2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) | 978-1-6654-7413-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSES55317.2022.9914155

Sign Language Recognition using Machine Learning

Manikandan J, Brahmadesam Viswanathan Krishna, Surya Narayan S, Surendar K

Computer Science and Engineering Rajalakshmi Engineering College Thandalam, Chennai, India

E-mail : jmanekandan@gmail.com krishna.bv@rajalakshmi.edu.insuryanarayan.s.2018.cse@rajalakshmi.edu.in surendar.k.2018.cse@rajalakshmi.edu.invijayk.btech@gmail.com

Abstract-One of the ways to communicate with deaf and dumb individuals is through sign language. So, in order to speak with deaf and dumb people, one should learn sign language; yet, because not everyone can learn it, communication becomes nearly impossible. The goal of this study is to use machine learning to break through these communication hurdles. The majority of existing technologies rely on external sensors, which are out of reach for most people. We utilize OpenCV to take images and the CNN technique to train the machine, with the output being text. Many previous studies have offered methods for partial sign language identification, however this study intended for the full acceptance of American Sign Language comprises of 26 letters and 10 numbers. The majority of ASL letters are static, however few are dynamic. As a result, the goal of this research was to extract features from finger and hand motions in order to distinguish between static and dynamic gestures.

Keywords—Computer vision, imbalanced distribution, ML algorithm, hyperparameters.

I. INTRODUCTION

Communication is an important component of our day-todaylives since it allows us to share information. Asspeculated Nelson Mandela's quote "Talking to a man in alanguage he knows goes straight to his head." If you speak tohim in his native tongue, you will touch his heart." Languagehas existed since the birth of civilization and is indisputablyfundamental to human communication. It is a mediumthrough which individuals express themselves andcomprehend real-world concepts. People with hearingimpairments are often overlooked and left out in today's fastpacedculture. They must fight to communicate themselvesto those who are different from them, to bring up theirthoughts, to speak their opinions, and to express themselves[1]. Even though sign language is a means of communicationfor deaf and dumb individuals, it has no significance for nonsignlanguage users. Around 60 million Indians are deaf anddumb. The majority of individuals communicate with normalpeople through signals, which are extremely difficult tointerpret and make communicating information impossible.We are proposing a sign language recognition technology toprevent this from happening. It will be a fantastic tool forpersons with Hearing impairments can use sign language tocommunicate their thoughts, while non-sign language userscan understand what the latter is saying. The first three stepsof a sign language recognition system are database creation,classification, and prediction [2]. We use the American signlanguage as a database in our project to allow deaf and dumbpeople to communicate. We need a sign database with 26English alphabet signs and 10 numeric signs with properimages for this project. A specific image is assigned to eachnumber or alphabet. These images are in.jpg format, and theywere collected as greyscale images during the collectionprocess. These excreted images are fed into the model asinput. CNN is used to train the machine, and the output is atext.

Technically, generating descriptors to convey hand formsand motion trajectory is the most difficult aspect of signlanguage recognition. Hand-shape description, in particular,entails tracking hand regions in a video stream, segmentinghand-shape images from a complicated background in eachframe, and problems with gesture detection. The tracking ofcritical points and curve matching are also related to motiontrajectory. Despite the fact that numerous research studieshave been undertaken on these two topics to date, SLRremains difficult to achieve satisfactory results due to thevariance and occlusion of hands and body joints.Furthermore, integrating the hand-shape and trajectoryinformation together is a difficult task. To overcome theseissues, we created a 3D CNN that can organically incorporatehand forms, action trajectory, and facial expression. Ratherthan using commonly used color images as input to networkslike [1, 2], We use a mixture of color images, depth images,and body skeleton data from the Microsoft Kinect to createour final product. When a person moves, the Kinect motionsensor produces both a color scheme and a depth stream. Thelocations of the body joints can indeed be procured in realtime by leveraging the public Windows SDK. As a result, wechose Kinect as the capture device for the dataset of signwords. Color and depth changes at the pixel level provideuseful information for distinguishing between distinct signactivities. And the trajectory of sign actions can be depictedby the fluctuation of bodily joints in time. CNNs pay attentionto changes not only in color, but also in depth and trajectory,when many types of visual sources are used as input. BecauseCNNs have the ability to learn features automatically fromraw data without any prior information, we may bypass thedifficulties of tracking hands, segmenting hands frombackground, and creating descriptors for hands [3]. In recentyears, 3D CNNs have been used to classify video streams [2,4, 5]. CNNs may be concerned about the amount of time ittakes to do a task. Training a CNN with a million-scale inmillion videos takes several weeks or months. Fortunately,using CUDA for parallel processing, it is still possible toattain real-time efficiency.

II. Related Work

The Following a literature review on the subject, it becomesclear that a variety of strategies have been developed to tacklethe issues of gesture identification in video. Hidden MarkovModels (HMM) were used in conjunction with BayesianNetwork Classifiers and the Gaussian Tree Augmented Naïve Bayes Classifier in [5] to differentiate between different facialexpressions captured in video clips.

In particular with regard, Francois et al. [6] presented theirwork on Human Posture Recognition in an Image SequencesUsing 2D and 3D Appearance Methods, on PatternRecognition. According to the research, PCA is being used toacknowledge silhouettes from a static camera, and then 3D isbeing used to model posture in order to identify the silhouette.This method has the drawback of necessitating the use ofintermediate step gestures, which can lead to ambiguity duringtraining as well as, as a result, decreased prediction accuracy.

The analysis of video segments, which encompasses theextraction of visual information in the form of feature vectors,is a common application of neural networks. Hand tracking,segmentation of people from the background andsurroundings, variability in lighting, occlusion, movements,and position are all issues that Neural Networks have to dealwith when working with images. According to Nandy et al.[7], they split the dataset into segments and extract featuresbefore categorising them using Euclidean Distance and KNearestNeighbors, respectively. In a similar vein, Kumud etal. [8] describe Continuous Indian Sign Language Recognitionin the same way.

III. Methodology

In our model we give our own data as the data set to trainthe modules and the data set will be pre-processed. We applythe pre-processing approach in MATLAB to normalise theluminance of individual particle pictures and remove low frequency background noise. Firstly, we convert RGB imagesinto grey scale images by eliminating the saturation and huewhile retaining the luminance. The Feature extraction takesplace, by using CNN to extract information from the framesand forecast hand gestures, a model is utilised. It's a multilayeredfeed forward neural network used primarily for imageidentification. The CNN design is made up of numerousconvolution layers, each of which has a different function. Hasa pooling layer. Our proposed system comprises of thefollowing major steps.

- Creating a Dataset
- Pre-processing
- Feature Extraction
- Applying Model

3.1 CREATING A DATASET

We are creating the dataset in this project. Each frame thatdetects a hand within the ROI (Region of Interest) formed maybe saved in a directory that has two folders, train and test, eachwith 36 files containing recorded images of numerals from 1to 10 and alphabets from a to z. Now, we'll use OpenCV tocollect the live cam stream in order to create the dataset. Andcreate a ROI, which is a We want to detect the hand for themotion in this part of the frame.

3.2 PREPROCESSING

After receiving the image from the user, we must preprocessit. The pre-processing procedure is used to removelow-frequency background noise and normalise the intensityof individual particle pictures. RGB photos are converted togreyscale images. We can execute noise removal andsegmentation operations at this step. The basic goal of preprocessingis to reduce input data distortion and development(sign language images). The image preprocessing approachtakes advantage of image redundancy. A neighbouring pixelin the actual image that corresponds to one object has beenmodified to a similar brightness value. The median filter isused to minimise salt and pepper noise in photos duringpreprocessing.

3.3 FEATURE EXTRACTION

Every image contains a vast amount of data, and featureextraction is the process of automatically extracting this datafrom the images. The input data that will be processed isreduced to a smaller collection of features. Feature extractionis the term for this method. The Feature Extraction stage isrequired because certain features must be extracted in orderfor each gesture or sign to be unique.

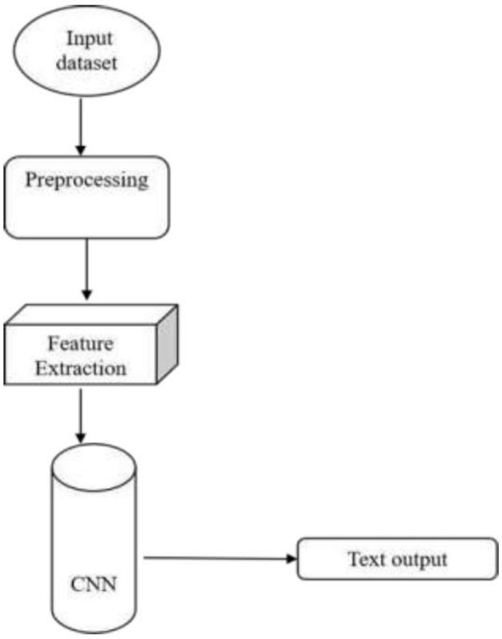


Fig. 1. Proposed System workflow

3.4 CONVOLUTIONAL NEURAL NETWORK

The goal of CNN is to understand the technical in theinformation with higher intelligence using convolution layerin addition to learning the engaging in positive in the data. TheProposed system performs admirably when it comes to objectrecognition, including image recognition. They are capable ofdistinguishing between individual people, continues to face,road signs, and other aspects of visual information. Asillustrated in Fig. 2, there are many different CNN variations,but each one is based just on sequence of layers that is presentin the system. The Convolution layer is made up of severalcomponents, each of which has a different set of two – layerand processing elements to perform. The following sectionexplains the objectives and functionality of some frequentlyused layers, which are described in greater detail below. Theconvolutional layer is a type of layer. The convolutional layeris one of the fundamental elements of CNN architecture. It ispossible to alter the input data using convolutional layers(Conv), which are made up of a patch of neurons that areconnected domestically from the preceding stage. In theoutput nodes, the dot product will indeed be calculated by thelayer that sits between both the area of neurons that areprevalent in the input nodes and the weight training that areprevalent in the output layer to which they would be regionallyconnected.

A convolution is a mathematical function that characterizes the principle for combining two sets of data into a single largerset of information. According to Fig. 3, the convolution readsthe input an extracted feature, appears to apply a convolutionfiltration or kernel, and gets back an extracted features as anoutput as has been shown. It is demonstrated in this operationhow sliding the kernel from across data input results in theconvoluted expected output. At every step, the data inputvalues are magnified by the operating system within itsboundaries, and a specific value in the predicted output iscreated from the result of the multiplication.

The identifying the determinants is the single most importantlayer. Because this automatic detection problem is a multiclassificationproblem, the softmax function is being used inthe output nodes to classify the results. Finally, the fully -connected layer with 100 neurons is being used to quantifythe class scores, which is the fully connected layer. Thenumber 1000 refers to the total number of classes in the dataset in this case. In general, the CNN architecture consists offour main layers: a convolutional layer, a pooling layer, aReLU layer, and a fully connected or output layer. Theconvolutional layer is the first of these layers. Testing of thesuggested model was carried out on approximately 50different CNN models by varying hyperparameters such asfilter size, stride, and padding in the manner described in thispaper, and the results were encouraging. A number ofdifferent combinations of convolutional and pooling layershave been tried out to see how well the system works. In orderto improve the effectiveness of the results, an additionallayer, referred to as the dropout layer, is included in theproposed approach. The dropout layer is a training algorithmwhich is used to dismiss randomly chosen neurons during thetraining process, thereby reducing the likelihood of overfitting.

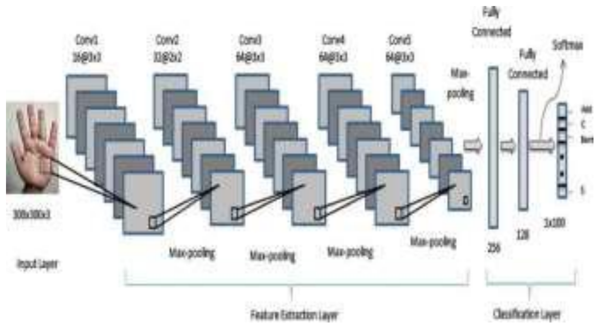


Fig. 2. Schematic representation of CNN

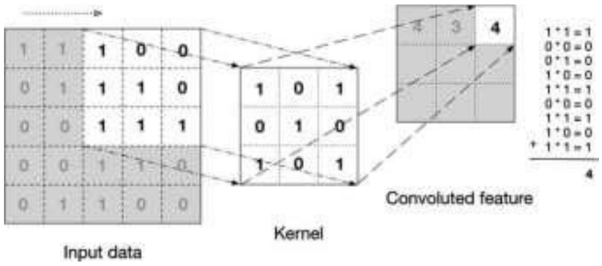


Fig. 3. General Convolutional operation

$$Outcome = \frac{W-G+2P}{T} + 1 \qquad \{1\}$$

The outcome with respect to convolutional operation is represented in eqn. 2.

$$b_j^t = g\left(\sum_{j \in b_j} x_j^{t-1} \times m_{ij}^t + c_j^t\right) \qquad \{2\}$$

$$g(y) = \max(0, y) \qquad \{3\}$$

IV. Experimentation & Result Discussion

Two separate experiments are used to assess therecognition method for Indian Sign Language performed well.To begin, the parameters used to train the model have beenfine-tuned, including the layer count, filters, and optimizersThe trained model's performance was tested in the secondexperiment.is assessed on both colour and grayscale imagedatasets. The ISL recognition system's average precision,recall, F1-score, and accuracy have also been calculated.

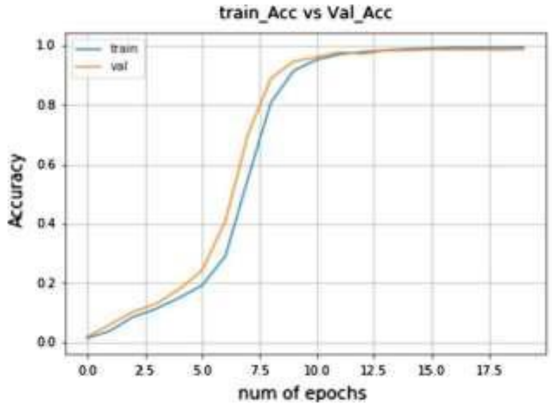


Fig. 4. Training Accuracy of our proposed model

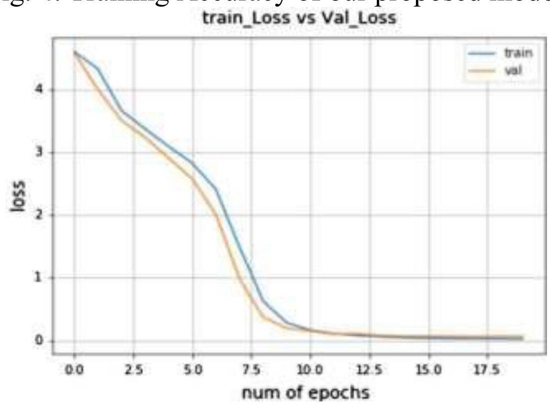
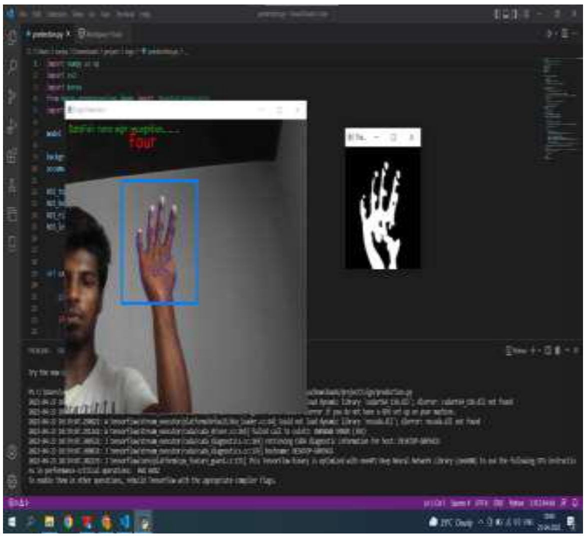
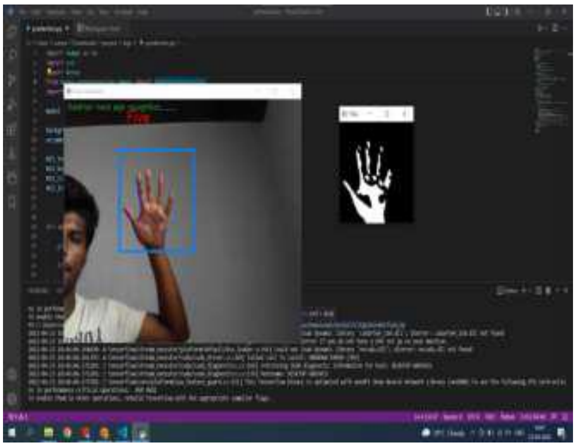


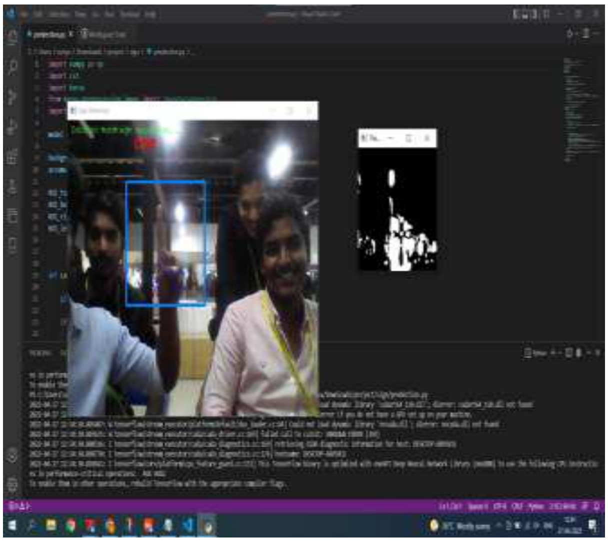
Fig. 5. Training loss of our proposed model

Result Screenshot

[Note: The red color text above the bounding box displaysthe output of the hand sign shown].







V. CONCLUSION

Number and character images are used to represent theidentified hand motions. Ten different numbers and alphabetsare used in the experiment. The accuracy of the systemdescribed above is more than 90%. For overall performanceanalysis of suggested Systems, general performancemeasures such as False Accept Rate and False Reject Rate arechosen. The rate of false acceptance and rejection is less than2%. In the dynamic recognition approach, 10 photos areincurred in real time (10 occurrences or 10 for each class).For the database, we receive 3100 photographs. During thetesting, all of the photos save the fifth one yielded the properresult. The rate of sign recognition is 95%. The rate of resultidentification varies depending on the testing sample'sConvolutional neural network deep learning model is used inthis Recognizing system. The coding is done in real time,using a consistent light source and a plain white background.

REFERENCES

[1] Nguyen Dang Binh and Toshiaki Ejima "Hand gesture recognition using fuzzy neural network (2005)" in Cognitive Informatics, 2006. ICCI 2006. 5th IEEE International Conference on Volume: 1

[2] LiuYun and Zhang Lifeng" A hand gesture recognition method based on multi-feature fusion and template matching (2012)" in Advances in neural information processing systems, 2012, pp. 1097–1105.

[3] Mahmoud Elmezain, Ayoub Al-Hamadi and Bernd Michaelis "Real-time capable system for hand gesture recognition using Hidden Markov models (2008)" in Pattern Recognition, 2008. ICPR 2008. 19th International Conference.

[4] Simon Lang, Marco Block- Berlitz and Raul Rajos " Sign language recognition and translation with kinect (2012)" in Proceedings of the 11th international conference on Artificial Intelligence and Soft Computing - Volume Part I.

[5] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S, Chen and Thomas S. Huang (2003, February). Facial expression recognition from video sequences: temporal and static modeling. Computer Vision and Image Undertaking 91.

[6] Bernard Boulay, Francois Bremond, MoniqueThonat, Human Posture Recognition in Video Sequence. IEEE International Workshop on VS PETS, Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2003, Nice, France.

[7] Recognition of Isolated Indian Sign Language Gesture in Real Time, Anup Nandy, Jay Shankar Prasad, Soumik Mondal, Pavan Chakraborty,G. C. Nandi, Communications in Computer and Information Science book series (CCIS, volume 70)

[8] Continuous dynamic Indian Sign Language gesture recognition with invariant backgrounds by Kumud Tripathi, Neha Baranwal, G. C. Nandi at 2015 Conference on Advances in Computing, Communications and Informatics (ICACCI).