International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5, June 2019

American Sign Language Video Hand Gestures Recognition using Deep Neural Networks

Shivashankara S, Srinath S

Abstract: ***In this paper an effort has been placed to translate / recognize some of the video based hand gestures of American Sign Language (ASL) into human and / or machine readable English text using deep neural networks. Initially, the recognition process is carried out by fetching the input video gestures. In the recognition process of the proposed algorithm, for background elimination and foreground detection, the Gaussian Mixture Model (GMM) is used. The basic preprocessing operations are used for better segmentation of the video gestures. The various feature extraction techniques like, Speeded Up Robust Features (SURF), Zernike Moment (ZM), Discrete Cosine Transform (DCT), Radon Features (RF), and R, G, B levels are used to extract the hand features from frames of the video gestures. The extracted video hand gesture features are used for classification and recognition process in forthcoming stage. For classification and followed by recognition, the Deep Neural Networks (stacked autoencoder) is used. This video hand gesture recognition system can be used as tool for filling the communication gap between the normal and hearing impaired people. As a result of this proposed ASL video hand gesture recognition (VHGR), an average recognition rate of 96.43% is achieved. This is the better and motivational performance compared to state of art techniques.***

Index Terms: American Sign Language, Deep Neural Networks, Hand Gestures Recognition, Radon Features, Stacked Autoencoder, SURF, Zernike Moment

I. INTRODUCTION

The communication is an essential and significant task for every human being. The normal people those who are able to speak and hear can easily and effortlessly communicate with the normal people but the people who are unable to speak and hear are very difficult to communicate with the normal people. Thus, the speech and hearing impaired people are well communicate with the other speech and hearing impaired people by making sign gestures as their communication language. This way of communication is called as Sign Language (SL). There are more than 70 million people in the world and around 10 million people across the India are suffering from speech and hearing disability problem. There are over 108 varieties of Sign languages are present in the universe. For example American Sign Language, Arabic Sign Language, British Sign Language, Egyptian Sign Language, German Sign Language, Korean Sign Language and many more.

Manuscript published on 30 June 2019.
***** Correspondence Author (s)
Shivashankara S, Research Scholar, Dept. of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysuru, India.
Srinath S, Dept. of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysuru, India.

Some countries uses single hand SL and some other countries uses two hand SL. The American Sign Language (ASL) is a most widely used language for communication of over 2 million hearing impaired people in United States of America and Canada. American SL is also used in Mexico, West Africa, Asia, and many other English speaking countries. In universe, more than 20 countries SL like Jamaica, panama, Thai, Malaysia and many more are derived by ASL. The ASL is founded by the American School for Deaf (ASD) by combining Old French Sign Language (OFSL) and some of the native village Sign Languages. The ASL consists of over 1000s of static and dynamic gestures in which some gesture signs are made using single and sometimes double hand sometimes with the facial expressions. ASL is quite notable and eye-opening for its substance, prominence, perspective expectations, and also for overall impression [1][2]. ASL is complex and complete language of communication for hearing impaired people and it has massive varieties of sign gestures of hand, finger, palm, and fist with and without the support of facial expressions and body movements [3]. In ASL, there are no specific sign gestures for several English words and sentences. Thus, those English words and sentences can be signed by spelling out the set of American Manual Alphabets (AMA) of 26 gestures from A to Z. ASL also consists of 10 static gestures of ASL numbers from 0 – 9 are depicted in Fig. 1.

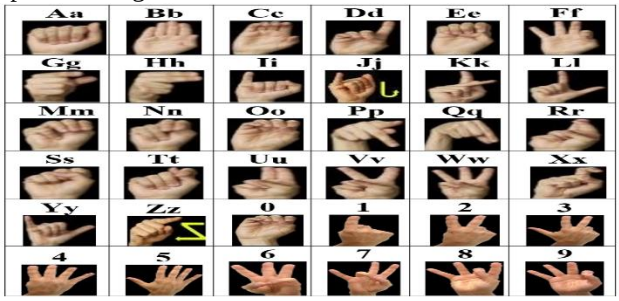


Fig. 1. Set of gestures of ASL Alphabets and Numbers

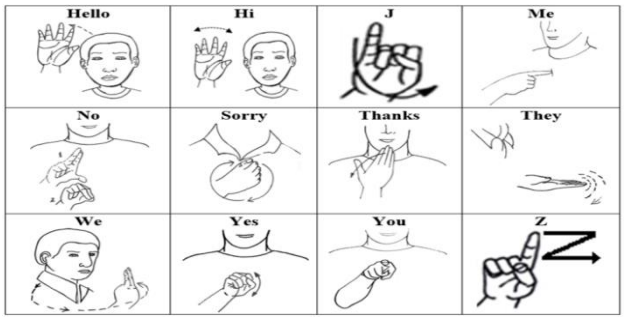


Fig. 2. Some of the Video Hand Gestures.

However, the ASL number from 10 and above involves hand movements i.e., video hand gestures. In ASL Alphabet gestures, the alphabets 'J' and 'Z' involves hand movements. The video hand gestures are of length from 1 to 3 seconds which depends on type of English word. Some of the video hand gestures (Hello, Hi, J, Me, No, Sorry, Thanks, They, We, Yes, You, and Z) of ASL is shown in Fig. 2. It is a tedious task to understand the many of the ASL gestures by the normal people. It is also a cumbersome task for normal people to learn the ASL for communicating with the hearing impaired people. The gesture recognition is the process of interpreting and recognizing human gestures by using computational algorithms. These algorithms are obviously distinct in recognizing the gestures.

The sign language recognition (SLR) is a system or process, in which the computer automatically understands the gestures and interprets them into their equivalent human and or machine recognizable or readable text. An automated SLR systems can be widely used in the applicable places like Industrial-Internet of Things (I-IoT) for Human Computer Interaction (HCI), Human Robot Interaction (HRI), smart homes for controlling the electronic devices by gestures, public places like hospitals, police station, court and other places. Also gesture recognition systems can be used in social assistive robotics, directional indication through pointing, and control through facial gestures, alternative computer interfaces, immersive game technology, virtual controllers, affective computing, and also remote control. The SLR system can also use in educational institutions, training and / or tutorial centers, and special education centers for specially abled children and many more places.

II. DATA COLLECTION

In proposed technique, an effort has been placed to collect the total of 168 video gestures from 14 datasets of video hand gestures of 12 most commonly used of ASL (Gestures mentioned in figure 2) were captured from the various resolution mobile cameras such as 8 and 13 Mega Pixel (MP) mobile cameras, plain and complex background, invariant location (indoor and outdoor), illumination (natural and artificial), signer (male and female), time (day and night) and Distance (considered 5 and 10 feet distances between the signer and camera).

Among these 14 datasets created, seven datasets were used for training purpose and remaining seven datasets were utilized for experimentation purpose. The 7 datasets used for testing, 4 datasets are Male signer datasets and remaining 3 datasets are Female Signer datasets. Also, 2 datasets (Set S1 and S3) are captured in plain background with one dataset using natural lightings and another one using artificial illumination. The remaining 5 datasets are captured in complex background using artificial illumination. From these 7 datasets, 4 datasets are captured from 5 feet distance and remaining 3 are from 10 feet distances. The Set 4 (S4) and Set 5 (S5) are captured at outdoor location in night artificial illumination condition. There are 2 ranges of mobile cameras with 8 MP and 13 MP are used

Table I shows the details of the video gesture datasets used for experimentations of the Proposed Video Gesture Recognition System (PVGRS). The result analysis of these experimentations are discussed in section 5.

Table I. Details of Datasets used for Testing

Set #	Details of the Dataset
S1	Indoor Plain Black Background with Natural Lightings from 10 Feet Distance using 8 MP Mobile Camera by Male Signer
S2	Indoor Complex Background with Artificial Lightings from 5 Feet Distance using 8 MP Mobile Camera by Male Signer
S3	Indoor Plain White Background with Artificial Lightings from 5 Feet Distance using 8 MP Mobile Camera by Female Signer
S4	Outdoor Complex Background with Artificial Lightings from 5 Feet Distance using 13 MP Mobile Camera by Male Signer
S5	Outdoor Complex Background with Artificial Lightings from 10 Feet Distance using 13 MP Mobile Camera by Female Signer
S6	Indoor Complex Background with Artificial Lightings from 5 Feet Distance using 13 MP Mobile Camera by Female Signer
S7	Indoor Complex Background with Artificial Lightings from 10 Feet Distance using 8 MP Mobile Camera by Male Signer

III. RELATED WORK

It is noticed that from the literature review, the very less amount of research works were carried out in recognizing video hand gestures. Some of the research works carried out by the various researchers for recognizing video hand gestures are discussed in this section.

A real time Hand Gesture recognition System for recognizing the twelve ASL gestures consisting Bathroom, Blue, Finish, Green, Hungry, Milk, Past, Pig, Store, Where, Letter J, Letter Z was implemented in using 'Action Graph Method' (AGM) [4]. All the gestures signs were carried by 10 signers 3 times using Microsoft Kinect device.

This recognition system is invariant to hand orientation, illumination, performing style and speed. This system obtains the average recognition rate of 87.7%. A human action recognition system called 'Actionlet Ensemble Model' (AEM) is developed [5] for recognizing 'twenty' human actions such as high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side throw, draw x, draw tick, draw circle, hand clap, two hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up & throw. All these human actions were carried out by 10 signers 3 times using depth cameras. These gestures are robust to noise, invariant to translational and temporal misalignments but background is plain.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

A total of 402 samples of 20 gestures used for recognition. This dataset is called MSR-Action 3D dataset and which yields an average recognition rate of 88.2%. The MSR-daily activity 3D dataset of 16 gestures is used for recognition [5] of daily activities of a human such as drink, eat, read book, call cell phone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on a sofa, walk, play guitar, stand up, and sit down. The total of 320 sample gestures are captured by Kinect device for this dataset and achieved an average recognition rate of 85.75%. A bag of 3D points based human action or gesture recognition system is developed in 2010 [6] using Projection Based Sampling Scheme (PBSS) from depth map sequences. This system consists a dataset which contains twenty actions such as high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pickup & throw. The dataset used here are categorized as 3 data subsets such as AS1, AS2, and AS3. Also 3 test cases made for recognition such as Test One, Test Two, and Cross Subject Test. This recognition system yields an overall average recognition rate of 86.82% for bag of 3D Points (BOPs) which includes 3 subsets and 3 test cases. The same 3 data subsets were used for conduction of experiments with the 3 same test cases for 2D Silhouettes (2DS) and obtained and overall average recognition rate of 70.61%. In [7], the HGR system is developed for recognizing continuous gesture with complex background using Hidden Markov Model (HMM). This HGR systems uses twenty distinct moving hand gestures. All the 20 hand gestures were performed 3 times by 20 different signers. This systems uses two techniques such as Fourier Descriptor (FD), and Fourier descriptor and motion vector (FD & MV). These two techniques yields an average recognition rate of 90.5% and 93.5% respectively. The 2 real-time HMM based ASL continuous word level recognition system was developed in [8] using 2 different camera setup of desk mounted camera (DMC) and cap mounted camera (CMC). This system uses 40 ASL words of pronoun, verb, noun, and objectives. These two camera setup recognition systems obtains and average recognition rate of 91.9% and 97.8% respectively in plain black background with good illumination. This recognition system is also used for sentence level recognition by combining the 5 words.

IV. PROPOSED WORK

A. Foreground Detection using Gaussian Mixture Models (GMMs)

The foreground detection (FD) is one among the main tasks of computer vision and digital image & video processing. Its major goal is to detecting the variations in the sequences of the image. The FD, permits an image's foreground to be extricated for further recognition and other processes. Describing the background of the image is a challenging if it consists shapes, shadows, moving objects, color variation in objects, intensity over time, vastly variable sequences, like objects with completely illumination, locations, and noisiness.

One among the FD techniques in image processing is GMMs. The GMMs are the parametric PDFs (Probability Density Functions). The GMMs approaches by displaying every picture elements (pixels) as a combination of Gaussians. Here, it is expected that the value of all the pixel's intensity of the video is modeled by GMMs [9]. A modest heuristic regulates that intensities are the utmost possibly of the background. Here, the pixels that don't equivalent for these are known as foreground pixels.

All the pixels in an image / video is shaped by the *K* Gaussian distribution mixtures [10]. The probability, in which the assured pixels has the values of X_T at the time *T* is written as (1)

$$p(X_T) = \sum_{i=1}^{N} w_i \, \eta(X_T ; \theta_i) \quad -(1)$$

Here, W_N is a weight parameter of N^{th} Gaussian component. $\eta(x; \boldsymbol{\theta})$ is a normal distribution of N^{th} element denoted as (2)

$$\eta(X;\theta_N) = \eta(X;\mu_N,\Sigma_N) = \frac{1}{(2\pi)^{D/2} |\Sigma_N|^{1/2}} e^{-\frac{1}{2}(X-\mu_N)^{Tr}\Sigma_N^{-1}(X-\mu_N)} \quad (2)$$

Here, μ_N = mean and $\Sigma_N = \sigma_N^2 I$ = covariance of N^{th} component.

The *N* distributions are ordered based on the fitness value W_N / σ_N. The 1st *B* distributions are utilized as a model of an image / video background. *B* is assessed as (3)

$$B = \arg_b \min\left(\sum_{i=1}^{b} w_i > Tr\right) \quad (3)$$

Where, *Tr* = threshold, which is the minimum fraction of the background model. In other words, it is a minimum prior likelihood in which the background is in an image/ video. The background elimination is carried out by marking the foreground pixels. Every pixels which are greater than 2.5 standard deviances away from any of the *B* distributions. The various Gaussians are expected to signify dissimilar colors. The mixture weight parameters signify the period magnitudes that those colors halt in an image / video. The components of the background are defined by supposing that the background consists the more possible colors. The possible background colors stays elongated and more static. The FD System using GMMs compares the color or grayscale images or video frames to the actual background model or color for determining whether every singular picture elements are the part of the image or video background or the foreground. Then the foreground mask will be obtained using the GMMs.

B. Gray Threshold Selection and Image Binarization using Otsu's Thresholding Algorithm

The Otsu's Thresholding Algorithm is a non-parametric, and non-learning technique for auto selection of threshold for image segmentation, presented by Noboyuki Otsu in 1979 [11],

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

which calculates the global thresholding, which is used for converting an intensity image into the black and white (binary) image.

Resultant value of this is the normalized intensity value, which lies between the range [0, 1]. In order of assessing the good thresholding in k^{th} level, the following discriminant criterion measure is introduced for using in the discriminant analysis:

$$\lambda = \frac{\sigma_B^2}{\sigma_W^2}, \qquad \kappa = \frac{\sigma_T^2}{\sigma_W^2}, \qquad \eta = \frac{\sigma_B^2}{\sigma_T^2}, \tag{4}$$

Where

$$\sigma_W^2 = \omega_0\sigma_0^2 + \omega_1\sigma_1^2 \tag{5}$$

$$\sigma_B^2 = \omega_0(\mu_0 - \mu_T)^2 + \omega_1(\mu_1 - \mu_T)^2 = \omega_0\omega_1(\mu_1 - \mu_0)^2 \tag{6}$$

And

$$\sigma_T^2 = \sum_{i=1}^{L}(i - \mu_T)^2 p_i \tag{7}$$

Here, the λ, κ, and η are the variance with the class, the variance between the class, and the total variance of the levels, respectively.

C. Feature Extraction:

1. Speeded Up Robust Features (SURF)

The local feature extraction and description technique called SURF is used for recognition of objects, registration of images, classification, 3D reconstruction, The SURF is an innovative scale and rotation invariant feature extraction technique initially presented by Herbert Bay et. al in 2006 [12], which overtakes the earlier implemented techniques with respect to repeatedness, uniqueness, and robustness.

This algorithm consists 3 major parts: Interest point detection, local neighborhood description, and matching. As a Gaussian smoothing approximation, the SURF uses the box type filters, which is faster if it is an integral image $I_\Sigma(x)$ at the location X=(x, y) denotes the summation of entire pixels in an input image I as in equation (8).

$$I_\Sigma(X) = \sum_{i=0}^{i\le x}\sum_{j=0}^{j\le y} I(i,j) \tag{8}$$

To find out the point of interest, SURF uses the blob detector based Hessian Matrix H(X, σ), in X, at the scale σ, and is defines as (9)

$$H(X,\sigma) = \begin{bmatrix} L_{xx}(X,\sigma) & L_{xy}(X,\sigma) \\ L_{xy}(X,\sigma) & L_{yy}(X,\sigma) \end{bmatrix} \tag{9}$$

Here, $L_{xx}(X, \sigma)$ is a convolution of Gaussian 2^{nd} order derivative as in the equation (10) with an image I in the point X.

$$\frac{\partial^2}{\partial x^2} g(\sigma) \tag{10}$$

The hessian's determinant of 9X9 box filter approximation D_{xx}, D_{yy}, and D_{xy} is obtained as (11)

$$\det(H_{approx}) = D_{xx}D_{yy} - (0.9D_{xy})^2 \tag{11}$$

2. Zernike moment

Zernike moment is a feature extraction technique used to extract the set of rotation invariant features proposed in 1934 by Von F Zernike [13]. The magnitudes of collection of orthogonal complex moments of a digital image is termed as Zernike moments [14]. To achieve invariance of scale and rotation, the digital image is 1^{st} gone through the normalization process by applying regular moments. It consists 2 important values: Amplitude value and Angle Value. The Zernike moment always provides a superior accurateness, minor information removal, and also enhanced image reconstruction. The 2-dimentional Zernike moment is calculated using (12).

$$Z_{nl} = \left[\left[\frac{n+1}{\pi}\right]\sum_a\sum_b V^*nl(a,b) f(a,b)\right] \tag{12}$$

In this equation, $a^2 + b^2 \le 1$, $0 \le l \le n$, and *n-l* is even, *f*(a, *b*) denotes the intensity values of the normalized image and *V*nl* is a complex conjugate of a Zernike polynomial of degree *n* and angular dependence *l*.

3. Discrete Cosine Transform (DCT)

The DCT is the most broadly used and powerful transform in digital image processing applications [15] for extracting proper and distinctive features from the object or image. The DCT finds for coefficients that has much capability for discriminating various classes comparing to other coefficients. The DCT takes the image transformation as a whole and extract the distinctive relevant coefficients. The 2-D DCT of an image is mathematically provided as:

$$F(u,v) = \alpha(u)\alpha(v)\sum_{x=0}^{N-1}\sum_{y=0}^{M-1} c1 \;\; c2 \;\; f(x,y) \tag{13}$$

$$\alpha(u)\alpha(v) = \begin{Bmatrix} \sqrt{\frac{1}{N}} & for \;\; u,v \neq 0 \\ \sqrt{\frac{2}{N}} & for \;\; u,v = 0 \end{Bmatrix} \tag{14}$$

Where,

$$c1 = \cos\left[\frac{\pi u}{2N}(2x+1)\right] \tag{15}$$

$$c2 = \cos\left[\frac{\pi u}{2M}(2y+1)\right] \tag{16}$$

Here, F(x, y) is the pixel intensity at (x, y) coordinates. The 'u' and 'v' varies from 0 to M-1 and 0 to N-1 respectively.

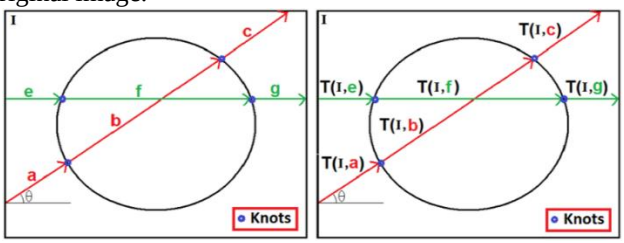
4. Radon Features

The Radon features (Radon-Like Features) are the feature extraction technique, used to enhance the cell boundaries, image segmentation in the form of intensities. These features can be used by the supervised or unsupervised learning methods. The Radon like features kept the key idea of Radon transform [16], is an integral transformation in 2-dimentional as (17)

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

$$R(m,\tau)[f(x,y)] = \int_{-\infty}^{\infty} f(x, m+\tau y)dx \quad \text{- (17)}$$

Here, m is a slope and τ is an intercept of the line. The f(x, y) is an integrated 2-D function. The inverse of the above transformation can be used for reconstruct the original image.



(a) (b)

Fig. 3. The Radon Like features

The value of Radon like features at the point P, with line segment l, and knots (t_1,…, t_n) as in figure 3, between (x (t_i), y (t_i)) and (x (t_{i+1}), y(t_{i+1})) is defined as (18)

$$\Psi(p,l,t_i,t_{i+1})[I(x,y)] = T(I,l(t)), t \in [t_i,t_{i+1}] \quad \text{- (18)}$$

Here, T is an extraction function.

The extraction of Radon like features using the extraction function T_1 as in (19)

$$T_1(f,l(t)) = \| l(t_{i+1}) - l(t_i) \|_2, t \in [t_i,t_{i+1}] \quad \text{- (19)}$$

Here, the extraction function T1 assigns all the pixels between the knots. The cell boundary enhancement of an image using Radon like features for an input image I(x, y) is as in (20)

$$R(x,y) = \max_{\sigma,\phi} \Delta G(\sigma,\phi) \otimes I(x,y) \quad \text{- (20)}$$

In equation (20), σ and ϕ are the scale and the boundary orientation to enhance Gaussian 2nd derivative filter $\Delta G(\sigma, \phi)$. The $\otimes$ represents the convolution.

D. Classification using Deep Neural Networks (DNN):

The neural networks consisting multiple hidden layers are very useful to solve the classification difficulties having complex data, i.e., images. Here, every layer will learn features at various levels of abstraction. An Artificial Neural Network (ANN) with two or more hidden layers between the input and output layers are termed as DNN, in which, the sophisticated mathematical model is used to process the data in intricate ways and computes the probability of each hidden layer output. All the layers executes precise kinds of arranging and assembling in a procedure, which is termed as "feature hierarchy". The main use of this DNN is working with unlabeled / unstructured complex data [17].

Gesture Classification using Stacked Auto-Encoders Neural Network

The Stacked Auto-Encoder (SAE) Neural Network is a DNN and can also called as Stacked Neural Network (SNN), consists multiple hidden layers between the input and output layers.The better way to obtain the beneficial and optimal features from the network is by efficiently train all the layers is by training one layer at a time. This can be achieved by training a special kind of network called an auto-encoder for all the hidden layers in an unsupervised manner. First train the 1st layer on raw input data to get the processed features. By using this 1st layer, for transforming the raw input data to the vector comprising of initiation of the hidden units. Next, train the 2nd layer on this vector to get the 2nd level of processed features. Repeat this training for all the layers one at a time, by inputting the output of each layer to the succeeding layer. This process trains the features of all the layer independently while fixing the features to the remainder of the model. After completion of training of each layers, the fine tuning of the features to be carried out to obtain the better result for classification process. The common practice for fine tuning is that just discard the 'decoding part' of each layer of the auto-encoders and connect the final hidden layer to the softmax classifier. Finally, train the softmax layer in supervised manner, and joins all the layers together with the output layer to form a stacked network [18][19]. The diagram of SNN is illustrated in Figure 4, which is molded with an input layer, 2 encoders from 2 auto-encoders, a softmax layer, and an output layer [19]. In the encoding stage, the trained features are used and express x_i *(1, 2,…,N)* as an input vector, h_i as the hidden layer representation. An input vector for calculating the x_i and joint Probability Distribution Function (PDF) of h_i. This will be utilized as an initial matrix weight. The PDF is calculated as:

$$p(h_i = 1 \mid x) = \sigma(b_i + \sum_j w_{ij} x_i) \quad \text{- (21)}$$

Where, σ is the sigmoid function. The w and b are the weight matrix and the bias respectively. The σ is defined as:

$$\sigma = \frac{1}{1+e^{-z}} \quad \text{- (22)}$$

The network input data is defined as z, the network output data is defined as $h_{w,b}$ *(z)* and the initial weighted matrix is defined as w_{ij} *(i=1, 2, …N)*. Input data z is activated through the mapping function to given m_f as follows,

$$m_f = \sigma(w_i z + b_i) \quad \text{- (23)}$$

Where, σ is an activation function also called as a sigmoid function.The 1st layer of the SNN supervises to learn the 1st order features from the raw input such as edges of an image.

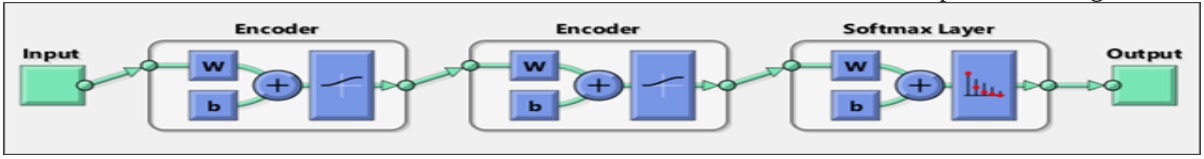


Fig. 4. Stacked Neural Network

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

The 2nd layer inclines to learn the 2nd order features conforming to patterns in the appearance of 1st order features. For example, in relations of what edges incline to take place together such as, for forming contour and / or corner detectors. Similarly, the succeeding layers of the network supervise to learn even higher order features.

E. Proposed Algorithm

The figure 5 shows the training and testing process of the proposed ASL video gesture recognition system.

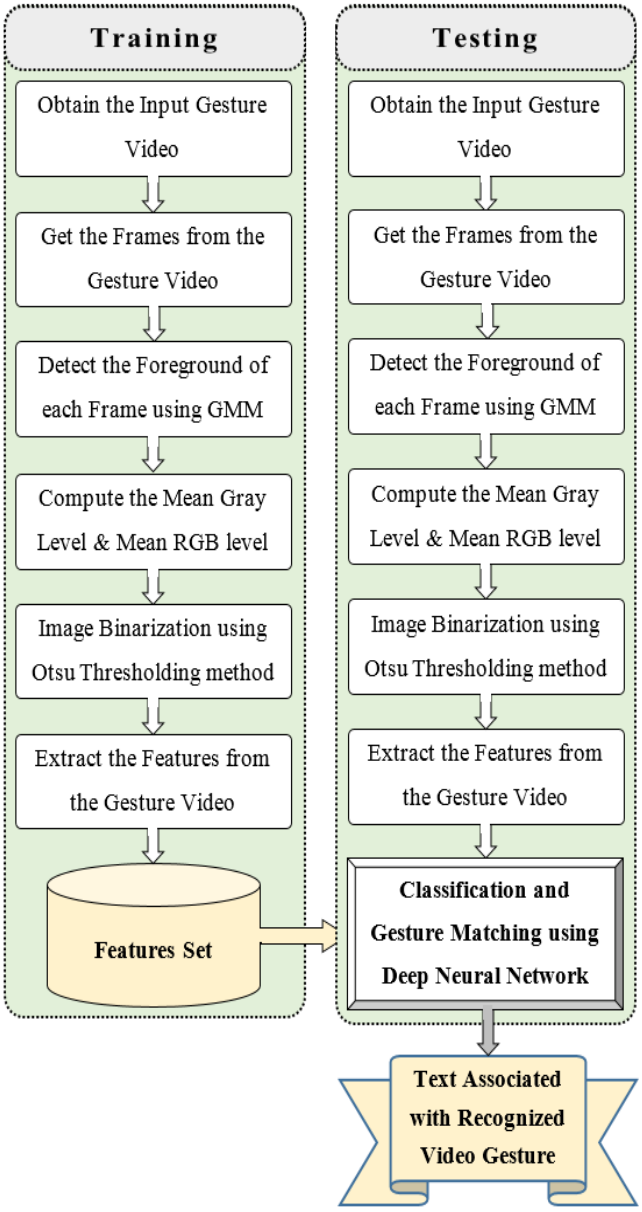


Fig. 5. The training and testing process

In the training phase of figure 5, initially, the input video gesture is read from the database folder, then this video will be converted as images frames. Here, we have selected 35 to 40 alternate frames from the each video gestures. Next, the foreground detection is carried out by using the GMM. Once the foreground of the video is detected, we have computed the mean gray level and mean R, G, and B level for further process.

The first 6 steps of the testing phase are exactly similar as in the testing phase. Once the necessary features are extracted from the testing gesture video, these extracted features will be matched and compared with the feature values of feature matrix for classification and gesture matching using deep neural network (Stacked Auto-Encoder Neural Network). Based on the DNN classification result, the text associated with the recognized video gesture is displayed.

The Otsu thresholding technique is used for gray threshold selection and image binarization. The necessary and useful features of the gesture video are extracted using 7 various feature extraction techniques such as SURF, Zernike moment, DCT, Radon, Red, Green, and Blue features. Once the features are extracted, will be loaded to the feature matrix for further utilizing them in the testing phase.

V. RESULTS AND DISCUSSION

The extracted feature values of video gestures dataset sets S1 through S7 are shown in table II through table VIII respectively. Each table consists the feature values of 12 video gestures from 2nd row to last row (Hello to Z), and 7 sets of features from column 2 to column 8.

Table II. Extracted Video Gestures' Feature Values of Dataset S1

Features ▸ Gestures ▾	R	G	B	SURF	ZM	DCT	Rad
Hello	66.525	74.498	67.169	0.034	0.009	-2.5E-06	0.510
Hi	66.371	75.311	66.976	0.033	0.006	-4.3E-07	0.527
J	67.544	82.370	71.982	0.035	0.007	-2.6E-07	0.921
Me	63.525	77.604	66.860	0.032	0.004	2.33E-07	0.242
No	69.884	76.911	69.178	0.032	0.007	1.6E-06	0.842
Sorry	65.247	77.686	68.016	0.037	0.010	8.7E-05	0.479
Thanks	63.004	76.418	67.598	0.039	0.027	4.81E-06	0.980
They	64.708	77.037	67.902	0.035	0.014	3.69E-05	0.740
We	64.989	77.311	68.286	0.033	0.004	3.52E-07	0.326
Yes	71.789	78.724	70.518	0.033	0.009	8.07E-07	1.060
You	62.293	75.950	66.427	0.034	0.005	1.4E-06	0.837
Z	61.409	77.138	66.498	0.038	0.025	3.47E-05	1.449

Table III. Extracted Video Gestures' Feature Values of Dataset S2

Features ▸ Gestures ▾	R	G	B	SURF	ZM	DCT	Rad
Hello	166.395	151.414	135.731	0.031	0.027	3.83E-06	0.761
Hi	168.084	152.841	137.358	0.034	0.015	4.34E-06	0.334
J	160.718	146.349	126.896	0.033	0.014	3.31E-06	0.818
Me	148.485	135.393	118.062	0.032	0.007	1.88E-06	0.242
No	166.708	151.455	135.719	0.033	0.017	2.69E-06	0.360
Sorry	175.181	158.131	140.898	0.034	0.015	0.000245	0.541
Thanks	152.208	138.785	121.217	0.032	0.012	2.23E-06	0.556
They	152.183	138.699	120.004	0.032	0.010	2.45E-06	0.629
We	150.841	137.914	120.342	0.032	0.011	2.10E-06	0.677
Yes	166.881	151.475	135.755	0.032	0.015	8.05E-07	0.320
You	159.476	146.021	127.722	0.032	0.007	1.78E-06	0.380
Z	155.923	142.171	124.420	0.033	0.008	6.03E-07	0.493

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Table IV. Extracted Video Gestures' Feature Values of Dataset S3

Features ▸ / Gestures ▾	R	G	B	SURF	ZM	DCT	Rad
Hello	141.086	141.970	147.747	0.035	0.028	2.92E-06	1.401
Hi	141.832	142.907	148.361	0.036	0.019	1.54E-06	0.929
J	141.007	140.680	145.044	0.031	0.016	1.14E-06	0.650
Me	146.017	146.805	151.043	0.035	0.017	9.69E-07	0.628
No	148.544	149.282	152.004	0.036	0.031	3.50E-06	1.567
Sorry	145.557	144.870	148.300	0.033	0.008	1.02E-06	0.699
Thanks	140.975	141.056	145.840	0.034	0.008	8.51E-07	0.561
They	144.227	143.256	146.482	0.030	0.039	3.20E-06	1.446
We	143.958	143.598	146.819	0.034	0.023	1.54E-06	0.866
Yes	135.609	136.515	143.650	0.034	0.023	1.59E-06	0.908
You	146.095	145.860	149.119	0.032	0.010	7.65E-07	0.370
Z	145.498	146.544	151.332	0.032	0.012	8.83E-07	0.630

Table V. Extracted Video Gestures' Feature Values of Dataset S4

Features ▸ / Gestures ▾	R	G	B	SURF	ZM	DCT	Rad
Hello	15.057	16.754	13.382	0.034	0.026	1.21E-06	0.682
Hi	18.734	20.367	16.263	0.034	0.011	1.12E-06	0.697
J	10.098	17.011	9.812	0.037	0.016	7.10E-07	0.381
Me	19.933	21.681	17.547	0.031	0.015	1.07E-06	0.626
No	19.697	21.865	17.692	0.034	0.006	9.89E-07	0.421
Sorry	22.542	24.640	20.204	0.033	0.012	1.64E-06	0.777
Thanks	21.631	24.086	19.897	0.033	0.016	9.61E-07	0.552
They	18.994	20.593	17.031	0.028	0.018	2.87E-06	0.741
We	22.210	24.046	19.690	0.035	0.025	3.01E-06	0.751
Yes	20.879	23.042	18.939	0.034	0.007	4.04E-07	0.298
You	21.920	23.718	19.741	0.034	0.014	1.11E-06	0.570
Z	11.799	10.461	13.241	0.032	0.010	6.59E-07	0.424

Table VI. Extracted Video Gestures' Feature Values of Dataset S5

Features ▸ / Gestures ▾	R	G	B	SURF	ZM	DCT	Rad
Hello	21.387	20.969	19.743	0.033	0.028	7.63E-06	0.707
Hi	22.375	21.836	20.929	0.028	0.017	1.36E-06	0.420
J	24.286	27.315	24.821	0.023	0.008	4.89E-06	0.473
Me	12.135	13.153	16.179	0.034	0.017	2.68E-06	0.725
No	19.741	19.480	18.375	0.031	0.016	4.77E-05	1.092
Sorry	26.172	26.428	25.772	0.031	0.010	1.80E-06	0.508
Thanks	25.284	25.990	24.667	0.032	0.018	2.90E-05	1.042
They	25.275	26.338	24.824	0.028	0.015	5.03E-05	0.535
We	25.152	26.838	24.743	0.032	0.024	5.89E-06	0.746
Yes	20.497	20.283	19.411	0.033	0.019	0.00019	0.934
You	20.739	27.313	24.598	0.031	0.009	1.06E-06	0.086
Z	27.349	30.291	28.734	0.033	0.018	8.00E-06	0.770

Table VII. Extracted Video Gestures' Feature Values of Dataset S6

Features ▸ / Gestures ▾	R	G	B	SURF	ZM	DCT	Rad
Hello	77.630	79.281	98.705	0.035	0.032	8.81E-06	1.460
Hi	78.972	80.525	98.565	0.030	0.014	8.06E-06	0.733
J	83.819	85.778	101.472	0.035	0.016	1.68E-05	2.276
Me	87.326	87.963	104.151	0.030	0.017	4.88E-05	1.806
No	71.485	73.598	94.594	0.032	0.009	0.00028	1.124
Sorry	82.728	85.208	102.834	0.033	0.016	9.30E-06	1.661
Thanks	84.167	86.841	104.109	0.025	0.007	3.46E-06	0.655
They	82.278	84.259	101.549	0.027	0.017	8.42E-06	0.983
We	83.221	85.371	101.969	0.034	0.028	6.30E-06	1.090
Yes	73.569	75.880	97.436	0.034	0.019	3.70E-06	0.930
You	86.684	87.502	102.911	0.030	0.020	8.38E-06	1.606
Z	78.134	80.245	96.276	0.031	0.016	6.61E-05	1.337

Table VIII. Extracted Video Gestures' Feature Values of Dataset S7

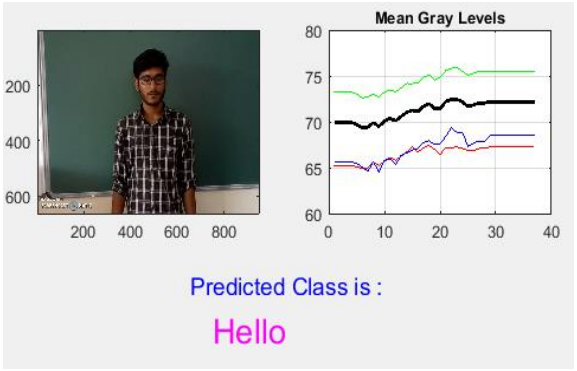
Features ▸ / Gestures ▾	R	G	B	SURF	ZM	DCT	Rad
Hello	65.737	67.494	61.952	0.032	0.035	4.64E-06	2.030
Hi	72.069	72.940	64.722	0.032	0.011	0.00038	2.624
J	71.685	72.760	64.651	0.031	0.012	6.10E-06	1.633
Me	66.415	67.377	63.287	0.032	0.012	9.25E-05	1.867
No	70.636	72.508	65.804	0.032	0.019	7.46E-06	2.654
Sorry	70.996	72.382	64.775	0.032	0.013	2.67E-05	2.771
Thanks	68.739	69.839	63.443	0.030	0.015	0.00028	4.239
They	70.228	71.766	67.468	0.031	0.014	7.46E-06	0.917
We	69.900	71.052	65.186	0.034	0.027	0.00018	4.073
Yes	68.620	70.846	64.082	0.033	0.032	5.95E-05	2.141
You	67.733	69.465	65.259	0.032	0.013	5.89E-06	1.843
Z	67.984	69.172	65.859	0.032	0.019	0.00015	4.109

Some of signer independent output samples of ASL video gestures of invariant background, location, illumination and distance shown in Figure 6, which shows the final frame of the video gestures, mean gray levels and corresponding predicted class/output labels.

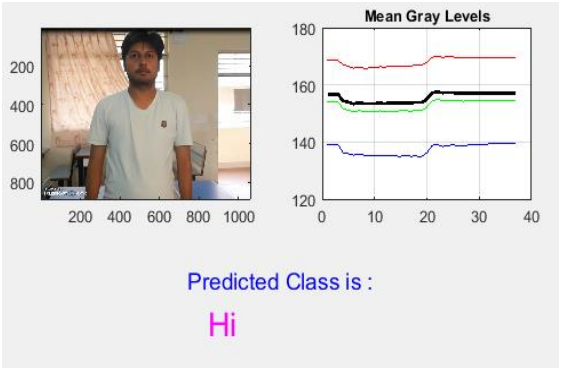
The output samples in Figure 6 (A), 6 (B), 6 (C), 6 (D), 6 (E), and 6 (G) are from video gesture Set S1, S2, S3, S4, S5 and Set S7 respectively. Also, Figure 6 (F) and (H) outputs are from Set S6.

The Videowise average accuracy (VAA) of 12 video gestures and Set-wise average accuracy (SAA) of the 7 datasets S1 to S7 are considered for testing, which is tabulated in table IX, the recognized and not recognized video gestures are represented by □and □ symbols respectively.

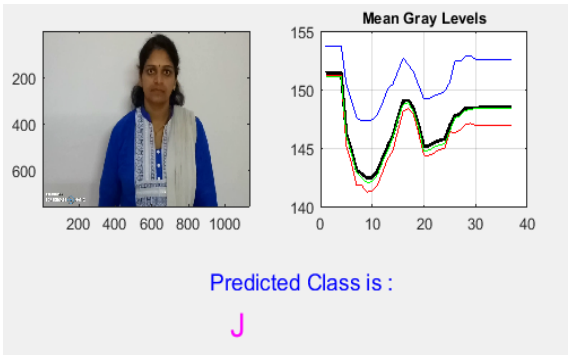
Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)



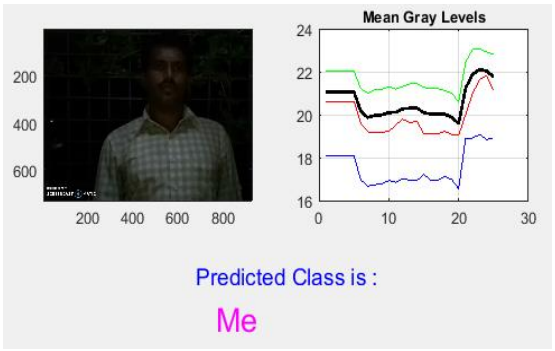
(A)



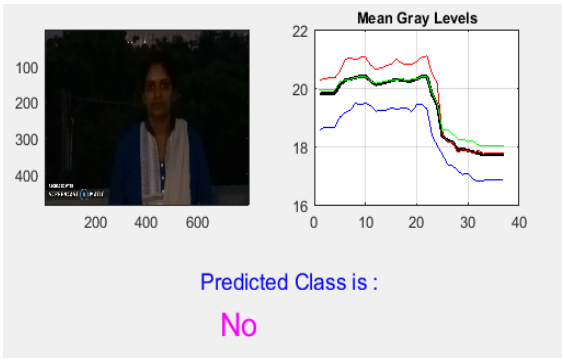
(B)



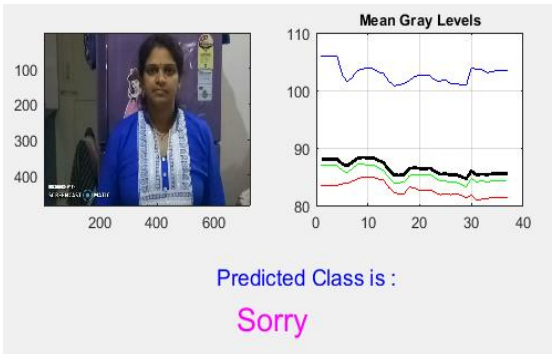
(C)



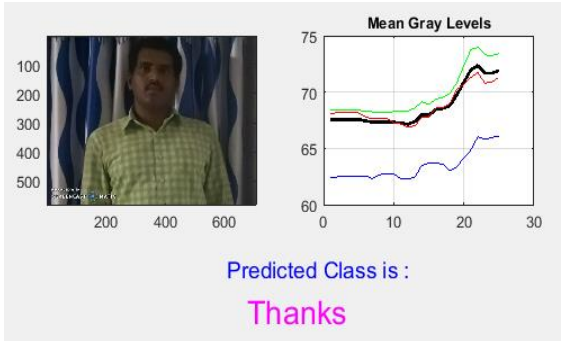
(D)



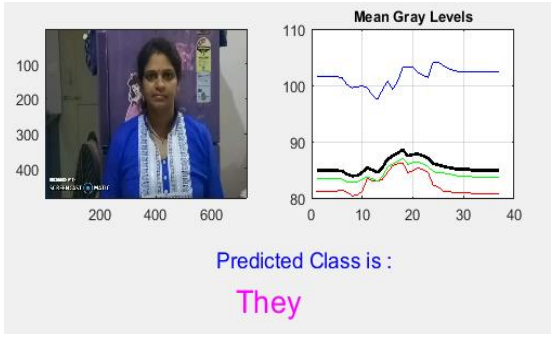
(E)



(F)



(G)



(H)

Fig. 6. Some of the output samples

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Table IX. Set-wise and Video-wise Average Accuracy

Datasets → / Gestures ▼	S1	S2	S3	S4	S5	S6	S7	VAA (%)
Hello	✔	✔	✔	✔	✔	✔	✔	100
Hi	✔	✔	✔	✔	✔	✔	✔	100
J	✔	✔	✔	✘	✔	✔	✔	85.7
Me	✔	✔	✔	✔	✔	✔	✔	100
No	✔	✔	✔	✔	✔	✔	✔	100
Sorry	✔	✔	✔	✔	✔	✔	✔	100
Thanks	✔	✔	✔	✔	✔	✔	✔	100
They	✔	✔	✔	✔	✔	✔	✔	100
We	✔	✔	✔	✔	✔	✔	✔	100
Yes	✔	✔	✔	✔	✔	✔	✔	100
You	✔	✔	✔	✔	✘	✔	✔	85.7
Z	✔	✔	✔	✘	✔	✔	✔	85.7
SAA (%)	100	100	100	83.3	91.7	100	100	96.43

Table IX shows that, overall, only three video gestures such as 'J' and 'Z' from the dataset S4, and the 'You' from dataset S5 were not recognized due to low illuminations of artificial lightings and faster hand movement of the video. The remaining 165 video gestures from the S1 to S7 were recognized properly. The average recognition rate achieved for plain background video gestures sets S1 and S3 is 100%, which is better compared with the average recognition rate of 95% for complex background video gestures. The gestures captured from 5 feet distance (S2, S3, S4, and S6) yields an average recognition rate of 95.83% whereas the gestures captured from 10 feet distance provides 97.23%.

The video gestures captured from 8 and 13 MP mobile cameras provides an average accuracy of 100% and 91.7% respectively. By considering all these invariants of the datasets, the PVGRS achieved an overall average recognition rate of 96.43%, which is better comparing with the Existing Video Gesture Recognition Systems (EVGRS).

Figure 7 shows the background-wise average accuracy of 7 video gesture sets. It is observed that, irrespective of natural or artificial illumination, and invariant to distance and signers, the video gestures captured using 8MP camera with plain background offers an excellent average accuracy of 100%. The video gestures captured with complex background with higher resolution camera (i.e., 13 MP) and artificial lightings conditions of day time provides the better average accuracy of 95%. It is clear from this, gestures captured in plain background offers better results than the gestures captured in complex background.

Table X highlights the comparative analysis of PVGRS and EVGRS, with their recognition accuracy and techniques used.

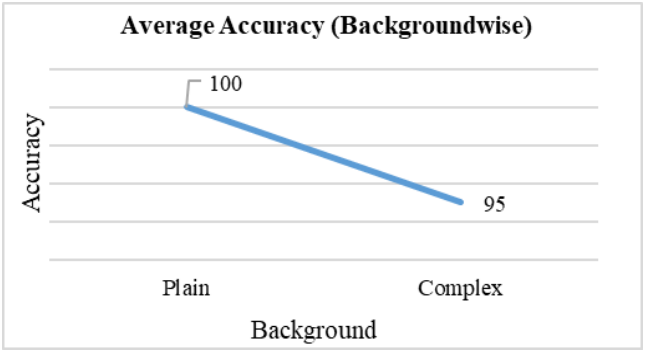


Fig. 7. Average Accuracy (Background considered)

Table X. Comparison of EVGRS and PVGRS

Year & Ref. No.	Method Used	Accuracy (%)	Remarks
2012 [4]	AGM	87.7	12 ASL gestures. Invariant to orientation, illumination.
2012 [5]	AEM - 20G	88.2	20 3D human actions. Robust to noise, invariant to translational and temporal misalignments but plain background.
2012 [5]	AEM - 16G	85.75	3D dataset of 16 daily activity gestures.
2010 [6]	PBSS - BOPs	86.82	20 actions of bag of 3D Points.
2010 [6]	PBSS - 2DS	70.61	20 actions of 2D Silhouettes.
2003 [7]	HMM, FD	90.5	20 hand gestures in complex background.
2003 [7]	HMM, FD & MV	93.5	20 distinct gestures in complex background.
1998 [8]	HMM, DMC	91.9	40 ASL words in plain black background with good illumination.
1998 [8]	HMM, CMC	97.8	40 ASL words in plain black background with good illumination.
PVGRS	**SNN**	**96.43**	**12 ASL words with invariant location, background, signer, illumination, distance, and also camera resolution**

In Table X, it is noticed that, the HMM, and CMC based gesture recognition system [8] achieved 97.8% of accuracy, which is bit more than the PVGRS but in [8], gestures were captured in plain black background with good illumination of controlled lab environment which cannot be comparable with PVGRS as the video gestures captured here are invariant location, background, signer, illumination, distance, and also camera resolution. Due to these invariants, an average accuracy of the PVGRS is better than the EVGRS.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

VI. CONCLUSION AND FUTURE SCOPE

In this paper, the sincere effort has been placed to recognize the some of the ASL videos into human or device identifiable English text. There are 14 datasets are created for recognition process. Among the 14 datasets, the 7 datasets were used for training and remaining 7 datasets were used for experimentation. All the training and testing datasets are signer independent, invariant to location, illumination, distance, background and pixel resolution of the camera. In both training and testing process, some common tasks such as foreground detection, frame selection, feature extraction were carried out. Further, in testing process, stacked auto-encoder neural network of DNN is used for classification and recognition of video gestures. As a result of experimentation, the PVGRS produces an overall average recognition rate of 96.43%. It is noticed that, due to the gestures captured in low illumination night time, there is a bit of loss of recognition rate. Many of the state of art ASL video recognition systems were carried out in controlled lab environments with plain background and good illumination conditions. Overall, the recognition rate obtained is better comparing with the state of art techniques. The dataset-wise and video gestures-wise recognition rate is illustrated in Table 9. The comparative overall recognition rate of SL video gestures is highlighted in Table 10. This PVGRS motivates other researchers for carrying out video restores recognition task with more robust and improved recognition rate.

As a future directions of this PVGRS, it can be scaled to more number of video gestures and also try out with developing the double handed video gestures considering various angles and much more invariant distances. Also, it can be extended for simple two to three words sentence recognition tasks.

REFERENCES

1. Srinath S, Ganesh Krishna Sharma, "Classification approach for sign language recognition", *in Proc. International Conference on Signal, Image Processing, Communication & Automation*, (2017), pp.1-11.
2. Shivashankara S, Srinath S, "A comparative study of various techniques and outcomes of recognizing American Sign Language: A Review", *International Journal of Scientific Research Engineering & Technology*, Vol. 6, Issue 2, (2017), pp.1013-1023.
3. Shivashankara S, Srinath S, "A review on vision based American Sign Language recognition, its techniques, and outcomes", *Presented at 7th IEEE International Conference on Communication Systems and Network Technologies*, (2017). [Online]. Available: https://ieeexplore.ieee.org /document /8418554
4. Alexey Kurakin, Zhengyou Zhang, Zicheng Liu, "A real time system for dynamic hand gesture recognition with a depth sensor", *20th European Signal Processing Conference*, (2012), pp.1975-1979.
5. Jiang Wang, Zicheng Liu, Ying Wu, Junsong Yuan, Mining, "Actionlet Ensemble for Action Recognition with Depth Cameras", I*EEE Conference on Computer Vision and Pattern Recognition, Providence*, (2012), pp.1290-1297. [Online]. Available: http://www.uow.edu.au /~wanqing/#Datasets
6. Wanqing Li, Zhengyou Zhang, Zicheng Liu, "Action Recognition Based on A Bag of 3D Points", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, (2010), pp.9-14. [Online]. Available: http://www.uow.edu.au/ ~wanqing/#Datasets
7. Feng-Sheng Chen, Chih-Ming Fu, Chung-Lin Huang, "Hand gesture recognition using a real-time tracking method and hidden Markov models", *Image and Vision Computing (Elsevier)*, 21, (2003), pp.745–758.
8. Thad Starner, Joshua Weaver, and Alex Pentland, "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 12, (1998).
9. Stauffer, C. and Grimson, W.E.L, Adaptive Background Mixture Models for Real-Time Tracking, Computer Vision and Pattern Recognition, *IEEE Computer Society Conference*, Vol. 2, (1999), pp. 2246-252.
10. P. KaewTraKulPong and R. Bowden, "An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection", *2nd European Workshop on Advanced Video Based Surveillance Systems*, (2001), pp.1-5.
11. Nobuyuki Otsu, "A Threshold Selection Method from Gray-Level Histograms", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-9, No. 1, (1979), pp.62-66.
12. Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "SURF: Speeded Up Robust Features", *9th European conference on computer vision (Springer)*, (2006), pp.404-417.
13. V F Zernike, "Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasen kontrast methode", *Physica*, Vol. 1, Issue 7-12, (1934), pp.689-704. https://doi.org/10.1016/S0031-8914 (34) 80259-5.
14. A Kothanzad, Y H Hong, "Invariant image recognition by Zernike moments", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, Issue 5, (1990), pp.489-497. [Online]. Available: https://ieeexplore.ieee.org/document/55109.
15. K Manikantan, Vaishnavi Govindarajan, V V S Sasi Kiran, S Ramachandran, "Face Recognition using Block-Based DCT Feature Extraction", *Journal of Advanced Computer Science and Technology*, Vol. 1, Issue 4, (2012), pp.266-283. [Online]. Available: www.sciencepubco.com/index.php/JACST
16. Ritwik Kumar, Amelio V´azquez-Reina, Hanspeter Pfister, Radon-Like Features and their Application to Connectomics, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops*, (2010), pp.186-193. [Online]. Available: https://ieeexplore.ieee.org/document/5543594
17. https://www.techopedia.com/definition/32902/deep-neural-network
18. http://ufldl.stanford.edu/wiki/index.php/Stacked_Autoencoders.
19. https://www.mathworks.com/

AUTHORS PROFILE

Shivashankara S, Research Scholar at Sri Jayachamarajendra College of Engineering, Mysuru (Visvesvaraya Technological University, Belagavi), India. Born in 1980. Received B E in computer science and engineering in 2005, and M.Tech in computer network engineering in 2008 from Visvesvaraya Technological University, Belagavi, India. He joined as Lecturer in 2005 as Assistant Professor in 2011 and having 12 years teaching experience and guided many UG students for their academic projects and has 4 years of research experience. Currently pursuing Ph.D degree in Visvesvaraya Technological University, Belagavi, India. He has published 7 research papers in national / international journals and conference proceedings. The main research interests includes Image Processing & Pattern Recognition, and Machine & Deep Learning.

Dr.Srinath S Ph.D., Associate Professor at JSS Science and Technology University, Mysuru, India. Received his Engineering degree in 1995, M.Tech (1st Rank with Gold Medal) in 2002 and Ph.D. in 2015. He has published 20 scientific papers in international and national journals and conference proceedings. Executed 3 research projects sponsored from VTU, AICTE and MHRD (Government of India). Organized several training programs sponsored by different government agencies. Given more than 50 invited talks at different colleges. He has been the resource person for many training programs. The main research interests include Pattern Recognition and Image Processing.