentation and classification method in gesture recognition
[59]	Moeslund et al.	2006	Discussed papers on approaches used in initialization, tracking, pose estimation and recognition between 2000 and 2006
[56]	Ribeiro and Gonzaga	2006	Compared real-time GMM background subtraction algorithm
[60]	Mitra and Acharya	2007	Discuss extensively the most commonly used classification method
[37]	Murthy and Jadon	2009	A general review on application domain, challenges, approaches and previous work on vision based gesture recognition system
[57]	Rautaray and Agrawal	2012	Comprehensively review on challenges, and approaches in gesture recognition
[39]	Khan and Ibraheem	2012	Various recognition system technique comparison on all stages
[55]	Ibraheem and Khan	2012	Compared different ANN approaches
[38]	Chaudhary et al.	2013	Discussed the approach in gesture recognition and several classification methods
[36]	Sharma et al.	2014	Surveyed different segmentation and feature extraction techniques
[95]	Mohandes et al.	2014	Survey on vision-based and sensor-based approach in ArSL recognition

Table 2 Vision-based static gesture recognition summary

References	Authors	Year	Classification	Feature extraction	Segmentation	Accuracy/sample size	Scope
[3]	Lockton	2002	Template score matching in canonical frame	Normalization of hand scaling and yaw angle Quantized skin concentration map	Manual placement of hand under bounding boxes Region growing Pixel position information to remove forearm pixel	99.1% 42 gestures	ASL and gestures (wrist band)
[4]	Lee et al.	2004	Maximum distribution value in centroidal profile	Chain code Distances of hand contours to centroid	PIM-based motion detection Skin color (HSI)	95% six gestures	Gestures
[5]	Nguyen and Toshiaki	2005	GRFNN	–	Skin color (HSV) Kalman filter and hand blob analysis	92.19 and 36×200 samples	ASL
[111]	Manresa et al.	2005	Finger posture recognition	Convexity defects	Skin color (HSI) Pixel labelling tracking	98% 8×40 samples	Gestures
[6]	Shin et al.	2006	Maximum distribution value in centroidal profile	Chain code Distances of hand contours to centroid	PIM-based motion detection Skin color (HSI)	95% six gestures	Gestures
[8]	Chakraborty et al.	2008	Template matching by shortest Euclidian distance	Rotation invariant (Comp) subtraction, gradient and PCA	Thresholding, Gaussian filter	4×4 samples	Gestures
[12]	Rokade et al.	2009	–	Angle, vertical distance, horizontal distance	Skin color (YIQ + YCbCr) Histogram matching Thinning method	92.13% 10×60 samples	ASL
[27]	Akmeliawati et al.	2009	ANN	PCA LPP	CAMShift	96.02% NZSL 13 signs	NZSL
[84]	Stergiopoulou and Papamarkos	2009	Finger likelihood based classification	SGONG Finger angle and distances to palm center	Skin color (YCbCr)	90.45% 31 gestures	Gestures
[66]	Dardas et al.	2010	BoF K-means clustering SVM	SIFT	–	90% 4×100 samples	Gestures
[15]	Dardas and Georganas	2011	BoF K-means clustering SVM	SIFT	Skin color (HSV) Viola Jones	96.23% 6×100 samples	Gestures and face
[26]	Rekha et al.	2011	*k*-NN + SVM	SURF + Hu moment invariant **(Comp)** SIFT, SURF and Hu moment invariant	Skin color (RGB) K-means clustering	15×80 samples	ASL
[14]	Hassan and Misra	2010	Block pixel matching	Brightness factor **(Comp)** edge detection	Block Scaling Normalization Thresholding	91% 6×4 samples	Gestures
[104]	Tofighi et al.	2010	Hu Moments	Convexity defects Solidity	Adaptive Histogram Template of Skin	89% 10×50 samples	Gestures

Table 2 (continued)

References	Authors	Year	Classification	Feature extraction	Segmentation	Accuracy/sample size	Scope
[69]	Karami et al.	2011	MLP-NN	DWT	Cropping image	94.06% 32 × 20 samples	PSL
[76]	Kurdyumov et al.	2011	SVM (Gaussian kernel) **(Comp)** SVM (Linear kernel), *k*-NN	Relative area, height and width of hand pixel Fourier transform	Background subtraction Thresholding Normalization of hand size	93% 25 × 100 samples	ESL
[18]	Lionnie et al.	2012	*k*-NN	–	Desaturation **(Comp)** Low pass filtering, Histogram equalization, Skin color (HSI), Sobel edge detection	83.78% 6 × 480 samples	Gestures
[19]	Pansare et al.	2012	Template matching by shortest Euclidian distance	Centroid and area of hand edges	Thresholding Blob and Crop Sobel edge detection	90.19% 26 × 100 samples	ASL
[25]	Liu et al.	2012	Template matching by shortest Euclidian distance	Hu moment invariant, angle count, skin color angle and non-skin color angle	Skin color (HSV)	90% 10 × 100 samples	Gestures
[91]	Sethi et al.	2012	Correlation matching	SIFT	Region growing **(Comp)** Skin color	–	ASL
[113]	Tariq et al.	2012	ANN	Convexity defects	Skin color (YCbCr)	62% 33 × 10 samples	113
[20]	Pansare et al.	2013	Template matching by shortest Euclidian distance	DCT **(Comp)** Centroid matching, Edge-Oriented-Histogram and Fourier transform	Otsu Thresholding Skin color (YIQ) Thinning method Blob & Crop	10 × 2 samples	ISL
[21]	Rajathi and Jothilakshimi	2013	Proximal SVM	Solidity, perimeter, convex area, major axis length, minor axis length, eccentricity and orientation	Thresholding	91% 20 samples	TSL and face
[110]	Ganapathyraju	2013	Finger posture recognition	Convexity defects	Skin color (YCbCr)	4 gestures	Gestures
[88]	Bhuyan et al.	2014	–	Metacarpo- phalangeal (MP) joints extraction using edge detection technique	Skin color (HSI + HSV) component labeling Distance transform	–	Gestures 2 hand and body
[131]	Yasir et al.	2014	ANN	LDA	Skin color (YCbCr)	100% 15 × 7 samples	Bangla SL
[142]	Jasim and Hasanuzzaman	2014	*k*-NN	LDA (Comp) LBP	–	92.4% (CSL) 88.6%(Bangla SL)	CSL Bangla SL
[145]	Hartanto et al.	2014	Template matching by Euclidean distance	SURF	Skin color (HSV)	63% 24 × 12 samples	Indonesian SL

Table 2 (continued)

References	Authors	Year	Classification	Feature extraction	Segmentation	Accuracy/sample size	Scope
[23]	Tharwat et al.	2015	SVM (Comp) *k*-NN (*k* = 5) and (*k* = 1)	SIFT LDA	–	99% 30 × 7 samples	ArSL
[112]	Lahiani et al.	2015	SVM	Convexity defects	Skin color	93% 10 × 10 samples	Gestures
[141]	Huong et al.	2015	Template matching by Mahalonobis distance	PCA	Skin color (HSV)	91.5% 25 × 10 samples	VSL
[147]	Bastos et al.	2015		HOG ZIM	MP-ANN on skin color (HSV, YCbCr and YIQ)	96.8% 40 × 240 samples	Libras
[146]	Gupta et al.	2016	*k*-NN	HOG + SIFT	–	26 × 30 samples	ISL Two hands
[149]	Pan et al.	2016	SVM	SIFT, Hu-moments and FD PCA and LDA	Skin color (YCbCr) with GMM	99.8%(CSL) 26 × 300 samples 94% (ASL) 36 × 2425 samples	CSL ASL

techniques, it is indicated in the *Scope* column. Some research compared performance of different techniques used. The techniques with most prominent result is presented first, and those techniques compared are stated after "**(Comp)**". Research which uses hybrid of techniques are indicated by "+". In the event of information not explicitly stated or are found to be vague by the authors of this paper, the information is left blank.

Pre-processing method are carried out to improve accuracy and processing time. The most commonly applied pre-processing techniques includes Median and Gaussian filter to remove noises. Downsizing the input image is often used prior to segmentation and the following stage in gesture recognition research to reduce the computational load. Tracking of hand movement are often carried out using Particle filtering, CAMShift method, and Adaboost tracking algorithm.

Skin color segmentation is a popular choice of segmentation method. The most commonly use color space are HSV, YCbCr, and CIE Lab as these color space easily differentiate skin color from the background. The research shown that skin color segmentation with other features such as edge detection and threshold improves the segmentation result. Skin color modelling approaches and adaptive skin model are more robust towards dynamic changing background than explicitly selected threshold in color space.

In feature extraction stage, appearance-based and model-based recognition uses different approaches. Appearance-based method in both time and frequency domain extracts useful information from pixels of the input image. Model-based method includes both volumetric and skeletal modelling in either 2D or 3D environment, this includes convexity defects and K-curvature techniques. SURF is more computational efficient as compared to SIFT. However, the performance of SURF is not as invariant as SIFT. PCA are mostly used in hybrid with other features to improve overall accuracy. PCA and LDA are also useful in dimensionality reduction, which serves to reduce the computational load. Hybrid feature extraction method has been used widely in recent gesture recognition research.

In dynamic gestures classification, some notable methods are DTW and HMM. Several variants of HMM are proposed such as PaHMM, CHMM, and LHMM to address scalability issues. PHMM on the other hand are proposed as a solution to reduce the parameters-dependent characteristic of a standard HMM. In classification of static gestures, some of the commonly used techniques are SVM and ANN. Many researches which performed comparison of classification method has shown SVM in overall have better performance.

In the context of sign language recognition specifically, the vocabulary of a sign languages system is tremendously vast. However, the vocabulary used in most research until today is little as compared to that of a sign language system. The scalability issue is another challenges exclusive

Table 3 Vision-based dynamic gesture recognition summary

References	Authors	Year	Classification	Feature extraction	Segmentation	Accuracy/sample size	Scope
[30]	Starner	1995	HMMs	X and Y position, angle of axis of least inertia, and eccentricity	–	95% 395 sentences	ASL Two hands (colored gloves)
[45]	Min et al.	1997	HMMs	Chain code	–	98.3% 12 × 20 gestures	Gestures
[48]	Grobel and Assan	1997	HMMs	X and Y position, center and size of hand	–	94% 262 signs	DSL Two hands (colored gloves)
[87]	Hong et al.	2000	FSMs	Center of hands and head	Skin-color tracking	–	gestures and body
[81]	Liu and Lovell	2003	HMM	Blob Ellipse Model K-means clustering	Skin color (HSV) CAMShift	8 × 50 samples	Gestures and face
[33]	Chen et al	2003	HMM	FD	Skin color (RGB) Kirsch edge detection	90% 20 Gesture	Gestures and body
[77]	Wong and Cipolla	2005	Sparse Bayesian Classifier RVM	Motion Gradient Orientation (MGO) PCA	–	91.8% 10 × 30 samples	Gestures
[86]	Rybach et al.	2006	HMMs	PCA LDA	Score function based image difference motion tracking	RWTH-Boston 104	ASL and body
[10]	Elmezain et al.	2008	HMMs	Quantization of hand orientation	Skin color (YCbCr) with GMM Blob analysis Depth information	94.72% 36 × 30 samples	Gestures and body (stereo camera)
[49]	Lichtenauer et al.	2008	SDTW + CDFDs SDTW + Q-DFFM	Discriminative feature (DF) selection	–	75 signers 120 signs	DSL
[47]	Elmezain et al.	2009	HMMs	Quantization of hand orientation Zero-code work detection	Skin color (YCbCr) with GMM Blob analysis Depth information	98.6% signs 94.29% sentences 10 × 30 samples	Gestures and body (stereo camera)
[13]	Appenrodt et al.	2010	HMMs	Zero-code work detection	Skin color (YCbCr) Depth information	98% 36 × 30 samples	Gestures and body (thermal camera) (stereo camera)
[34]	Bao et al.	2011	Correlation analysis	SURF Trajectory directions	–	84.6% 26 × 40 samples	Gestures and body
[51]	Rekha et al.	2011	*k*-NN SVM (1-vs-all) DTW	PCBR + WPD-2 + convexity defect	Skin color (YCbCr) with single Gaussian model	86.3% (static) 23 × 40 sample 77.2% (dynamic) 3 × 22 sample	ISL Two hands
[70]	Zaki and Mahmoud	2011	HMMs	Kurtosis position + PCA + motion chain code	Skin color Connected component labelling	89.1% RWTH-Boston 50	ASL
[79]	Zhang et al	2011	Simulated annealing back propagation NN	Hand center Edge information	Skin color (RGB) Robert gradient sharpening Thresholding	92.7% 20 × 40 samples	Gestures

Table 3 (continued)

References	Authors	Year	Classification	Feature extraction	Segmentation	Accuracy/sample size	Scope
[29]	Wang et al	2012	HMMs	Hand edges Boundaries of regions in skin color	Adaboost with HOG Condensation partitioning sampling	800 samples with 10 gestures	Gestures
[82]	Barros et al	2013	HMMs DTW Elman RNN	SURF + local contour sequence (LCS)	–	93% 4 gestures	Gestures
[89]	Baranwal and Nandi	2016	SVM (Comp) *k*-NN	WD and Mel Sec frequency Cepstral coefficients (MFCC)	Otsu thresholding edge detection	23 × 8 static 19 × 8 dynamic	ISL

for sign language recognition. Although many researches on gesture and sign language recognition have been done, however none has deployed on a large scale to date [95]. Despite most researches done on gesture and sign language recognition shown promising results, a practical implementation of such system is still far from reality as there is several underlying assumptions in most researches. Most researches done might be suitable in a controlled lab setting but does not generalize to arbitrary setting [37]. One common assumption in most researches is to assume a high contrast and stationary backgrounds with constant ambient lighting conditions.

4.3 Benchmark databases

In sign language recognition research, benchmark databases are available as standard reference for future researches. Benchmark databases allows comparison of a model-free and person-independent approaches [122]. These includes Purdue RVL-SLLL [124], RWTH-PHOENIX-Weather [125], ATIS Sign Language Corpus [127], SIGNUM Corpus [78], RWTH-BOSTON-50, RWTH-BOSTON-104, and RWTH-BOSTON-400 [128]. Standard accuracy measurement is introduced for performance to be compared. RWTH-BOSTON-50 used error rate (ER) as accuracy measurement. RWTH-BOSTON-104 and ATIS used tracking error rate (TER), Word error rate (WER) and Independent word error rate (PER) as assessment for accuracy. There are several researches conducted using the benchmark databases and their result are shown in Table 5. Nevertheless, these databases are not widely referenced in research of sign language recognition. The recognition results presented in most papers reviewed are based on each author's own collection of data.

5 Conclusion and future work

Gesture recognition has been an on-going research driven by its wide potential for applications such as sign language recognition, remote control robots and human–computer interaction in virtual reality. Nevertheless, the barriers to achieving an accurate and robust system persist, namely the occlusion of hand, presence of affine transformation, scalability of database, different background illumination and high computational cost.

There are growing numbers of emerging technology such as EMG, LMC, and Kinect which capture gesture information more readily. The common pre-processing method used are Median and Gaussian filter as well as downsizing of images prior to subsequent stages. Skin color segmentation is one of the most commonly used segmentation method. Color space which are generally more robust towards illumination condition are CIE Lab, YCbCr and HSV. More recent research utilizes combination of several others spatial features and modeling approaches to improve segmentation performance.

Common feature extraction with appearance-based approaches includes SIFT, SURF, PCA, LDA and DWT. Model-based approaches includes both volumetric and skeletal modelling and convexity defects techniques. Hybrid of feature extraction method has been widely used to provide more robust feature for recognition.

From previous works, HMMs appear as promising approaches towards dynamic gesture recognition as it has been successfully implemented in many researches. In static gesture recognition, SVM is the most popular method as it has shown to have better performance in several researches. Several variants are proposed towards existing method and hybrids of methods are becoming more widely used as it can overcome the shortfall of the single method. There are significant gaps to be filled for gesture recognition to be able to be put into actual use. The numbers of research using benchmark database are far less

Table 4 Active techniques and sensor-based gesture recognition summary

References	Authors	Year	Classification	Feature extraction	Sensor type	Accuracy/sample size	Scope
[22]	Chai et al	2013	Template matching by Euclidian distance	3D trajectory of hand	Kinect	96.32% 239 × 5 sentences	CSL Two hands and body
[102]	Maisto et al	2013	Comparing Mean and Standard deviation	Convexity defects	Kinect	98% Three gestures	Gesture
[106]	Keskin *et al.*	2013	RDF	Mean shift parameters Per pixel value	Kinect	82.1% 40 gestures	Gesture
[107]	Billiet et al	2013	Finger posture recognition	Convexity defects	Kinect	95.5% 10 gestures	Gesture
[144]	Sykora et al	2014	SVM	SURF **(Comp)** SIFT	Kinect	82.2% 15 × 100 samples	Gesture
[103]	Yeo et al	2015	FSM	Convexity defects K-Curvature	Kinect	86.66% Nine gestures	Gesture
[80]	Molchanov et al	2015	3D Convolutional NN	Depth threshold image gradient value	Kinect	77.5% 19 gestures	Gestures
[96]	Chuan et al	2014	SVM (Gaussian RBF kernel) **(Comp)** *k*-NN	Seven features from LMC	LMC	79.83% 26 × 4 samples	ASL
[97]	Mohandes et al	2014	MLP-NN **(Comp)** Naïve Bayes	12 features from LMC	LMC	99.1% 28 × 10 samples	ArSL
[98]	Funasaka et al	2015	Decision Tree	Hand position, velocity and movement	LMC	82.71% 24 static samples	ASL
[100]	Marin et al	2014	SVM (Gaussian RBF kernel)	Angle, distance, orientation (LMC) Color and depth (Kinect)	LMC and Kinect	91.3% 10 × 140 samples	ASL
[41]	Kim et al	1996	Fuzzy min–max NN	x and y axis movement 16 raw angle, orientation, position data	VPL Data Glove	85% 25 words	KSL Two hand
[53]	Nam and Wohn	1996	HMM	10 flex angles, 3D absolute position Plane Fitting	VPL Data Glove	80% Ten gestures	Gestures
[42]	Liang and Ouhyoung	1998	HMM	10-finger flexion, position, angle, motion trajectory	Data Glove	80.4% 50 static 30 dynamic	TSL
[90]	Bukhari et al	2015	Template matching by Euclidian distance	PCA	Data Glove	92% 26 × 250 samples	ASL
[35]	Kim et al	2008	*k*-NN + Bayes	Variance, mean value, and standard deviation and Fourier variance	EMG	94% Four gestures	Gestures
[24]	Ahsan et al	2011	Back-propagation ANN	MAV, RMS, VAR, SD, ZC, SSC and WL	EMG	88.40%	Gestures
[73]	Zhang et al	2011	Fuzzy K-means clustering	3-axis ACC and 5-channel EMG signals	Accelerometer and EMG	93.1% 72 × 2 sentences	CSL

Table 4 (continued)

References	Authors	Year	Classification	Feature extraction	Sensor type	Accuracy/sample size	Scope
[74]	Kainz and Jakab	2014	Deep Believe Network	Hand position and orientation EMG signals	LMC and surface electromyography (SEMG)	95%	Gestures
[140]	Naik et al	2014	LDA classifier	LDA ICA	sEMG Cyberglove	90%	Gesture
[143]	Abhishek et al	2016	Finger posture recognition	Electrode signals	Capacitive touch sensor	92% 26 static	ASL
[150]	Gabriel et al	2016	SVM	EMG signals	Myo Armband EMG and IMU	98.6% 20 static	Libras
[93]	Wan et al	2014	*k*-NN (k = 3)	Magnitude difference + Doppler shift	Radar	98% Four gestures	Gestures
[92]	Abdelnasser et al	2015	Compare string pattern with gesture template	DWT edge detection	WiFi (WiGest)	96% Seven gestures	Gestures

Table 5 List of Research using Benchmark Database

References	RWTH-BOSTON-50	RWTH-BOSTON-104	ATIS	Purdue-ASL
[70]	ER 10.9%			
[130]	ER 28.4%			
[121]	ER 17%	WEB 22%		
[86]		WEB 22%		
[123]		PER 2.2% WER 2% TER 2.3%		
[126]		WER 17%		
[129]		PER 23.8% WER 26.5%	PER 34.7% WER 45.1%	
[148]				95.5%

compared to those collected their own database. Future works using benchmarked databases are advised to allow for direct comparison between algorithms used.

Acknowledgements This research was made possible by the funding of the Ministry of Higher Education Malaysia and Universiti Teknologi Malaysia through the Research University Tier 1 Grant (Vote No. 09H75).

Compliance with ethical standards

Funding This study was funded by Ministry of Higher Education Malaysia and Universiti Teknologi Malaysia through the Research University Tier 1 Grant (Vote No. 09H75).

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Starner T, Weaver J, Pentland A (1998) Real-time American sign language recognition using desk and wearable computer based video. IEEE Trans Pattern Anal Mach Intell 20:1371–1375
2. Starner T, Pentland A (1997) Real-time American sign language recognition from video using hidden Markov models. In: Motion-based recognition. Springer, pp 227–243
3. Lockton R (2002) Hand gesture recognition using computer vision 4th year project report, pp 1–69
4. Lee J, Lee Y, Lee E, Hong S (2004) Hand region extraction and gesture recognition from video stream with complex background through entropy analysis. In: Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th annual international conference of the IEEE, IEEE, pp 1513–1516
5. Binh ND, Ejima T (2005) Hand gesture recognition using fuzzy neural network. In: Proc. ICGST conf. graphics, vision and image process, Cairo. pp 1–6
6. Shin J-H, Lee JS, Kil SK, Shen DF, Ryu JG, Lee EH, Min HK, Hong SH (2006) Hand region extraction and gesture recognition using entropy analysis. IJCSNS Int J Comput Sci Netw Secur 6:216–222

7. Bay H, Tuytelaars T, Van Gool L (2006) Surf: speeded up robust features. In: European conference on computer vision. Springer, pp 404–417
8. Chakraborty P, Sarawgi P, Mehrotra A, Agarwal G, Pradhan R (2008) Hand gesture recognition: a comparative study. In: Proceedings of the international multiconference of engineers and computer scientists, Citeseer, pp 19–21
9. Zhang Q, Chen F, Liu X (2008) Hand gesture detection and segmentation based on difference background image with complex background. In: Embedded software and systems, 2008. ICESS'08. International conference, IEEE, pp 338–343
10. Elmezain M, Al-Hamadi A, Michaelis B (2008) Real-time capable system for hand gesture recognition using Hidden Markov models in stereo color image sequences. J WSCG 16(1–3):65–72
11. Kim D, Dahyot R (2008) Face components detection using SURF descriptors and SVMs. In: Machine vision and image processing conference, 2008. IMVIP'08 international, IEEE, pp 51–56
12. Rokade R, Doye D, Kokare M (2009) Hand gesture recognition by thinning method. In: Digital image processing, 2009 international conference, IEEE, pp 284–287
13. Appenrodt J, Al-Hamadi A, Michaelis B (2010) Data gathering for gesture recognition systems based on single color-, stereo color-and thermal cameras. Int J Signal Process Image Process Pattern Recognit 3:37–50
14. Hasan MM, Misra PK (2011) HSV brightness factor matching for gesture recognition system. IJIP 4(5):456–467
15. Dardas NH, Georganas ND (2011) Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. IEEE Trans Instrum Meas 60:3592–3607
16. Schmitt D, McCoy N (2011) Object classification and localization using SURF descriptors. CS 229:1–5
17. Ghotkar AS, Kharate GK (2012) Hand segmentation techniques to hand gesture recognition for natural human computer interaction. Int J Hum Comput Interact IJHCI 3:15
18. Lionnie R, Timotius IK, Setyawan I (2012) Performance comparison of several pre-processing methods in a hand gesture recognition system based on nearest neighbor for different background conditions. J ICT Res Appl 6:183–194
19. Pansare JR, Gawande SH, Ingle M (2012) Real-time static hand gesture recognition for American sign language (ASL) in complex background. J Signal Inf Process 3:364
20. Pansare JR, Dhumal H, Babar S, Sonawale K, Sarode A (2013) Real time static hand gesture recognition system in complex background that uses number system of Indian sign language. Int J Adv Res Comput Eng Technol IJARCET 2:1086–1090
21. Rajathi P, Jothilakshmi S (2013) A static Tamil sign language recognition system. Int J Adv Res Comput Commun Eng 2(4):1–7
22. Chai X, Li G, Lin Y, Xu Z, Tang Y, Chen X, Zhou M (2013) Sign language recognition and translation with kinect. In: IEEE Conf, AFGR
23. Tharwat A, Gaber T, Hassanien AE, Shahin M, Refaat B (2015) Sift-based arabic sign language recognition system. In: Afro-European conference for industrial advancement, Springer, pp 359–370
24. Ahsan MR, Ibrahimy MI, Khalifa OO (2011) Electromygraphy (EMG) signal based hand gesture recognition using artificial neural network (ANN). In: Mechatronics (ICOM), 2011 4th international conference, IEEE, pp 1–6
25. Yun L, Lifeng Z, Shujun Z (2012) A hand gesture recognition method based on multi-feature fusion and template matching. Procedia Eng 29:1678–1684
26. Rekha J, Bhattacharya J, Majumder S (2011) Hand gesture recognition for sign language: a new hybrid approach. In: Proc. conference on image processing computer vision and pattern recognition, pp 1–7
27. Akmeliawati R, Dadgostar F, Demidenko S, Gamage N, Kuang YC, Messom C, Ooi M, Sarrafzadeh A, SenGupta G (2009) Towards real-time sign language analysis via markerless gesture tracking. In: Instrumentation and measurement technology conference, I2MTC'09, IEEE, pp 1200–1204
28. Vogler C, Metaxas D (1999) Parallel hidden markov models for american sign language recognition. In: The Proceedings of the seventh IEEE international conference, IEEE, pp 116–122
29. Wang X, Xia M, Cai H, Gao Y, Cattani C (2012) Hidden-Markov-models-based dynamic hand gesture recognition. Math Prob Eng 2012:986134. doi:10.1155/2012/986134
30. Starner TE (1995) Visual recognition of American sign language using hidden Markov models. Dept of Brain and Cognitive Sciences, Massachusetts Inst of Tech, Cambridge
31. Wilson AD, Bobick AF (1999) Parametric hidden Markov models for gesture recognition. IEEE Trans Pattern Anal Mach Intell 21:884–900
32. Vogler C, Metaxas D (2001) A framework for recognizing the simultaneous aspects of American sign language. Comput Vision Image Underst 81:358–384
33. Chen F-S, Fu C-M, Huang C-L (2003) Hand gesture recognition using a real-time tracking method and hidden Markov models. Image Vis Comput 21:745–758
34. Bao J, Song A, Guo Y, Tang H (2011) Dynamic hand gesture recognition based on SURF tracking. In: Electric information and control engineering (ICEICE), international conference, IEEE, pp 338–341
35. Kim J, Mastnik S, André E (2008) EMG-based hand gesture recognition for realtime biosignal interfacing. In: Proceedings of the 13th international conference on Intelligent user interfaces, ACM, pp 30–39
36. Jones MJ, Rehg JM (2002) Statistical color models with application to skin detection. Int J Comput Vis 46:81–96
37. Murthy G, Jadon R (2009) A review of vision based hand gestures recognition. Int J Inf Technol Knowl Manag 2:405–410
38. Chaudhary A, Raheja JL, Das K, Raheja S (2013) Intelligent approaches to interact with machines using hand gesture recognition in natural way: a survey. arXiv preprint arXiv:13032292
39. Khan RZ, Ibraheem NA (2012) Survey on gesture recognition for hand image postures. Comput Inf Sci 5:110
40. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60:91–110
41. Kim J-S, Jang W, Bien Z (1996) A dynamic gesture recognition system for the Korean sign language (KSL) IEEE Trans Syst Man Cybern Part B Cybern 26:354–359
42. Liang R-H, Ouhyoung M (1998) A real-time continuous gesture recognition system for sign language. In: Automatic face and gesture recognition, 1998. Proceedings. Third IEEE international conference, IEEE, pp 558–567
43. Delac K, Grgic M, Grgic S (2005) Independent comparative study of PCA, ICA, and LDA on the FERET data set. Int J Imaging Syst Technol 15(5):252–260
44. Yang R, Sarkar S, Loeding B (2010) Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming. IEEE Trans Pattern Anal Mach Intell 32:462–477
45. Min B-W, Yoon H-S, Soh J, Yang Y-M, Ejima T (1997) Hand gesture recognition using hidden Markov models. In: Systems, Man, and Cybernetics, 1997. Computational cybernetics and simulation. 1997 IEEE international conference, IEEE, pp 4232–4235
46. Bellugi U, Fischer S (1972) A comparison of sign language and spoken language. Cognition 1:173–200
47. Elmezain M, Al-Hamadi A, Appenrodt J, Michaelis B (2009) A hidden Markov model-based isolated and meaningful hand gesture recognition. Int J Electr Comput Syst Eng 3:156–163

48. Grobel K, Assan M (1997) Isolated sign language recognition using hidden Markov models. In: Systems, Man, and Cybernetics, 1997. Computational cybernetics and simulation. 1997 IEEE international conference, IEEE, pp 162–167
49. Lichtenauer JF, Hendriks EA, Reinders MJ (2008) Sign language recognition by combining statistical DTW and independent classification. IEEE Trans Pattern Anal Mach Intell 30:2040–2046
50. Bahlmann C, Burkhardt H (2004) The writer independent online handwriting recognition system frog on hand and cluster generative statistical dynamic time warping. IEEE Trans Pattern Anal Mach Intell 26:299–310
51. Rekha J, Bhattacharya J, Majumder S (2011) Shape, texture and local movement hand gesture features for indian sign language recognition. In: 3rd international conference on trendz in information sciences and computing (TISC2011), IEEE, pp 30–35
52. Darrell T, Pentland A (1993) Space-time gestures. Comput Vis Pattern Recognit. Proceedings CVPR'93. 1993 IEEE computer society conference, IEEE, pp 335–340
53. Nam Y, Wohn K (1996) Recognition of space-time hand-gestures using hidden Markov model. In: ACM symposium on Virtual reality software and technology, pp 51–58
54. Thomas G (2011) A review of various hand gesture recognition techniques. VSRD Int J Electr Electron Commun Eng 1(7):374–383
55. Ibraheem NA, Khan RZ (2012) Vision based gesture recognition using neural networks approaches: a review. Int J Hum Comput Interact IJHCI 3:1–14
56. Ribeiro HL, Gonzaga A (2006) Hand image segmentation in video sequence by GMM: a comparative analysis. In: 19th Brazilian symposium on computer graphics and image processing, IEEE, pp 357–364
57. Rautaray SS, Agrawal A (2015) Vision based hand gesture recognition for human computer interaction: a survey. Artif Intell Rev 43:1–54
58. Moeslund TB, Granum E (2001) A survey of computer vision-based human motion capture. Comput Vis Image Underst 81:231–268
59. Moeslund TB, Hilton A, Krüger V (2006) A survey of advances in vision-based human motion capture and analysis. Comput Vis Image Underst 104:90–126
60. Mitra S, Acharya T (2007) Gesture recognition: a survey. IEEE Trans Syst Man Cybern Part C Appl Rev 37:311–324
61. Wu Y, Huang TS (1999) Vision-based gesture recognition: a review. In: International gesture workshop, Springer, pp 103–115
62. Wu Y, Huang TS (1999) Human hand modeling, analysis and animation in the context of HCI. In: Image processing, ICIP 99. Proceedings. 1999 international conference, IEEE, pp 6–10
63. Wang L, Hu W, Tan T (2003) Recent developments in human motion analysis. Pattern Recognit 36:585–601
64. Brand M, Oliver N, Pentland A (1997) Coupled hidden Markov models for complex action recognition. In: Computer vision and pattern recognition, proceedings. 1997 IEEE computer society conference, IEEE, pp 994–999
65. Ghahramani Z, Jordan MI (1997) Factorial hidden Markov models. Mach Learn 29:245–273
66. Dardas N, Chen Q, Georganas ND, Petriu EM (2010) Hand gesture recognition using bag-of-features and multi-class support vector machine. In: Haptic audio-visual environments and games (HAVE), 2010 IEEE international symposium, IEEE, pp 1–5
67. Pu Q, Gupta S, Gollakota S, Patel S (2013) Whole-home gesture recognition using wireless signals. In: Proceedings of the 19th annual international conference on Mobile computing and networking, ACM, pp 27–38
68. Vogler C, Metaxas D (1998) ASL recognition based on a coupling between HMMs and 3D motion analysis. In: computer vision, 1998. Sixth international conference, IEEE, pp 363–369
69. Karami A, Zanj B, Sarkaleh AK (2011) Persian sign language (PSL) recognition using wavelet transform and neural networks. Expert Syst Appl 38:2661–2667
70. Zaki MM, Shaheen SI (2011) Sign language recognition using a combination of new vision based features. Pattern Recognit Lett 32:572–577
71. Vogler C, Metaxas D (1997) Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods. In: Systems, Man, and Cybernetics, Computational cybernetics and simulation. 1997 IEEE international conference, IEEE, pp 156–161
72. Gavrila DM (1999) The visual analysis of human movement: A survey. Comput Vis Image Underst 73:82–98
73. Zhang X, Chen X, Li Y, Lantz V, Wang K, Yang J (2011) A framework for hand gesture recognition based on accelerometer and EMG sensors. IEEE Trans Syst Man Cybern Part A Syst Hum 41:1064–1076
74. Kainz O, Jakab F (2014) Approach to hand tracking and gesture recognition based on depth-sensing cameras and EMG monitoring. Acta Inf Prag 3:104–112
75. Vyas KK, Pareek A, Tiwari S (2013) Gesture recognition and control. Int J Recent Innov Trends Comput Commun 1(7):575–581
76. Kurdyumov R, Ho P, Ng J (2011) Sign language classification using webcam images
77. Wong S-F, Cipolla R (2005) Real-time adaptive hand motion recognition using a sparse Bayesian classifier. In: Int Workshop Hum Comput Interact, Springer, pp 170–179
78. Von Agris U, Kraiss KF (2007) Towards a video corpus for signer-independent continuous sign language recognition. Gesture Hum Comput Interact Simul, Lisbon
79. Zhang H, Wang Y, Deng C (2011) Application of gesture recognition based on simulated annealing BP neural network. In: Electronic and mechanical engineering and information technology (EMEIT), 2011 international conference, IEEE, pp 178–181
80. Molchanov P, Gupta S, Kim K, Kautz J (2015) Hand gesture recognition with 3D convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 1–7
81. Liu N, Lovell BC (2003) Gesture classification using hidden Markov models and viterbi path counting. In: VIIth digital image computing: techniques and applications
82. Barros PV, Júnior NT, Bisneto JM, Fernandes BJ, Bezerra BL, Fernandes SM (2013) An effective dynamic gesture recognition system based on the feature vector reduction for SURF and LCS. In: International conference on artificial neural networks, Springer, pp 412–419
83. Kumar G, Bhatia PK (2014) A detailed review of feature extraction in image processing systems. In: 2014 fourth international conference on advanced computing and communication technologies, IEEE, pp 5–12
84. Stergiopoulou E, Papamarkos N (2009) Hand gesture recognition using a neural network shape fitting technique. Eng Appl Artif Intell 22:1141–1158
85. Graham J, Starzyk JA (2008) A hybrid self-organizing neural gas based network. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), IEEE, pp 3806–3813
86. Rybach D, Ney IH, Borchers J, Deselaers D-IT (2006) Appearance-based features for automatic continuous sign language recognition. Master's thesis, Human Language Technology and Pattern Recognition Group. RWTH Aachen University, Aachen
87. Hong P, Turk M, Huang TS (2000) Gesture modeling and recognition using finite state machines. In: Automatic face and gesture recognition proceedings. fourth IEEE international conference, IEEE, pp 410–415

88. Bhuyan MK, Ramaraju VV, Iwahori Y (2014) Hand gesture recognition and animation for local hand motions. Int J Mach Learn Cybern 5:607–623
89. Baranwal N, Nandi G (2017) An efficient gesture based humanoid learning using wavelet descriptor and MFCC techniques. Int J Mach Learn Cybern 8(4):1369–1388
90. Bukhari J, Rehman M, Malik SI, Kamboh AM, Salman A (2015) American sign language translation through sensory glove; signspeak. Int J u-e-Serv Sci Technol 8
91. Sethi A, Hemanth S, Kumar K, Bhaskara Rao N, Krishnan R (2012) SignPro—an application suite for deaf and dumb. IJCSET: 1203–1206
92. Abdelnasser H, Youssef M, Harras KA (2015) Wigest: a ubiquitous wifi-based gesture recognition system. In: 2015 IEEE conference on computer communications (INFOCOM, IEEE, pp 1472–1480
93. Wan Q, Li Y, Li C, Pal R (2014) Gesture recognition for smart home applications using portable radar sensors. In: 2014 36th annual international conference of the IEEE engineering in medicine and biology society, IEEE, pp 6414–6417
94. Murakami K, Taguchi H (1991) Gesture recognition using recurrent neural networks. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, pp 237–242
95. Mohandes M, Deriche M, Liu J (2014) Image-based and sensor-based approaches to Arabic sign language recognition. IEEE Trans Hum Mach Syst 44:551–557
96. Chuan C-H, Regina E, Guardino C (2014) American Sign Language recognition using leap motion sensor. In: Machine learning and applications (ICMLA), 13th international conference, IEEE, pp 541–544
97. Mohandes M, Aliyu S, Deriche M (2014) Arabic sign language recognition using the leap motion controller. In: 2014 IEEE 23rd international symposium on industrial electronics (ISIE), IEEE, pp 960–965
98. Funasaka M, Ishikawa Y, Takata M, Joe K (2015) Sign language recognition using leap motion controller. In: Proceedings of the international conference on parallel and distributed processing techniques and applications (PDPTA), The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), p 263
99. Potter LE, Araullo J, Carter L (2013) The leap motion controller: a view on sign language. In: Proceedings of the 25th Australian computer–human interaction conference: augmentation, application, innovation, collaboration, ACM, pp 175–178
100. Marin G, Dominio F, Zanuttigh P (2014) Hand gesture recognition with leap motion and Kinect devices. In: 2014 IEEE international conference on image processing (ICIP), IEEE, pp 1565–1569
101. Shukla J, Dwivedi A (2014) A method for hand gesture recognition. In: Communication systems and network technologies (CSNT), 2014 fourth international conference, IEEE, pp. 919–923
102. Maisto M, Panella M, Liparulo L, Proietti A (2013) An accurate algorithm for the identification of fingertips using an RGB-D camera. IEEE J Emerg Sel Top Circuits Syst 3(2):272–83
103. Yeo HS, Lee BG, Lim H (2015) Hand tracking and gesture recognition system for human–computer interaction using low-cost hardware. Multimed Tools Appl 74(8):2687–715.
104. Tofighi G, Monadjemi SA, Ghasem-Aghaee N (2010) Rapid hand posture recognition using adaptive histogram template of skin and hand edge contour. In: 2010 6th Iranian conference on machine vision and image processing, IEEE, pp. 1–5
105. Han G, Choi H (2014) MPEG-U based advanced user interaction interface system using hand posture recognition. In: 16th international conference on advanced communication technology, IEEE, pp. 512–517
106. Keskin C, Kıraç F, Kara YE, Akarun L (2013) Real time hand pose estimation using depth sensors. In: Consumer depth cameras for computer vision 2013, Springer, London, pp 119–137
107. Billiet L, Mogrovejo O, Antonio J, Hoffmann M, Meert W, Antanas L (2013) Rule-based hand posture recognition using qualitative finger configurations acquired with the Kinect. In: Proceedings of the 2nd international conference on pattern recognition applications and methods, pp 1–4
108. Mo Z, Neumann U (2006) Real-time hand pose recognition using low-resolution depth images. CVPR 2:1499–1505
109. Vančo M, Minárik I, Rozinaj G (2012) Gesture identification for system navigation in 3D scene. In: ELMAR, 2012 proceedings, IEEE, pp 45–48
110. Ganapathyraju S (2013) Hand gesture recognition using convexity hull defects to control an industrial robot. In: Instrumentation control and automation (ICA), 2013 3rd international conference, IEEE, pp. 63–67
111. Manresa C, Varona J, Mas R, Perales FJ (2005) Hand tracking and gesture recognition for human–computer interaction. ELCVIA Electron Lett Comput Vis Image Anal 5(3):96–104
112. Lahiani H, Elleuch M, Kherallah M (2015) Real time hand gesture recognition system for android devices. In: Intelligent systems design and applications (ISDA), 2015 15th international conference, IEEE, pp. 591–596
113. Tariq M, Iqbal A, Zahid A, Iqbal Z, Akhtar J (2012) Sign language localization: learning to eliminate language dialects. In: Multitopic conference (INMIC), 2012 15th international, IEEE, pp 17–22
114. Pedersoli F, Benini S, Adami N, Leonardi R (2014) XKin: an open source framework for hand pose and gesture recognition using kinect. Vis Comput 30(10):1107–1122
115. Shaik KB, Ganesan P, Kalist V, Sathish BS, Jenitha JM (2015) Comparative study of skin color detection and segmentation in HSV and YCbCr color space. Procedia Comput Sci 57:41–48
116. Kaur A, Kranthi BV (2012) Comparison between YCbCr color space and CIELab color space for skin color segmentation. IJAIS 3(4):30–3
117. Tsagaris A, Manitsaris S (2013) Colour space comparison for skin detection in finger gesture recognition. Int J Adv Eng Technol 6(4):1431
118. Qiu-yu Z, Jun-chi L, Mo-yi Z, Hong-xiang D, Lu L (2015) Hand gesture segmentation method based on YCbCr color space and K-means clustering. Interaction 8:106–16
119. Kaur G, Kaur P. Face recognition using YCbCr and CIElab skin color segmentation methods: a review
120. Sun HM (2010) Skin detection for single images using dynamic skin color modeling. Pattern Recognit 43(4):1413–1420
121. Zahedi M, Gorgan I (2007) Robust appearance based sign language recognition, Doctoral dissertation. RWTH Aachen University
122. Dreuw P, Forster J, Ney H (2010) Tracking benchmark databases for video-based sign language recognition. In: European conference on computer vision, Springer, Berlin, pp 286–297
123. Dreuw P, Stein D, Ney H (2007) Enhancing a sign language translation system with vision-based features. In: International gesture workshop, Springer, Berlin, pp 108–113
124. Kak AC (2002) Purdue RVL-SLLL ASL database for automatic recognition of American sign language. In: Proceedings of the 4th IEEE international conference on multimodal interfaces, IEEE Computer Society, pp. 167
125. Forster J, Schmidt C, Hoyoux T, Koller O, Zelle U, Piater JH, Ney H (2012) RWTH-PHOENIX-weather: a large vocabulary sign language recognition and translation corpus. In: LREC, pp. 3785–3789

126. Dreuw P, Rybach D, Deselaers T, Zahedi M, Ney H (2007) Speech recognition techniques for a sign language recognition system. Hand 60:80
127. Bungeroth J, Stein D, Dreuw P, Ney H, Morrissey S, Way A, van Zijl L (2008) The ATIS sign language corpus
128. Dreuw P, Neidle C, Athitsos V, Sclaroff S, Ney H (2008) Benchmark databases for video-based automatic sign language recognition. LREC
129. Stein D, Dreuw P, Ney H, Morrissey S, Way A (2007) Hand in hand: automatic sign language to English translation
130. Zahedi M, Keysers D, Ney H (2005) Pronunciation clustering and modeling of variability for appearance-based sign language recognition. In: International gesture workshop, Springer, Berlin, pp. 68–79
131. Yasir R, Khan RA (2014) Two-handed hand gesture recognition for Bangla sign language using LDA and ANN. In: Software, knowledge, information management and applications (SKIMA), 2014 8th international conference, IEEE, pp 1–5
132. Suriya M, Sathyapriya N, Srinithi M, Yesodha V (2016) Survey on real time sign language recognition system: an LDA approach. In: International conference on exploration and innovations in engineering and technology, ICEIET, pp. 219–225
133. Nummiaro K, Koller-Meier E, Van Gool L (2003) An adaptive color-based particle filter. Image Vis Comput 21(1):99–110
134. Shan C, Wei Y, Tan T, Ojardias F (2004) Real time hand tracking by combining particle filtering and mean shift. In: Automatic face and gesture recognition, 2004. Proceedings. Sixth IEEE international conference, IEEE, pp. 669–674
135. Bretzner L, Laptev I, Lindeberg T (2002) Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. In: Automatic face and gesture recognition, 2002. Proceedings. Fifth IEEE international conference, IEEE, pp. 423–428
136. Kakumanu P, Makrogiannis S, Bourbakis N (2007) A survey of skin-color modeling and detection methods. Pattern Recognit 40(3):1106–1122
137. Li P, Zhang T, Pece AE (2003) Visual contour tracking based on particle filters. Image Vis Comput 21(1):111–123
138. Czyz J, Ristic B, Macq B (2007) A particle filter for joint detection and tracking of color objects. Image Vis Comput 25(8):1271–1281
139. Shan C, Tan T, Wei Y (2007) Real-time hand tracking using a mean shift embedded particle filter. Pattern Recognit 40(7):1958–1970
140. Naik GR, Acharyya A, Nguyen HT (2014) Classification of finger extension and flexion of EMG and Cyberglove data with modified ICA weight matrix. In: 2014 36th annual international conference of the IEEE engineering in medicine and biology society, IEEE, pp. 3829–3832
141. Huong TN, Huu TV, Le Xuan T (2015) Static hand gesture recognition for Vietnamese sign language (VSL) using principle components analysis. In: 2015 International conference on communications, management and telecommunications (ComManTel), IEEE, pp. 138–141
142. Jasim M, Hasanuzzaman M (2014) Sign language interpretation using linear discriminant analysis and local binary patterns. In: Informatics, electronics and vision (ICIEV), 2014 international conference, IEEE, pp 1–5
143. Abhishek KS, Qubeley LC, Ho D (2016) Glove-based hand gesture recognition sign language translator using capacitive touch sensor. In: Electron devices and solid-state circuits (EDSSC), 2016 IEEE international conference, IEEE, pp 334–337
144. Sykora P, Kamencay P, Hudec R (2014) Comparison of SIFT and SURF methods for use on hand gesture recognition based on depth map. AASRI Procedia 9:19–24
145. Hartanto R, Susanto A, Santosa PI (2014) Real time static hand gesture recognition system prototype for Indonesian sign language. In: Information technology and electrical engineering (ICITEE), 2014 6th international conference, IEEE, pp 1–6
146. Gupta B, Shukla P, Mittal A (2016) K-nearest correlated neighbor classification for Indian sign language gesture recognition using feature fusion. In: 2016 international conference on computer communication and informatics (ICCCI), IEEE, pp 1–5
147. Bastos IL, Angelo MF, Loula AC (2015) Recognition of Static Gestures applied to Brazilian Sign Language (Libras). In: 2015 28th SIBGRAPI conference on graphics, patterns and images, IEEE, pp 305–312
148. Ding L, Martinez AM (2009) Modelling and recognition of the linguistic components in American sign language. Image Vis Comput 27(12):1826–1844
149. Pan TY, Lo LY, Yeh CW, Li JW, Liu HT, Hu MC (2016) Real-time sign language recognition in complex background scene based on a hierarchical clustering classification method. In: Multimedia big data (BigMM), 2016 IEEE second international conference, IEEE, pp 64–67
150. Gabriel J, Marcelo J, Figueiredo LS, Teichrieb V (2016) Evaluating sign language recognition using the Myo Armband. In: Virtual and augmented reality (SVR), 2016 XVIII symposium, IEEE, pp 64–70