ISSN: 2581-4419 Volume 1 Issue 1

SIGN LANGUAGE RECOGNITION USING NEURAL NETWORK

KAUSTUBH JADHAV[1], ABHISHEK JAISWAL[2], ABBAS MUNSHI[3], MAYURESH YERENDEKAR[4]
[1-4] Students, Department of Electronics and Telecommunication Engineering K.C. College of Engineering & Management studies & Research,

Kopri, Thane (E)-400 603, India.

Abstract*— *A practical sign language translator is an essential way for communication between the deaf community and the general public. So here we present the development and implementation of an American Sign Language (ASL) fingerspelling translator based on a convolutional neural network. We utilize a pre-trained GoogLeNet architecture trained . We produced a robust model which classifies letters a-z correctly with first-time users and another that correctly distinguish letters a-k in a majority of cases. The limitations of the dataset and the encouraging results achieved, we are confident that with further research and more data, we can produce a generalized translator for all ASL letters.

Keywords- ASL, Sign Language Character Recognition, Convolution Neural Network, Computer Vision, Machine Learning.

I. INTRODUCTION

Although sign languages have emerged naturally in deaf communities alongside or among spoken languages, they are unrelated to spoken languages and have different grammatical structures at their core.One might expect digital technologies will play a huge role in human's daily routines and whole world will be interacting via machines either with the means of gestures or speech recognition within a few decades. If we are in a position to predict such a future, we ought to think about the physically challenged and do something for them. Sign language is the natural language of the deaf and aphonic people. It is the basic method for the communication of deaf person. American Sign Language (ASL) is the language chosen by almost all the deaf communities of United States of America. Different Sign languages are evolved depending on the regions such as GSL (German Sign Language), CSL (Chinese Sign Language), Auslan (Australian Sign Language), ArSL (Arabic Sign Language), and many more [1].

II. RELATED WORK

Characterization of sign language is between two parameter one being manual and other non-manual. The manual parameter consists of motion, location, hand shape, and hand orientation. The non-manual parameter includes facial expression, mouth movements, and motion of the head [2]. Sign language does not include the environment which kinesics does. Few terms are use in the sign language like signing space, which refers to signing taking place

in 3D space and close to truck and head. Signs are either one-handed or two-handed. When only the dominant hand is in use to perform the signs they are denoted as one-hand signs else when the non-dominant hand also comes in the phase it is termed as two- handed signs [3].

Sign language when evolved is different from spoken language so the grammar of the sign language is primarily different from spoken language. Inspoken language, the structureof the sentence is one-dimensional; one word followed by another, while in sign language, a simultaneous structure exists with aparallel temporal and spatial configuration. As based on these characteristics, the syntax of sign language sentence is not as strict as in spoken language. Formation of a sign language sentence includes or refers to time, location, person, base. In spoken languages, a letter represents a sound. For deaf nothing comparable exists. Hence the people, who are deaf by birth or became deaf early in their lives, have very limited vocabulary of spoken language and faces great difficulties in reading and writing.

III. AMERICAN SIGN LANGUAGE (ASL):

American Sign Language (ASL) is the non-verbal way of communication based on English language. Which can be expressed by movements of the hands and face.It is the primary language of many North Americans who are deaf and find difficulties in hearing . It is not a universal sign language. Different sign languages are used in different countries or regions. For example, British Sign Language(BSL) is a different language compared to ASL so the person who knows ASL may not understand BSL.ASL is forth most commonly used Language in US.

Figure 1:Sign Language Alphabets of the ASL

ASL is a language completely segregated and different from English. ASL contains all the significant features of language, with its own rules for pronunciation, word formation, and word order. While every language has ways of indicating different functions, such as asking question instead of making a statement

IV. CONVOLUTIONAL NEURAL NETWORK

A convolutional neural network (CNN) is a type of artificial neural network which uses perceptron learning rule along with supervised learning, to analyze the data.CNN is applied to process the image, natural language processing and other kinds of cognitive tasks. Like other kinds of artificial neural networks, a convolutional neural network contains an input layer, an output layer and various hidden layers.Some of these layers are convolutional, using a mathematical model to pass on results to successive layers. This simulates some of the actions in the human visual cortex. CNN are a fundamental example of deep learning algorithm.

1. Input Layers : The layer in which input is given to our model. The number of neurons in this layer is equal to total number of features in our data.
2. Hidden Layer: In this layer the input from Input layer is feed into the hidden layer.There can be many hidden layers depending upon our model and data size. Each hidden layers can have different numbers of neurons which are generally greater than the number of features. The output from each layer is computed by matrix multiplication of output of the previous layer with learn able weights of that layer and then by addition of learn able biases followed by activation function which makes the network nonlinear.
3. Output Layer: The output from the hidden layer is then fed into a logisticfunction like sigmoidal or softmax layer which converts the output of each class into probability score of each class.

V. LITERATURE

The hand gesture recognition is a well contributed research area with a lot of different approaches in implementing it, in this chapter, we review a variety of such approaches for hand gesture recognition. Our entire literature review on hand gesture recognition can be categorized into three groups, image processing/statistical modelling based recognition ,classic machine learning based recognition and deep learning based recognition.
Triesch and Malsburg in 1996 developed an ASL hand gesture recognition system ,aiming at performing accurate gesture recognition even for images captured with complex background. In this system, they have eliminated hand
segmentation assuming hand images input. For hand representation or feature extraction, the system uses Gabor filters as the Gabor filters are known to resemble the receptive fields of visual cortex. Upon obtaining the Gabor features, they performed a similarity matching technique for gesture recognition called Elastic Bunch- Graph Matching (EBMS) technique. For experimentation they have

considered a subset of 10 gestures from American sign language and the dataset consists of 657 images captured from 24 people in three different backgrounds (Complex, uniform light and uniform dark). The system has achieved an accuracy of 86.2% under complex background condition and overall accuracy of 91% under all background conditions. Many researches have taken the path of HMM based modeling for hand gesture recognition pursuing it as action recognition problem[5].

VI. METHODOLOGY

VI.1. ACQUISITION OF DATA (CAMERA INTERFACING)

This is a primary and essential step in sign recognition whole process. Camera interfacing is necessary task to capture images with the help of Webcam. Now a days lots of Laptops are coming with inbuilt camera system so that's helps lot for capturing images to process it further. Gestures can be captured by inbuilt camera to detect hand movements and position. Capturing 30fps will be sufficient to process images; more input images may lead to higher computational time and will make system slow and vulnerable.

VI.2. IMAGE PROCESSING

Image pre-processing contains removing unwanted noise, adjusting brightness and contrast of the image, cropping the image as per requirement [li]. In this process contains image enhancement, segmentation and color filtering process [6].

VI.3. IMAGE ENCHANCEMENT AND SEGMENTATION

As images captured by webcam is RGB images, but RGB images are very much sensitive for various light conditions therefore RGB information convert into YCbCr. Where Y is luma component which denotes luminance information of image, and Cb , Cr are chromo components which give color information of images red difference and blue difference. Luminance component may create problems so only chrominance components get process further. After that YCbCr image converted to binary image.

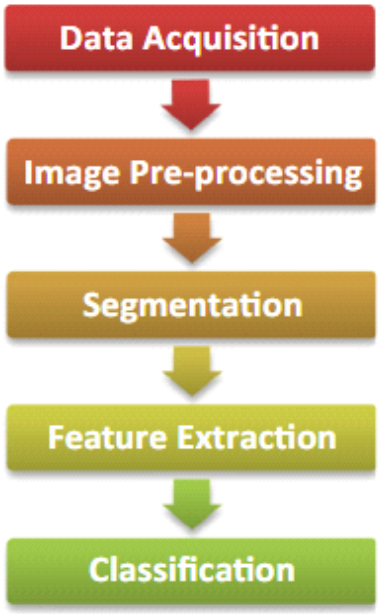

Figure 2 : Image processing FlowChart

VI.4. COLOUR FILTERING AND SKIN SEGMENTATION

As real time image capture by web camera contains collection of frames.There is need to convert RGB image frames into HSV images, because it is related to human color perception. Basically the color spaces. differentiate into three components: hue (H), saturation(S), value (V). Image segmentation is typically performed to locate the hand object and boundaries of images, for this HSV features helps user to specify boundary of skin color in terms of hue and saturation value. V value gives brightness information so therefore it is easy to classify skin color and non-skin color information in images. In this approach adjusting value of HSV within range 0 to 255 to extract and get accurate boundary of object.[4]

VI.5. NOISE REMOVAL EROSION AND DILATION

The set of operations which performs on the image based on shapes are known as Morphological operations. There are two most basic morphological operations: Erosion and Dilation, it uses for Removing noise, Separation of individual elements and joining misaligned elements in an image, even Finding of intensity bumps or holes in an image. Erosion shrinks boundaries of an image and enlarges holes; Erosion can be used to remove noises from an image. And Dilation is used to add pixels at region of boundaries or to fill in holes which generate during erosion process. Dilation can also be used to connect disjoint pixels and add pixels at edges.[5]

VI.6. THRESHOLDING

Thresholding is the simplest method of image segmentation. From a grayscale image, thresholding process can be used to generate binary images. In thresholding each pixel in image

replace into black pixel, if image intensity is less than some constant and a white pixel if intensity is greater than constant value. A primary property which pixels in image can share its intensity. Hence in thresholding images separate into regions depending on Light and dark regions.

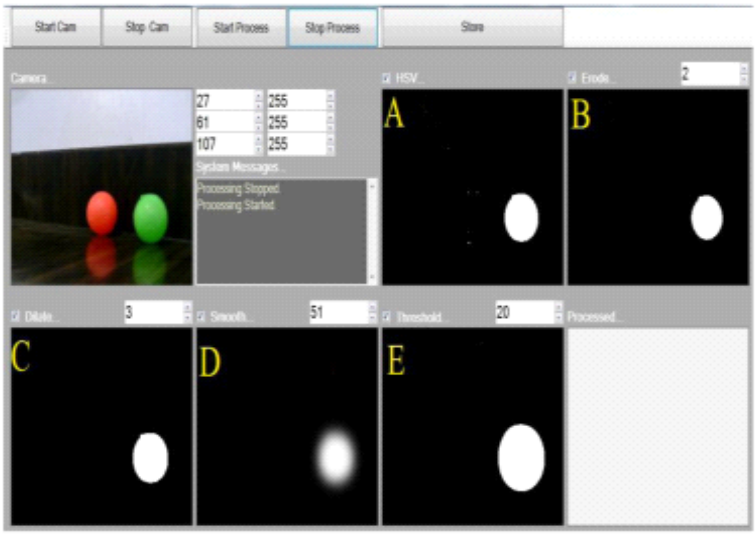

Figure 3 :Image Preprocessing: Color Filtering A)HSV image B)Erosion C)Dilation D)Smoothing E)Thresholding

VI.7. BLOB DETECTION

In the field of computer based vision, blob detection is refers to detection of points/regions in the image which either brighter or darker than surrounding region. Basically blob is defined as a collection of pixels organized into a structure. It is a detection of points/regions in the images which differs in features like brightness and color. Title Blob Detection part thoroughly divided into Fill Holes() and Detect And Recognize Hand() Methods. In Fill Holes() Biggest Contours are created which are used as final Contour. In Detect And Recognized Hand the biggest contours will be drawn using HSV format .It will be represented in BITMAP(0,1) and rectangular shape.

VI.8. CONTOUR DETECTION

In contour detection convexity hull algorithm uses for drawing contour around the palm and finger points detection. In convexity hull algorithm adaptive boosting algorithm use for hand detection and can use haar classifier algorithm to train classifier. Initial step in convexity hull algorithm is to segment image in which hand is located. For this some feature must be assume. Here assumed shape of hand but that may change according to movement of hand. Therefore, skin color of hand is considered, because it is invariant to scale and movement of hand. The next phase of a tracking system contains separating hand pixels from non-hand pixels. Before segmentation occurs, filter all captured images with a Gaussian filter [6] and then scales this filtered image by the non-changing background scene. And after segmentation contour is extracted [7].

Figure4 :Contour Extraction

Sign language when evolved is different from spoken language so the grammar of the sign language is primarily different from spoken language. Inspoken language, the structureof the sentence is one-dimensional; one word followed by another, while in sign language, a simultaneous structure exists with aparallel temporal and spatial configuration. As based on these characteristics, the syntax of sign language sentence is not as strict as in spoken language. Formation of a sign language sentence includes or refers to time, location, person, base. In spoken languages, a letter represents a sound. For deaf nothing comparable exists. Hence the people, who are deaf by birth or became deaf early in their lives, have very limited vocabulary of spoken language and faces great difficulties in reading and writing.

VI.9. NEURAL NETWORK DESGIN

A neural network is basically modelled as the structure shown in figure2, in which can be observed a group of elements that interact to generate an output vector from an input vector which is described by the variable x. The training information is stored in the set of synaptic weight values of the neural network, and the output neuron is limited to a specific range of values of the activation function.

Output neuron can be described by

$$y_k = \phi(\sum w_{ki}.x_i + w_o), \qquad (1)$$

or

$$y_k = \phi(v_k) \qquad (2)$$

where the subscript i indexes units in the input layer, k in the hidden; w ki denotes the input to hidden layer weights at the hidden unit k. An adder ∑ which produces the weighted sum of inputs according to the respective weights of the connections. A activation function defines the output amplitude of that node given an input or set of inputs $\varphi(v_k)$, and w_o is a threshold value.

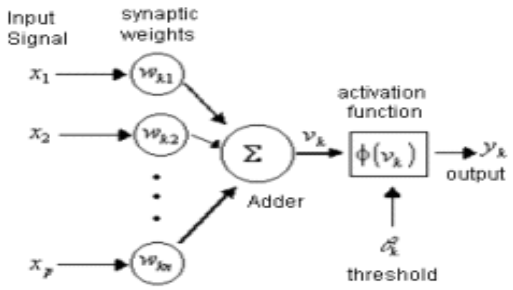

Figure 5 : Neural Network Model

A multilayer neural network was used in the design with a backpropagation algorithm. The structure of the network is formed by three layers, called the input layer, hidden layer and output layer; the basic components can be seen in figure 3 in which was used a simplified graphic notation. For the input and hidden layer neurons were employed a hyperbolic tangent activation function with 5 neurons each one and one neurons at the output layer with a linear activation function.

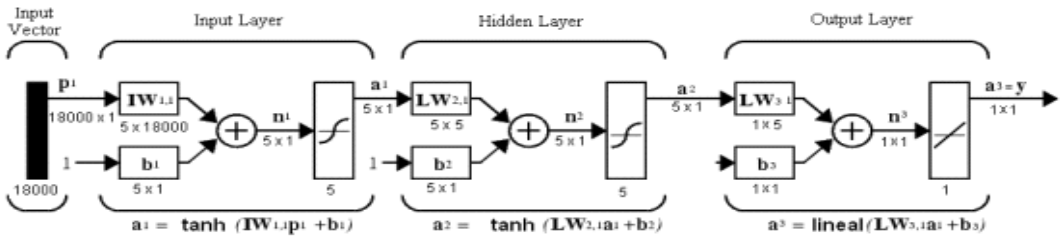

Figure 6: Employed Multilayer Neural Network

VII. TRAINING PROCESS

During the training process of the first stage, it is used a backpropagation algorithm. The supervised backpropagation learning scheme modifies the weight in the opposite direction of the gradient of the error function to minimize a mean squared error of whole patterns, which are used to train the neural network. These algorithms build models that predict the desired values. Its an gradient based algorithm, which start with the initial weight vector, estimates the error function and its gradient for training, and it is obtained a new modified weight vector. This is repeated till the error finds the set limit [8]. Therefore, by definition, the weights are updated through the expression:

$$w^{m+1}=w^{m}+\alpha(-\nabla^{m}), \qquad (3)$$

where is α the learning rate of the network, and ∇ gradient of error function about to w_m .In the backpropagation algorithm is used the mean squared error that is calculated from a desired output m d as:

$$(e^{m})^{2}=(d^{m}-w^{m}.x^{m})^{2} \qquad (4)$$

therefore, the gradient is obtained from the error

$$\nabla^{m}=-2.e^{m}\,\phi(v^{m}).x^{m} \qquad (5)$$

Replacing in (3), it is obtained the following expression:

$$w^{m+1}=w^{m}+2.\alpha.\phi(v^{m}).x^{m} \qquad (6)$$

This process is made for all the neurons of each layer in the network.

VII.1. PRE TRAINING CNN MODEL

The concept of Transfer learning is used, in which the model is first pre-trained on a dataset and then it is different from the original. In this way the model gains knowledge to other neural networks can be transferred. The knowledge gained by the model, is in the form of "weights"can be saved and it can be loaded into some other
model.For feature extraction pre-trained model are used by adding fully-connected layers on top of it. The model are trained with the original dataset after loading the saved weights.

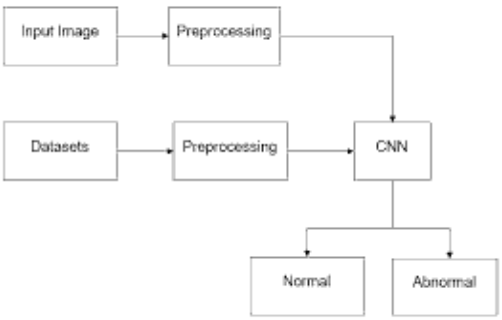

Figure 7:CNN pre-training model

VII.2. PCA (PRINCIPAL COMPONENT ANALYSIS)

Using PCA, data is projected over a lower dimension for reduction of dimension.The most important feature is the one with the largest spread or, as it corresponds to the largest entropy and thus encodes the most
information. Thus the dimension with the largest variance are kept and others are reduced.

VII.3. LBP (LOCAL BINARY PATTERNS):

It computes a local representation of texture which is constructed by comparing each pixel by its surrounding or neighbouring pixels. The results of this are stored in the form of array and then it is converted into decimal and stored as an LBP 2D array**.**

VII.4. HoG (HISTOGRAM OF GRADIENTS):

A feature descriptor is a representation of an image or an image patch which simplifies the image by extracting useful information and throwing away extraneous or non useful information.Hog is a feature descriptor which calculates a histogram of gradient for the image pixels, which is a vector of 9 bins (numbers) to the corresponding angles such as 0, 20, 60... 160. These images are divided into cells, (usually, 8x8), and for each cell, by which gradient magnitude and gradient angle is calculated, using which a histogram is created for a cell. The histogram of a block of cells are normalized, and the final feature vector for the entire image is calculated.

VIII. ALGORITHM USED:

VIII.1 CONVOLUTIONAL NUERAL NETWORK:

Convolutional Neural Network is a deep learning technique which is developed from the inspiration of visual cortex which are the fundamental blocks of human vision. It is observed from the research that, the human brain performs a
large-scale convolutions to process the visual signals received by eyes, based on this observation CNNs are constructed and observed to be outperforming all the prominent classification techniques .Two major operations performed in CNN are convolution ($wT* X$) and pooling(max()) and these blocks are wired in a highly complex fashion to mimic the human brain .The neural network is constructed in layers, where the increase in the number of layers increases the network complexity and is observed to improve the system accuracy .The CNN architecture
consists of three operational blocks which are connected as a complex architecture.

The functional blocks of Convolutional Neural Network:

1.Convolutional Layer
2.Max Pooling layer
3.Fully-Connected layer

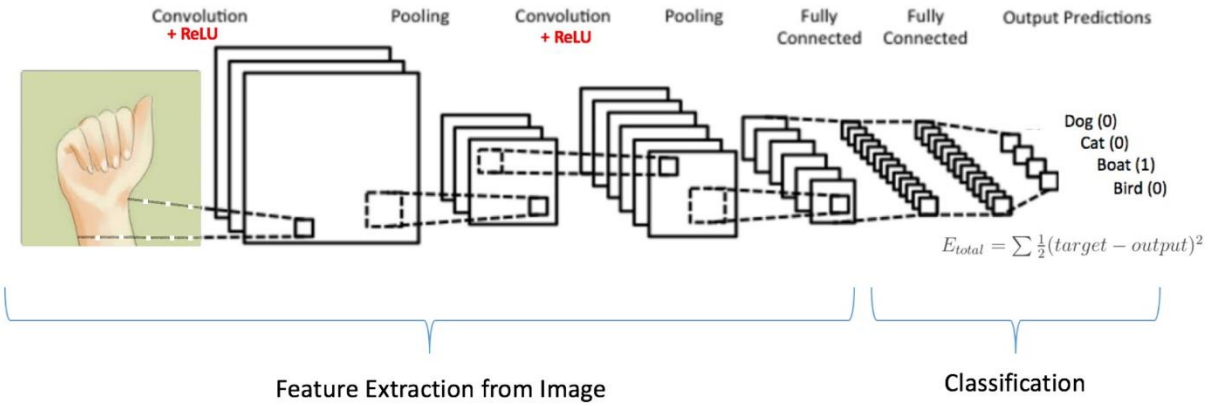

Figure 8:Convolutional Neural Network

IX. CONCLUSION:

Sign Language Recognition System has been developed from classifying only static signs and alphabets, to the system that can successfully recognize dynamic movements that comes in continuous sequences of images. Researcher nowadays are paying more attention to make a large vocabulary for sign language recognition systems.

Many researchers are developing their Sign Language Recognition System by using small vocabulary and self-made database. Large database build for Sign Language Recognition System is still not

available for some of the country that involved in developing Sign Language Recognition System.The neural networks are one of the more powerful tools in the identification system and pattern recognition. The system presents a performance pretty good to identify the static images of the sign alphabetic language. The system shows that the first stage can be useful for deaf persons or with speech disability for communicating with the rest of the people who do not know the language. In this work, the developed hardware architecture is used as image recognizing system but it is not only limited to this application, it means, the design can be employed to process other type of signs. As future work, it is planned to add to the system a learning process for dynamic signs, as well as to prove the existing system with images taken in different position. Several applications can be mention for this method: finding and extracting information about human hands, which can be apply in sign language recognition that it is transcribed to speech or text, robotics, game technology, virtual controllers and remote control in the industry and others.

ACKNOWLEDGMENT

We express our sincere thanks to guide Asst. Professor Aarti Bakshi whose supervision, inspiration and valuable guidance helped us a lot to complete our work. Her guidance proved to be the most valuable to overcome all the hurdles in the fulfillment of this paper work. Also, we are thankful to all those who have helped us in the completion of paper work.

REFERENCES

[1] . Goodman J W 1968 Introduction to Fourier optics McGraw Hill

[2] . N.Otsu, “A Threshold Selection Method from Gray-LevelHistograms”, IEEE transactions on systems, man, and

[3] . Rafiqul Zaman Khan and Noor Adnan Ibraheem , “HandGesture Recognition: A Literature Review”, InternationalJournal of Artificial Intelligence & Applications (IJAIA),Vol. 3, No. 4, pp. 162-174, July 2012

[4] . M. Panwar, "Hand gesture based interface for aiding visuallyimpaired," Proc. IEEE Int. Conf. Recent Adv. Comput. Softw.Syst. RACSS2012, pp. 80-85.

[5] . Aliaa A. A.Youssif, AmalElsayedAboutabl, HebaHamdyAli,"Arabic Sign Language (ArSL) Recognition System UsingHMM ", (lJACSA) International Journal of AdvancedComputer Science and Applications, Vo1.2, Issue. 11,2011

[6] .L. Gu, X. Yuan, and T. Ikenaga, "Hand gesture interface basedon improved adaptive hand area detection and contoursignature," IEEE Int. Symp. Intel!. Signal Process. Commun.Syst. (ISPACS 2012), no. Ispacs, pp. 463--468

[7] . H. Y. Lai and H. J. Lai, "Real-Time Dynamic Hand GestureRecognition," IEEE Int. Symp. Comput. Consum. Control,2014no. 1,pp. 658-661

[8] .Maeda Y, Wakamura M 2005 Simultaneous perturbation learning rule for recurrent neural networks and its FPGA implementation IEEE Trans. Neural Network 16 6 1664 – 1672.

[9] .Gallaudet University Research Institute: A Brief Summary of Estimates for the Size of the Deaf Population