Vision-based Pakistani Sign Language Recognition Using Bag-of-Words and Support Vector Machines

Muhammad Shaheer Mirza (shaheer.mirza@zu.edu.pk)
Ziauddin University
Sheikh Muhammad Munaf
Ziauddin University
Shahid Ali
Ziauddin University
Fahad Azim
Ziauddin University
Saad Jawaid Khan
Ziauddin University

Research Article

Keywords: Urdu Sign Language, Gesture Recognition, Sign Language Dataset, Pattern Recognition, Image Processing, Machine Learning

Posted Date: January 3rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1204236/v1

Vision-based Pakistani Sign Language recognition using Bag-of-Words and Support Vector Machines

Muhammad Shaheer Mirza*[1], Sheikh Muhammad Munaf[2], Shahid Ali[3], Fahad Azim[4], Saad Jawaid Khan[1]

[1] Department of Biomedical Engineering, Faculty of Engineering, Science, Technology and Management, Ziauddin University, Karachi, Pakistan.

[2] Department of Software Engineering, Faculty of Engineering, Science, Technology and Management, Ziauddin University, Karachi, Pakistan.

[3] Department of Speech Language and Hearing Sciences, Faculty of Health Sciences, Ziauddin University, Karachi, Pakistan.

[4] Department of Electrical Engineering, Faculty of Engineering, Science, Technology and Management, Ziauddin University, Karachi, Pakistan.

***Corresponding author: Muhammad Shaheer Mirza (e-mail: shaheer.mirza@zu.edu.pk)**

Abstract

In order to perform their daily activities, a person is required to communicating with others. This can be a major obstacle for the deaf population of the world, who communicate using sign languages (SL). Pakistani Sign Language (PSL) is used by more than 250,000 deaf Pakistanis. Developing a SL recognition system would greatly facilitate these people. This study aimed to collect data of static and dynamic PSL alphabets and to develop a vision-based system for their recognition using Bag-of-Words (BoW) and Support Vector Machine (SVM) techniques. A total of 5,120 images for 36 static PSL alphabet signs and 353 videos with 45,224 frames for 3 dynamic PSL alphabet signs were collected from 10 native signers of PSL. The developed system used the collected data as input, resized the data to various scales and converted the RGB images into grayscale. The resized grayscale images were segmented using Thresholding technique and features were extracted using Speeded Up Robust Feature (SURF). The obtained SURF descriptors were clustered using K-means clustering. A BoW was obtained by computing the Euclidean distance between the SURF descriptors and the clustered data. The codebooks were divided into training and testing using 5-fold cross validation. The highest overall classification accuracy for static PSL signs was 97.80% at 750×750 image dimensions and 500 Bags. For dynamic PSL signs a 96.53% accuracy was obtained at 480×270 video resolution and 200 Bags.

Keywords: Urdu Sign Language, Gesture Recognition, Sign Language Dataset, Pattern Recognition, Image Processing, Machine Learning

1. Introduction

In today's fast-growing world, communication is key, whether it is communication between different machines, between people or both of them combined. A person cannot perform their everyday tasks without communicating with others. This poses a major problem for the deaf population of the world. According to the World Health Organization, around 466 million people worldwide have disabling hearing loss, which are estimated to increase to over 900 million people by 2050 (1).

The deaf people rely on sign languages (SL), native to their countries, to communicate with others and this is an issue that still remains because not all people are familiar with their local sign languages. Researchers around the world have been working to bridge this communication gap between the deaf and the normal population and have come up with a solution, i.e., automated sign language recognition systems.

According to the Pakistan Association of the Deaf, there are approximately 250,000 hearing-impaired Pakistanis (2), and many of them use Pakistani Sign Language (PSL) as a medium of communication. Developing a SL recognition system would be greatly beneficial for these people. In all the studies mentioned in the next section, only a few have used PSL in their SL recognition systems which means that vision-based Pakistani SL recognition is still a relatively unexplored area of research.

The studies mentioned in the literature review, give us the overall layout of all the techniques used for various SL recognition systems. These techniques can be explored for developing PSL recognition systems. Vision-based PSL alphabets' datasets consisting of bare-handed images and videos, i.e., without any sensors, are not publicly available so researchers have to collect their own dataset in order to perform their studies. The datasets that are available either use sensors to detect PSL signs or are of PSL words. The proposed system will use image for static (still) signs and

videos for dynamic (signs that require movement of the hand) signs of PSL alphabets. All previous PSL studies only focused on static PSL alphabets and none have used dynamic PSL alphabets and only dynamic PSL words have previously been classified. Feature extraction techniques such as SURF, have not yielded good accuracies while being used with SVM and Bag-of-Words (BoW) technique has yet to be applied on vision-based PSL recognition systems.

Therefore, a vision-based PSL alphabets recognition system will be developed in this study, that will form BoW using SURF features and K-means clustering and classify the obtained codebooks of static and dynamic PSL alphabets using Support Vector Machines.

The objectives of this research are as following:

i. To create a dataset containing static and dynamic PSL alphabets, with uniform background and lighting conditions.
ii. To develop a vision-based system for the recognition of Pakistani Sign Language (PSL) alphabets using Bag-of-Words (BoW) and Support Vector Machine (SVM) techniques.

The paper is organized as follows: Section 2 explains the methods used for the literature review and the related studies obtained; Section 3 describes the approach in this study, including the data collection protocol used, and the techniques used for the recognition of PSL alphabets; Section 4 provides the experimental results and their discussion; Section 5 concludes this paper.

2. Literature Review

Several studies have been performed to develop SL recognition systems using different image processing and learning methods. Most of these studies extract specific features and then use machine learning algorithms to classify the SL images. Many different SL have been used in these studies, namely American SL[1-8], Arabic SL[9-12], British SL[13], Chinese SL[14], German SL[15,16], Indian SL[17], Irish SL[18], Pakistani SL[19-24], Persian SL[25], and more in combination such as American & German SL[26], American & Thai SL[27] and American & Indian SL[28].

The literature review done of the SL mentioned, was focused between the time period of 2010 and 2021. Instead of sensor-based recognition systems, i.e., systems that use Cyber-gloves, leap motion controller, accelerometers or EMG sensors, vision-based SL recognition systems were focused. Specifically, those systems that used images and videos from a single camera of bare hands, instead of those that used multiple cameras or different object tracking technologies, were used for this study.

Many systems used a combination of image and video-based datasets as input and used different classifiers, such as, Neural Networks like Convolutional Neural Network (CNN) and Multilayer Perceptron (MLP), Support Vector Machine (SVM), K Nearest Neighbor (KNN), Hidden Markov Model (HMM), etc.

Zadghorban, e.t.al, used dynamic Persian SL with shape-based and motion-based features, DTW to compare features, and HMM to classify motion features with 92.4% accuracy, and a hybrid of KNN-DTW to classify hand shape features to obtain 92.3% accuracy[25]. Dynamic American and Thai SL were used by Klomsae, e.t.al, with Scale Invariant Feature Transform (SIFT) and String Grammar Unsupervised Possibilistic C-Median (sgUPCMed) to obtain accuracies of 90.85% for signer-dependent Thai SL and 91.35% using RWTH-BOSTON-50 dataset using Fuzzy KNN (FKNN)[27].

Static Irish SL was used by Kelly, e.t.al, with weighted eigenspace size function & Hu moments used as features and SVM used to achieve a classification accuracy of 97.3%[18]. Joshi, e.t.al, used static American and Indian SL, using shape-based features and using SVM obtained accuracies of 98.6% using Indian SL with uniform background, and 98.8% using Jochen–Triesch static hand posture with uniform background datasets[28]. Athira, e.t.al, used Indian SL with Zernike moments and centroid of signs to recognize static signs with 90.1% and dynamic signs with 89% accuracies using SVM[17]. Dardas, e.t.al, used the Bag-of-features technique with Scale Invariant Feature Transform (SIFT) and SVM to achieve 96.23% accuracy using 10 signs of static American SL[7]. Singha, e.t.al, used dynamic American SL and features including location, position, velocity, acceleration, orientation, distance and many more to obtain an accuracy of 92.23% using a fusion of classifiers like KNN, SVM and Artificial NN[4].

The literature review was done for Pakistani SL (PSL) to identify the protocols used for the collection of data for static and dynamic PSL alphabets and the methods used for the recognition of PSL alphabets. The protocol used by the researchers of all the included PSL studies used RGB images and single-handed static signs of PSL alphabets except for Saqib, e.t.al, who used dynamic PSL words[24]. The studies used various lighting conditions and studies by Kausar, e.t.al[19], and Shah, e.t.al[23], mentioned that the clothing should be separate from the skin colour of the participant. Khan, e.t.al[22], and Ahmed, e.t.al[21], used complex backgrounds to collect the data while the rest used uniform backgrounds.

Khan, e.t.al, collected a total of 500 (426 training / 74 testing) images of 37 PSL alphabets, converted the RGB images to grayscale, segmented based on skin colour, resized the images to 300×400 pixels, applied Discrete Wavelet Transform (DWT) to extract features and achieved 84.6% classification accuracy using MLP[22]. Ahmed, e.t.al, used 10 PSL alphabets and collected 600 (360 training / 240 testing) images from 60 participants, resized them to 640×480, used ROI segmentation in HSV color space to extract skin pixels, extracted global features including length, area, rectangularity, eccentricity, and more and shape features and used multi-class SVM to obtain an 83% accuracy[21]. 80% accuracy was obtained by Kausar, e.t.al, using 37 Urdu alphabets & 9 numbers, totaling to 455 images (245 training / 210 testing), K-means clustering based segmentation, centroid distance signature in mathematical modelling (polynomial, sinusoidal, exponential, gaussian) and KNN[19]. Multiclass SVM was used by Shah, e.t.al, to achieve 77.18% accuracy, with six statistical features of local binary pattern histogram i.e., standard deviation, variance, skewness, kurtosis, entropy and energy, with skin detection being done in HSV domain from 3,414 images (2,384 training / 1,030 testing), using 37 PSL alphabets[20].

Saqib, e.t.al, used 20 dynamic PSL words, with 8,000 videos (6,480 training /1,520 testing) collected from 15 participants, resized the images to 234×234 and converted them to grayscale, and used CNN with Convolution layers and fully connected layers, along functional layers such as max pooling Layers, Rectified Linear Units layer (ReLU layer) and SoftMax activation function to achieve a 90.79% accuracy[24]. Shah, e.t.al, classified 36 PSL alphabets, with 6,633 images (4,643 training /1,990 testing) collected from 6 participants using SVM and using K-means clustering-based segmentation and converting them to grayscale, obtained classification accuracies of 15.41% using Speeded Up Robust Features (SURF), 87.67% using Edge Orientation Histogram (EOH), 45.71% using Local Binary Patterns (LBP), and 89.52% using Histogram of Oriented Gradient (HOG) and the final reported accuracy of 91.98%[23].

3. Methodology

The methodology for this study was divided into 2 parts:

- Data Collection and,
- Data Analysis.

Data Collection

The data was collected for this study over the course of three months at Ziauddin College of Speech Language and Hearing Sciences, Ziauddin University, Clifton, Karachi. The data collection protocols were approved by the Ziauddin University Ethical Review Committee (Reference Code: 4611221SJBME) and the data collected was in accordance to their guidelines and regulations. Native signers of PSL were selected as participants for this study, irrespective of their race, gender, age, height and skin colour and their written informed consent was obtained. The protocol used for the collected data is mentioned in ***Table 1***. A total of 39 signs of PSL alphabets were collected for this study, i.e., 36 static signs and 3 dynamic signs, as specified in the ***Figure 1*** and ***Figure 2***, respectively.

Table 1 – PSL data collection protocol

Parameters	Our Dataset
Imaging Technique used	48MP smart phone camera
Image Dimensions	3000×3000
Video Resolution and Frames per second	1920×1080 (1080p) at 60fps
Image and Video Type	RGB
Hands used in performing signs	One Hand
Static Signs	Images of the hand
Dynamic Signs	Videos of the signer

Clothing Requirements	Uniform clothing for all the participants
Lighting Conditions	Uniform lighting
Background Conditions	Uniform background
Total Number of Signs	36 Static Urdu alphabets + 3 Dynamic Urdu alphabets
Number of Images / Videos	At least 10 samples per sign per participant
Number of Participants	10
Selection of Participants	Native PSL users who can perform the required signs

The participants were provided with a black lab coat to keep the same clothing conditions and asked to stand in front of the camera with black background. A separate white light source was attached with the camera with uniform intensity for all the participants. The height and the distance between the camera and the participant were not constant. The participants were then asked to perform the signs as they naturally would and the images and videos were captured.

Figure 1 – PSL static alphabets

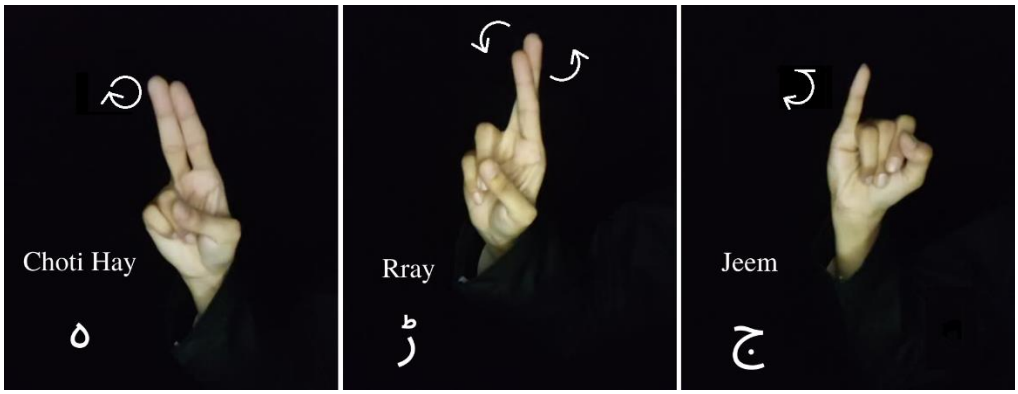

Figure 2– PSL dynamic alphabets

Data Analysis

The images and videos from the collected data were stored in labelled folders. The videos were processed frame by frame, act as static images. The flowchart for the entire data analysis processing is shown in ***Figure 3***.

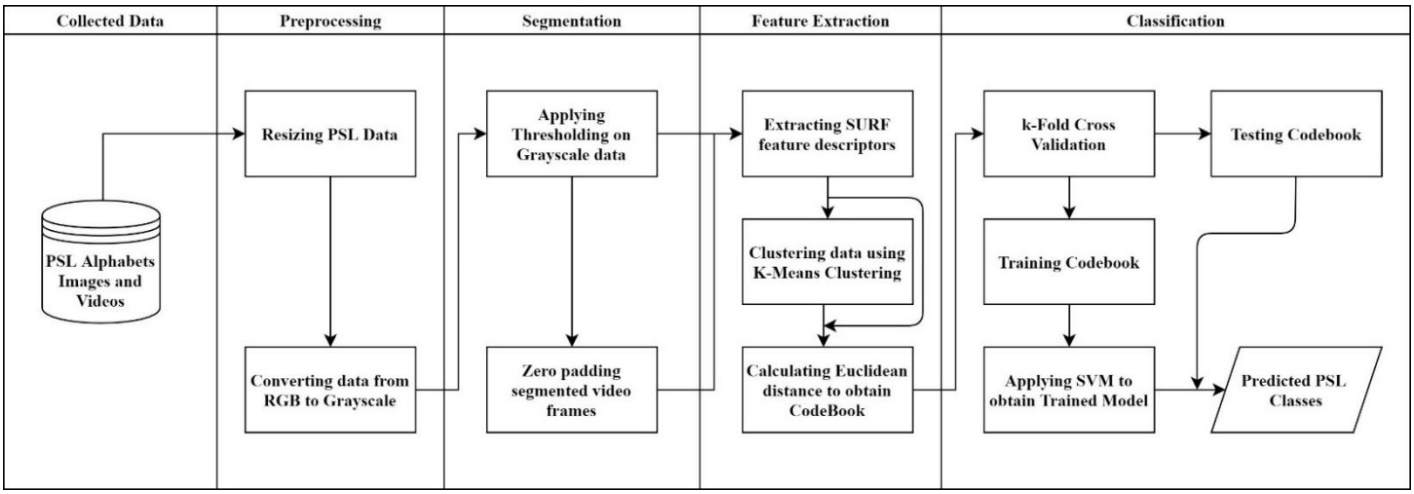

Figure 3 – PSL recognition flowchart

Preprocessing

The collected data was resized to different scales of the original images and videos, i.e., 0.125 (375×375), 0.25 (750×750), 0.375 (1,225×1,225) and 0.5 (1,500×1,500) for images and 0.125 (240×135), 0.25 (480×270) and 0.375 (720×405) for videos. Once the images were resized, they were converted from RGB to grayscale, in order to reduce their complexity and computation time.

Segmentation

The hand sign was detected by applying a threshold on the grayscale images whose value was set low enough to capture all the skin components in that image. The black background and the black clothing conditions facilitated this process of thresholding.

To crop the segmented hand sign, the bounding box technique was used. The thresholded signs were bounded in boxes and their areas were calculated. The bounded box that had the largest area in the image was cropped from each image and saved as the segmented image. The segmented images obtained were of different dimensions, according to the signs being performed in the images. For videos, zero padding was applied to obtain a uniform resolution size for segmented videos.

Feature Extraction

The SURF algorithm was applied on the images to extract their SURF features. The SURF points were detected for each image and then these points were used to extract the key point descriptors which are also called the SURF features.

The SURF algorithm is based on the Hessian matrix [29], because of its better performance in the required computation time and the overall detection accuracy. It relies on the determinant of Hessian for the selection of both, the scale and the location. Given a point $x = (x, y)$ in an image *I,* the Hessian matrix $H(x, \sigma)$ in x at scale σ is defined as follows

$$H(x,\sigma) = \begin{bmatrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma) \end{bmatrix} \quad (1)$$

where, $L_{xx}(x, \sigma)$ is the convolution of Gaussian second order derivative $\frac{\partial^2}{\partial x^2} g(\sigma)$ with the image I in point x, and similarly for $L_{xy}(x, \sigma)$ and $L_{yy}(x, \sigma)$.

The key point descriptors in SURF were detected by first, constructing a circular region around the key points and then computing the Haar-wavelet responses in both x and y directions to get the orientation. Then using this orientation, a square region was constructed around the interest points. The square regions were split into 4×4 sub regions, to contain the relevant spatial information. Haar-wavelet responses d_x and d_y were weighted with a Gaussian centered at the interest point and summed over each sub region. The sum of the absolute values of the responses were also calculated $|d_x|$ and $|d_y|$, to extract information about the polarity of intensity changes. With this, each sub region had a four-dimensional descriptor vector,

$$v = \left(\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|\right) \quad (2)$$

This produced the standard SURF descriptor of length 64 for all 4×4 sub regions.

These extracted features of all the images were then clustering using unsupervised learning algorithm, K-means++ clustering. The k-means++ algorithm uses a heuristic method to find centroid seeds[30].

The algorithm chooses seeds as follows, assuming the number of clusters is k. It then selects a descriptor at random from the images features dataset, X. The chosen descriptor is the first centroid, and is denoted c_1. It then computes the distances from each descriptor to c_1. The distance between c_j and the descriptor k as is denoted as $d(x_m, c_j)$. Then it selects the next centroid, c_2 at random from X with probability

$$\frac{d^2(x_m, c_1)}{\sum_{j=1}^{n} d^2(x_j, c_1)} \quad (3)$$

In order to choose center j, it computes the distances from each descriptor to each centroid, and assign each descriptor to its closest centroid. For $m = 1, \dots, n$ and $p = 1, \dots, j-1$, it selects the centroid j at random from X with probability

$$\frac{d^2(x_m, c_p)}{\sum_{\{h; x_h \in C_p\}} d^2(x_h, c_h)} \quad (4)$$

where C_p is the set of all descriptor closest to centroid c_p and x_m belongs to C_p, i.e., it selects each subsequent center with a probability proportional to the distance from itself to the closest center that was already chosen. The process to choose the center j, is repeated until k centroids are chosen.

A set of K-cluster values were used to form Bags (clusters) for the extracted features and each Bag is called a visual word. A set of these Bags form the visual vocabulary which are in-turn used to form the codebook or Bag-of-words. To select the K-cluster values for Bag formation, the maximum number of SURF descriptors were found for each scale of images and videos used, which were 90, 202, 307 and 444 for 375×375 (0.125), 750×750 (0.250), 1,225×1,225 (0.375) and 1,500×1,500 (0.500), image dimensions (scale), respectively, for static signs and 84 for all video resolutions (scale) used i.e., 240×135 (0.125), 480×270 (0.250), 720×405 (0.375) for dynamic signs. Using these maximum descriptors, 500 K-cluster value (Bag) was selected for static signs and 200 K-cluster value (Bag) was selected for dynamic signs.

An empty codebook was used to start the process. The Euclidean distance between each surf descriptor or feature and the centroid for each Bag and the feature was calculated. The least value of Euclidean distance was then assigned to the codebook as a part of that Bag using the formula,

$$d(x_i, c_i) = \sqrt{\sum (x_i - c_i)^2} \tag{5}$$

Where, $d(x_i, c_i)$ is the distance between and the descriptor x_i and the centroids c_i.

The same procedure was repeated until each and every feature of all the images was assigned a Bag. If a specific Bag matched with more than one descriptor, the number of descriptors were added up. The final codebook obtained contained the number of features that each centroid had the least distance with, or the number of times each centroid was activated. The codebook obtained had the dimensions of the K-cluster value used and the total number of images. The labels for each image were then added to the codebook. This process of generating the codebook is shown in ***Figure 4***. The obtained codebook was then used for the classification of these images.

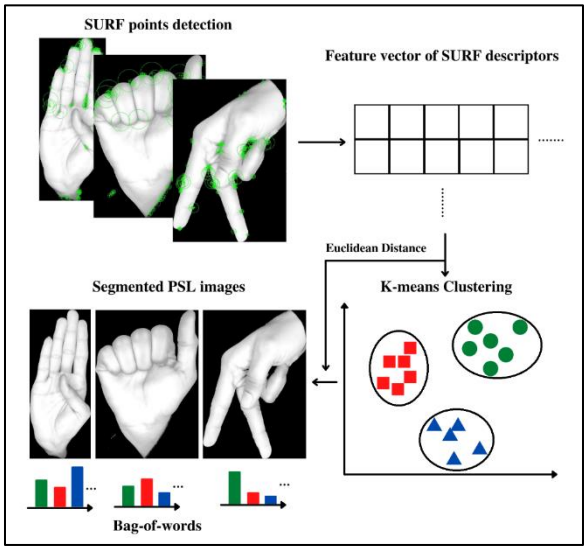

Figure 4 – Bag-of-words generation

Classification

In k-fold cross-validation, the dataset being was partitioned into k disjoint subsets, known as folds, of approximately equal size. This partitioning is randomly performed by sampling the dataset without replacement.

The Support Vector Machine classifier (SVM) was used for classification. SVM used a part of the partitioned dataset, the training set, to find the optimal separating hyperplane between classes of the training data. The feature vectors near the hyperplane, the support vectors, are shown in ***Figure 5***. The SVM classifier used the training dataset to build a model that predicted whether the given example fell into one class of the target variable or the other.

The value of $k = 5$ was chosen for k-fold cross-validation in this study, which partitioned the dataset into 80% for training and 20% for testing. As the dataset was folded five times, five training and five testing datasets were obtained, and the five training datasets were used to train five SVM models.

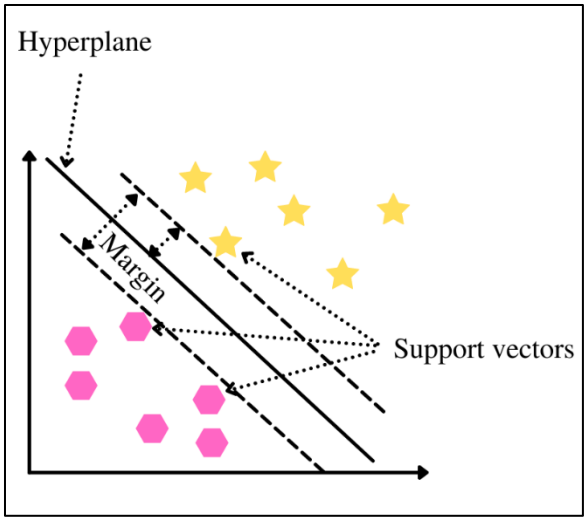

Figure 5 – Margin Optimization

The validation or testing dataset was applied on the trained models, and the performance was measured. This process was repeated until all of the k subsets served as testing sets. The cross-validated accuracy was obtained, by averaging the five accuracies achieved on the test sets. The cross-validated estimate of the prediction error, $\hat{\epsilon}_{cv}$, is then given as

$$\hat{\epsilon}_{cv} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, \hat{f}_{-k}(x_i)) \quad (6)$$

Where, $\hat{f}_{-k}$ is the model trained on all but the k^{th} test subset, $\hat{y}_i = \hat{f}_{-k}(x_i)$ is the predicted value for the real class label, y_i, of case x_i, which is an element of the k^{th} subset [31].

4. Results and Discussion

The samples and details of the data collected per participant are mentioned in ***Table 2***. In this study, 5-fold cross validation was applied on the obtained codebook for static and dynamic signs, yielding five training codebooks and five testing codebooks for each K-cluster value of Bags used. As a size of 500 Bags was used for static signs with four different image scale sizes, as previously mentioned, a total of 20 models were trained for static images. The number of images used in each model were 4,096 for training and 1,024 for testing. The subsequent training and testing accuracies obtained from these 20 models are shown in ***Table 3.*** The overall accuracies were obtained by averaging the training and testing accuracies of each model. The image scale size of 0.250 with 750×750 image dimensions and using 500 Bags yielded the highest overall classification accuracy for static signs of PSL alphabets, i.e., 97.80%. ***Figure 6*** shows the confusion matrix of the testing model, which was obtained by averaging the testing confusion matrices of all the five models.

Table 2 – PSL data per participant and total collected data

Participant	Image Samples Total (Min, Max)	Video Samples Total (Min, Max)	Video Duration in seconds Total (Min, Max)	Video Frames Total (Min, Max)
1	511 (10, 16)	32 (10, 11)	80.26 (1.33, 3.68)	4,755 (78, 216)
2	392 (10, 16)	38 (12, 14)	58.20 (0.71, 2.55)	3,447 (42, 149)
3	423 (10, 15)	34 (10, 12)	69.16 (1.15, 3.57)	4,138 (68, 215)
4	520 (11, 17)	35 (10, 14)	45.24 (0.70, 3.12)	2,700 (40, 187)
5	514 (12, 16)	35 (10, 13)	66.31 (1.29, 3.07)	3,915 (78, 185)
6	547 (15, 16)	40 (13, 14)	91.99 (1.08, 3.67)	5,523 (65, 221)
7	548 (15, 16)	36 (11, 13)	91.82 (1.03, 5.09)	5,532 (62, 307)

8	549 (13, 17)	34 (10, 13)	59.09 (0.92, 2.87)	3,560 (55, 173)
9	547 (14, 17)	31 (10, 11)	61.71 (1.33, 3.22)	3,717 (80, 194)
10	569 (15, 16)	38 (12, 13)	131.69 (1.96, 4.93)	7,937 (118, 297)
Total	5,120	353	755.47	45,224

Table 3 – Classification accuracies for static signs at 500 bags

Image Dimensions (Scale)		**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Overall**
1,500×1,500 (0.500)	Training	94.80	94.90	95.10	95.60	95.10	95.10
	Testing	95.21	96.19	96.09	96.00	96.68	96.03
	Overall	95.01	95.55	95.60	95.80	95.89	95.57
1,225×1,225 (0.375)	Training	96.60	96.60	96.10	96.40	96.10	96.36
	Testing	97.66	97.07	97.66	96.39	98.05	97.37
	Overall	97.13	96.84	96.88	96.40	97.08	96.86
750×750 (0.250)	Training	97.40	97.30	97.80	97.40	97.40	97.46
	Testing	98.24	98.05	97.95	98.73	97.75	98.14
	Overall	97.82	97.68	97.88	98.07	97.58	**97.80**
375×375 (0.125)	Training	96.00	95.80	96.00	95.80	95.90	95.90
	Testing	96.48	96.88	96.29	96.39	96.39	96.49
	Overall	96.24	96.34	96.15	96.10	96.15	96.19

Similarly, a size of 200 Bags was used for dynamic signs with three different video scale sizes, a total of 15 models were trained for dynamic signs. The number of video frames used for training in one model were 36,180 and 36,179 for the other four models and for testing in one model were 9,044 and 9,045 for the other four models. The subsequent training and testing classification accuracies obtained from these 15 models are shown in ***Table 4.*** The video scale size of 0.250 with 480×270 video resolution and using 200 Bags yielded the highest overall classification accuracy for dynamic signs of PSL alphabets, i.e., 96.53%. ***Figure 7*** shows its testing confusion matrix, which was obtained by averaging the testing confusion matrices of all the five models.

Table 4 – Classification accuracies for dynamic signs at 200 Bags

Video Resolution (Scale)		**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Overall**
720×405 (0.375)	Training	96.10	96.90	96.30	96.20	96.40	96.38
	Testing	96.93	96.45	96.21	97.00	96.40	96.60
	Overall	96.52	96.68	96.26	96.60	96.40	96.49
480×270 (0.250)	Training	96.40	96.30	96.30	96.40	96.30	96.34
	Testing	96.56	97.06	96.64	96.54	96.75	96.71
	Overall	96.48	96.68	96.47	96.47	96.53	**96.53**
240×135 (0.125)	Training	96.20	96.20	96.20	96.30	96.10	96.20
	Testing	96.80	96.63	96.67	96.46	96.61	96.63
	Overall	96.50	96.42	96.44	96.38	96.36	96.42

Static PSL Signs		True Classes																																							
		Aa'in	Alif	Bari Ye	Bay	Chay	Choti Ye	Daal	Ddaal	Do Chashmi Hay	Fay	Gaaf	Gha'in	Hamza	Hay	Kaaf	Khay	Laam	Meem	Noon	Noon Ghunna	Pay	Qaaf	Ray	Say	Seen	Sheen	Suaad	Tay	Ttay	Tua'ay	Wow	Zaal	Zay	Zhay	Zua'ay	Zuaad
Predicted Classes	Aa'in	27.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Alif	0.0	26.6	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Bari Ye	0.0	0.0	28.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Bay	0.0	0.0	0.0	27.2	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Chay	0.0	0.0	0.0	0.0	28.2	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Choti Ye	0.0	0.0	0.0	0.0	0.0	28.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Daal	0.0	0.0	0.0	0.0	0.0	0.0	28.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Ddaal	0.0	0.0	0.0	0.0	0.0	0.0	0.0	26.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Do Chashmi Hay	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Fay	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
	Gaaf	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0
	Gha'in	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Hamza	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.2	0.0	28.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Hay	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Kaaf	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
	Khay	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Laam	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	27.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Meem	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.0	0.6	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Noon	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	27.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Noon Ghunna	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Pay	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	26.6	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Qaaf	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Ray	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Say	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Seen	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	28.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Sheen	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	29.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Suaad	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	28.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Tay	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Ttay	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	27.8	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0
	Tua'ay	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Wow	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	28.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Zaal	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	27.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Zay	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.4	0.0	0.0	0.0	0.0	0.0	0.0
	Zhay	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.8	0.0	0.0	0.0	0.0	0.0
	Zua'ay	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	28.2	0.0	0.0	0.0	0.0
	Zuaad	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.8	0.0	0.0	0.0

Figure 6 – Confusion matrix of static PSL signs at 750×750 image dimensions and 500 Bags

Dynamic PSL Signs		True Classes		
		Choti Hay	Jeem	Rray
Predicted Classes	Choti Hay	2888.4	122.0	12.2
	Jeem	125.8	3681.2	8.2
	Rray	19.4	10.0	2177.6

Figure 7 – Confusion matrix of dynamic PSL signs at 480×270 video resolution and 200 Bags

For the collection of data, recruiting participants of different race, gender, age, height and skin colour, added variations to the collected dataset, such as different skin colours, hand size and so on. Asking the participants to perform the hand signs as they naturally would, caused variations in the orientation of the signs being performed, and minor variations due to different joint flexibility of the participants. By varying the height and distance between the camera and the participant according to the participants comfort also added variations in the scale of the data being collected. The data collected only required the hand to be captured. If the data of PSL sentences was captured, also collecting the facial expressions of the participants would increase the complexity of the system being developed.

The black background and clothing conditions helped in the thresholding technique used during segmentation, as the skin colour in grayscale was easily distinguished from the background and clothes. During the video segmentation, all the frames in the video had to be of the same size, in order save them for further processing. This issue was resolved

by applying zero padding to the videos. This was done by finding the maximum dimensions from each video's segmented frames and using that reference value to apply zero padding to the frames with lesser dimensions. This resulted in a uniform resolution size for that specific video. Zero padding was an effective technique for the dataset used in this study as the background chosen for the collected data was black and by applying zero padding black pixels were added to the videos as 0 represents black when the pixels of images are visualized.

A similar study by, Nasser H. Dardas, e.t.al[7], used the Bag-of-features technique with SIFT and SVM to obtain 96.23% accuracy using 10 signs of static American SL with cluttered background. Another study by Farman Shah, e.t.al[23], used SURF with SVM but obtained 15.41% accuracy and the final reported accuracy using Histogram of Oriented Gradient (HOG) and SVM was 91.98%, which was also the highest classification accuracy reported, to the best of my knowledge, using static PSL alphabets. Shazia Saqib, e.t.al[24], used dynamic PSL words with CNN with Levenshtein distance to obtain 90.79% accuracy.

The highest classification accuracy obtained in this study for static PSL signs, was 97.80% as compared to 96.23% by Nasser H. Dardas, e.t.al[7], who used 1,000 static American SL images with non-uniform lighting, background, scale and rotation, and 15.41% by Farman Shah, e.t.al[23], who used SURF directly with SVM, instead of using the BoW technique. No previously performed study has classified dynamic PSL, to the best of my knowledge, so the classification accuracy of 96.53% for dynamic PSL signs cannot be compared to any PSL study.

The limitations of this study were that the dataset collected used only uniform lighting and uniform background conditions and the data was only captured with the participant facing the camera, i.e., only from one angle using their dominant right hand. Furthermore, the system was developed in such a way that it used offline testing along with the offline training.

For future work, a PSL dataset could be created that uses various lighting and complex background conditions. The data of the signer could be captured from multiple angles. More participants can be recruited, to increase the size of the dataset. The system could also be implemented using real-time testing of the trained models. The developed system can be implemented in comparison other sign languages.

5. Conclusion

The purpose of this study was to collect data of static and dynamic PSL alphabets and to develop a vision-based system for their recognition using BoW and SVM techniques. 36 static PSL alphabet signs and 3 dynamic PSL alphabet signs were collected with uniform background, uniform lighting at various orientations and scale, from 10 native signers of PSL and used as input in the developed system. The data was resized to various scales, segmented and converted into Bag-of-Words by finding the Euclidean distance between SURF descriptors and clustered value obtained by K-means clustering. The obtained codebooks were trained using SVM and tested to obtain the highest overall classification accuracy of 97.80% for static PSL signs and 96.53% for dynamic PSL signs.

Author contributions

M.S.M.'s contributions in the study were conceptualization, formal analysis, investigation, software, validation, visualization and original draft preparation. S.M.M.'s contributions in the study were conceptualization, methodology, software, supervision, review and editing. S.A.'s contributions in the study were investigation, visualization and review and editing. F.A.'s contributions in the study were investigation, resources, and review and editing. S.J.K.'s contributions in the study were conceptualization, supervision, visualization, review and editing. All authors have given their approval of this version of article to be published and agree to be accountable for all aspects of this work.

Competing interests

The authors declare no competing interests.

References

1 Ameen, S. & Vadera, S. A convolutional neural network to classify American Sign Language fingerspelling from depth and colour images. *Expert Systems* **34**, doi:10.1111/exsy.12197 (2017).

2 Athitsos, V., Wang, H. & Stefan, A. A database-based framework for gesture recognition. *Personal and Ubiquitous Computing* **14**, 511-526, doi:10.1007/s00779-009-0276-x (2010).

3 Zaki, M. M. & Shaheen, S. I. Sign language recognition using a combination of new vision based features. *Pattern Recognition Letters* **32**, 572-577, doi:10.1016/j.patrec.2010.11.013 (2011).

4 Singha, J., Roy, A. & Laskar, R. H. Dynamic hand gesture recognition using vision-based approach for human-computer interaction. *Neural Computing & Applications* **29**, 1129-1141, doi:10.1007/s00521-016-2525-z (2018).

5 Nasri, S., Behrad, A. & Razzazi, F. Spatio-temporal 3D surface matching for hand gesture recognition using ICP algorithm. *Signal Image and Video Processing* **9**, 1205-1220, doi:10.1007/s11760-013-0558-7 (2015).

6 Hikawa, H. & Kaida, K. Novel FPGA Implementation of Hand Sign Recognition System With SOM-Hebb Classifier. *Ieee Transactions on Circuits and Systems for Video Technology* **25**, 153-166, doi:10.1109/tcsvt.2014.2335831 (2015).

7 Dardas, N. H. & Georganas, N. D. Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques. *Ieee Transactions on Instrumentation and Measurement* **60**, 3592-3607, doi:10.1109/tim.2011.2161140 (2011).

8 Rastgoo, R., Kiani, K. & Escalera, S. Multi-Modal Deep Hand Sign Language Recognition in Still Images Using Restricted Boltzmann Machine. *Entropy* **20**, doi:10.3390/e20110809 (2018).

9 Elons, A. S., Aboul-Ela, M. & Tolba, M. F. 3D object recognition technique using multiple 2D views for Arabic sign language. *Journal of Experimental & Theoretical Artificial Intelligence* **25**, 119-137, doi:10.1080/0952813x.2012.680073 (2013).

10 Elons, A. S., Abull-ela, M. & Tolba, M. F. A proposed PCNN features quality optimization technique for pose-invariant 3D Arabic sign language recognition. *Applied Soft Computing* **13**, 1646-1660, doi:https://doi.org/10.1016/j.asoc.2012.11.036 (2013).

11 Mohandes, M., Deriche, M., Johar, U. & Ilyas, S. A signer-independent Arabic Sign Language recognition system using face detection, geometric features, and a Hidden Markov Model. *Computers & Electrical Engineering* **38**, 422-433, doi:https://doi.org/10.1016/j.compeleceng.2011.10.013 (2012).

12 Ibrahim, N. B., Selim, M. M. & Zayed, H. H. An Automatic Arabic Sign Language Recognition System (ArSLRS). *Journal of King Saud University - Computer and Information Sciences* **30**, 470-477, doi:https://doi.org/10.1016/j.jksuci.2017.09.007 (2018).

13 Han, J., Awad, G. & Sutherland, A. Boosted subunits: a framework for recognising sign language from videos. *IET Image Processing* **7**, 70-80, doi:10.1049/iet-ipr.2012.0273 (2013).

14 Jiang, X. & Zhang, Y.-D. Chinese Sign Language Fingerspelling Recognition via Six-Layer Convolutional Neural Network with Leaky Rectified Linear Units for Therapy and Rehabilitation. *Journal of Medical Imaging and Health Informatics* **9**, 2031-2038, doi:10.1166/jmihi.2019.2804 (2019).

15 Cui, R., Liu, H. & Zhang, C. A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. *IEEE Transactions on Multimedia* **21**, 1880-1891, doi:10.1109/TMM.2018.2889563 (2019).

16 Koller, O., Zargaran, S., Ney, H. & Bowden, R. Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. *International Journal of Computer Vision* **126**, 1311-1325, doi:10.1007/s11263-018-1121-3 (2018).

17 Athira, P. K., Sruthi, C. J. & Lijiya, A. A Signer Independent Sign Language Recognition with Co-articulation Elimination from Live Videos: An Indian Scenario. *Journal of King Saud University - Computer and Information Sciences*, doi:https://doi.org/10.1016/j.jksuci.2019.05.002 (2019).

18 Kelly, D., McDonald, J. & Markham, C. A person independent system for recognition of hand postures used in sign language. *Pattern Recognition Letters* **31**, 1359-1368, doi:10.1016/j.patrec.2010.02.004 (2010).

19 Kausar, S., Javed, M. Y., Tehsin, S. & Anjum, A. A Novel Mathematical Modeling and Parameterization for Sign Language Classification. *International Journal of Pattern Recognition and Artificial Intelligence* **30**, doi:10.1142/s0218001416500099 (2016).

20 Shah, S. M. S. *et al.* Shape Based Pakistan Sign Language Categorization Using Statistical Features and Support Vector Machines. *IEEE Access* **6**, 59242-59252 (2018).

21 Ahmed, H., Gilani, S., Jamil, M., Ayaz, Y. & Shah, S. Monocular Vision-based Signer-Independent Pakistani Sign Language Recognition System using Supervised Learning. *Indian Journal of Science and Technology* **9**, doi:10.17485/ijst/2016/v9i25/96615 (2016).

22 Khan, N. *et al.* A Vision Based Approach for Pakistan Sign Language alphabets Recognition. *La Pensée* **76** (2014).

23 Shah, F. R. *et al.* Sign Language Recognition Using Multiple Kernel Learning: A Case Study of Pakistan Sign Language. *Ieee Access* **9**, 67548-67558, doi:10.1109/access.2021.3077386 (2021).

24 Saqib, S., Ditta, A., Khan, M. A., Kazmi, S. A. R. & Alquhayz, H. Intelligent Dynamic Gesture Recognition Using CNN Empowered by Edit Distance. *Cmc-Computers Materials & Continua* **66**, 2061-2076, doi:10.32604/cmc.2020.013905 (2021).

25 Zadghorban, M. & Nahvi, M. An algorithm on sign words extraction and recognition of continuous Persian sign language based on motion and shape features of hands. *Pattern Analysis and Applications* **21**, 323-335, doi:10.1007/s10044-016-0579-2 (2018).

26 Elakkiya, R. & Selvamani, K. Subunit sign modeling framework for continuous sign language recognition. *Computers & Electrical Engineering* **74**, 379-390, doi:10.1016/j.compeleceng.2019.02.012 (2019).

27 Klomsae, A., Auephanwiriyakul, S. & Theera-Umpon, N. A Novel String Grammar Unsupervised Possibilistic C-Medians Algorithm for Sign Language Translation Systems. *Symmetry-Basel* **9**, doi:10.3390/sym9120321 (2017).

28 Joshi, G., Vig, R. & Singh, S. DCA-based unimodal feature-level fusion of orthogonal moments for Indian sign language dataset. *IET Computer Vision* **12**, 570-577, doi:10.1049/iet-cvi.2017.0394 (2018).

29 Bay, H., Tuytelaars, T. & Van Gool, L. in *COMPUTER VISION - ECCV 2006 , PT 1, PROCEEDINGS* Vol. 3951 (eds A. Leonardis, H. Bischof, & A. Pinz) 404-417 (2006).

30 Arthur, D., Vassilvitskii, S. & Siam/Acm. *k-means plus plus : The Advantages of Careful Seeding*. (2007).

31 Berrar, D. in *Encyclopedia of Bioinformatics and Computational Biology* (eds Shoba Ranganathan, Michael Gribskov, Kenta Nakai, & Christian Schönbach) 542-545 (Academic Press, 2019).