e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

A REVIEW ON SIGN LANGUAGE RECOGNITION

Aleena Shaji^{*1}**, Ms. Resija P.R**^{*2}

^{*1}Department Of Computer Science, Vimala College (Autonomous) Thrissur, Kerala, India.

^{*2}Assistant Professor, Department Of Computer Science Vimala College
(Autonomous) Thrissur, Kerala, India.

DOI : https://www.doi.org/10.56726/IRJMETS45608

ABSTRACT

True impairment is the inability to converse with others. [1]Deaf and dumb people mostly utilise [1] sign language to [1] communicate inside [1]their group [1]and with other people. People who are hesitant to converse or hear use hand gestures to communicate in this language. Hands, eyes, facial emotions, and movement are all used as visual clues within this form of language. Even though sign language has grown to be increasingly widespread in recent years, it can still be difficult for non- signers to interact with signers. A technique for [2]sign language recognition is presented [2]in this study. Recognizing [2]sign language through actions [2]is referred to as sign language recognition. For this, capturing sign language gestures for words, numerals, and alphabets using a camera. The recorded images are processed using supervised algorithms and various data sets. The pre- trained models and the different layers of these algorithms are used for feature extraction and classification. The system will provide the equivalent sign language in text or speech as its output.

Keywords: SL, DATA SET, Cross Validation, Artificial Neural Network, HU's Moments, Skin Segmentation, SVM, Deep Learning, Convolutional Neural Networks, Computer Vision, Sign Language, Pre-Trained SSD Mobile Net V2, Sign Language Recognition, Text To Speech.

I. INTRODUCTION

People with impairments communicate their feelings and thoughts to others through gestures rather than words. [3] Deaf people utilise sign language as their primary form of communication. People who are deaf have very little, if any, hearing. [4]Sign language consists of various [4]hand gestures using hands [4]and facial expressions. In contrast to spoken or written language, [5]sign language has its own vocabulary, meaning, [5]and syntax. Around the world, 466 million people lack hearing. 34 million of them are kids. This language is difficult for ordinary people to comprehend. During the educational and training sessions needs trained sign language expertise. Around 138 to 300 types of sign languages are used globally. Everyone uses sign language according to their place.

By recognizing sign language, we mean converting such languages into corresponding spoken language as text or speech. This uses the growth in [1]deep learning, machine learning, and [1]computer vision. It uses various supervised learning algorithms like SVM, CNN, etc. The existing data sets like MNIST, ASL Alphabet, and Sign language gesture images are also user-defined data sets according to the corresponding methods. [6] For the series of image processing the images are captured using webcams. After image capture, pre-processing, conversion, and feature extraction are done. Following the feature extraction this is put into the supervised learning algorithms for the conversion of appropriate form. The pretrained model SSD Mobile net v2 and custom CNN model are mainly used for gesture identification from sign language. This proposed strategy overcomes many existing methods and limitations and provides comfort to normal people

II. RELATED WORKS

A. Sign language Recognition Using Machine Learning Algorithm [2020]

[7] An automatic [8]system for recognizing gestures in sign language is demonstrated in this paper. The goal [8]of this paper is to identify [8]the movements in [8]sign language and translate them into text format. There are mainly three stages: In the first stage the skin part is segmented from the images and the remaining parts are considered as noise. The second stage [9]is used to extract relevant [9]features from [9]the segmented [9]image and in the last stage [9]the extracted features are put into various supervised learning algorithms for training then the classification part is done. The above-mentioned stages are performed on the collected data

sets. UCI skin segmentation data set is used which contains 2,00,000 points. SIFT features are used to find the key points from the image. SVM, Random Forest methods are used to extract features. SVM achieves 46.45% accuracy and the larger number of classes is the reason for the lower accuracy

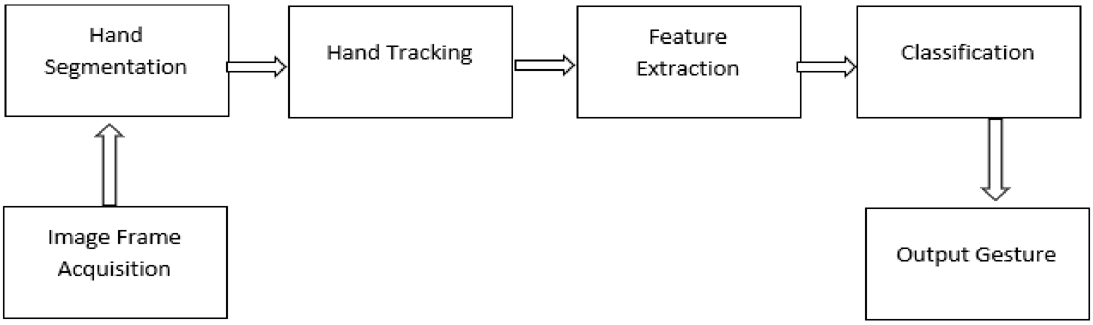

Fig. 1. Proposed Method

As mentioned in the above Fig.1 Webcam [10] is used to capture the hand gesture. The next [10]phase is [10] image pre- processing which means it includes cropping, filtering, brightness, etc. [10]Image cropping and [10]Image segmentation methods are used for this. The [10]captured image is [10]in RGB form [10] so it needs to convert [10] into Binary format. Then cropping is done by removing unwanted parts from the image. [10]In image segmentation, the [10]Edge detection method is used which identifies [10]the boundary of the [10]cropped image. For letter image similarity characterization both [11]global visual features and local visual features are extracted. In [11]sign recognition mainly two types of feature extractions are there, one [12]is Contour-based shape representation and description methods and [12]another is Region-based shape representation and description methods. Seven moments are identified using the method 7Hu.

The algorithm used in this is SVM. It is used for classification. On the basis of labelling the given data is divided as follows: if the data is labelled it puts into the supervised category otherwise it puts into the unsupervised category.

B. Sign Language Recognition Using Deep Learning and Computer Vision [2020]

[13] This work tries [1]to recognize sign language motions [1]and translate [1]them into text [10]using computer vision and deep learning [14]. Collection of handwritten digits data set namely MNIST is [15]used for training different [15]image processing systems. The MNIST [15] data contains 60,000 training images, 10,000 testing images, etc. In this, only 24 classes of letters are there so J and Z are excluded. Here, the video is captured as input using a webcam, and the image is segmented from that. Then it is converted into grayscale images.

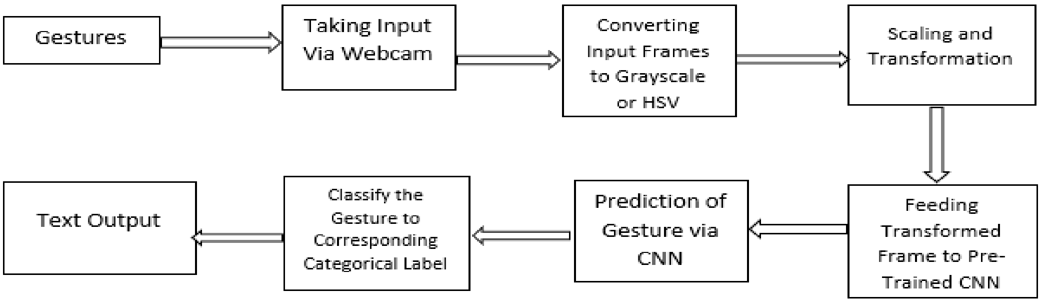
Gestures
Taking Input Via Webcam
Converting Input Frames to Grayscale or HSV
Scaling and Transformation
Feeding Transformed Frame to Pre-Trained CNN
Prediction of Gesture via CNN
Classify the Gesture to Corresponding Categorical Label
Text Output

Fig. 2. Methodology

As per the Fig.2 Converted images are given into the pretrained custom CNN model. The gesture prediction comes from this model and later it is classified based on the label. In the final stage, the classified [16]gesture is displayed as text. [16]CNN model contains [16]11 different layers. [16]Four Convolutional layers, Three Pooling (Max) layers, Two Dense Connected, One Flatten, [16]and One Dropout layer. The initial convolutional layer has a convolutional kernel (7,7) followed by a convolutional layer (3,3). (2,2) Pool efficient layers exist for each max pooling layer. To avoid negativities the flattening layer followed by the fully connected layer with ReLu also boards a pool (2,2). To avoid overfitting of the model the dropout layer gives 20% probability. MNIST ASL data set is used to train the CNN model.

C. Real Time Sign Language Detection [2021]

[17] This paper presents a sign language detection [2] system that take input through a webcam and processes it for detection. A user-defined data set is used for this purpose. It includes 2000 images and [18]a total of 5 symbols that are Hello, Yes, No, I Love You, and Thank You [18]. OpenCV is used to capture images of hand gestures using a webcam. Labelling is performed as the next step, using a [19]pre-trained model called [19]SSD Mobile Net v2 for sign recognition [19]. Capturing different images [19]of multiple [19]sign language symbols from various angles. Bounding boxes are enclosed in the particular sign language that is selected from the entire image. Labelling is performed with the help of a labelling tool. The next step is the [8]selection of images for training and testing then [8]TF records are [8]created. In the final stage, [8]the classification is carried out.

The algorithm used in this paper is the convolutional neural network. It selects images and puts importance on various aspects of that image. It requires less amount of pre-processing as compared to other classification techniques. [20] The main aim of CNN is to extract relevant features from images for easier image processing. CNN [21]is made up of three different [21]layers: convolutional layers, pooling layers, [21]and fully connected layers. The tool used for this paper is TensorFlow, Object Detection API, Open CV, and Labelling. [8]Transfer Learning and pre-trained model SSD Mobile [8]net v2 is used for training the model.

After training the models we get the results like: [8]Yes88.7, No- 88.6, Thank You- 84.1, Hello- 91.0, I Love You- 82.4.

D. Sign Language Recognition [2021]

[22] In this paper create [4]a practical and meaningful system that can understand sign language and translate it into text. Because of this it uses different data sets to achieve good accuracy. The first [23] ASL [24]Alphabet dataset [24]is a collection of alphabet [24]images from American sign language. It has [24]29 classes, 26 [24]of which are English alphabets, and the rest are SPACE, DELETE, and NOTHING. The second dataset, Sign Language Gesture Images, contains 37 different hand gestures. They are the alphabet A- Z, numbers 0- 9, and a gesture for SPACE. Each gesture contains 1500 images. CNN is used to train this dataset.

Images are 2- dimensional arrays of pixels [25]ranging from 0 to 255, where 0 represents black and 255 represents white. Images are also represented as mathematical functions. Image pre-processing is required before it is sent to the model training. The means engaged with image pre-processing are reading images, resizing or reshaping images, eliminating commotion, and changing all image pixel arrays from 0 to 255. CNN has two main aspects, feature extraction, and classification.

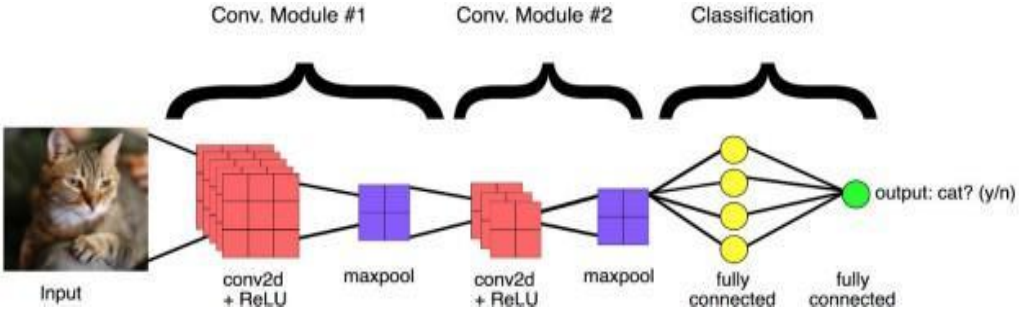

Fig. 3. CNN Architecture

The above Fig.3 describes the main steps of CNN that are Convolution, Pooling, flattening, and Full connection. Convolution is used to extract features using filters like randomly selected edges, highlighted patterns, etc. [24]The pooling layer is used to make [24]the image smaller. [26]Max pooling and Average pooling are the two different forms of pooling layers. These two categories of layers are employed to extract the matrix's maximum and average pixel values, respectively. By [24]inputting the layer to the next layer, the data is converted into 1-a dimensional array using flattening. The prediction is performed using a fully connected layer. In this architecture we trained 3 models which are LeNet-5, [27] Mobile Net V2, and our own architectures with 2 different data sets. The ensemble technique, Horizontal Voting is used to train these models and make predictions using models. [28]The class with the most votes will be [28]the final prediction [28].

III. COMPARATIVE STUDY

Sign languages are languages in which meaning is conveyed through the visual-manual mode rather than spoken words. Manual articulation and non-manual markers are used to express sign language. With aids for the hearing impaired, sign language recognition aims to provide an effective and precise method for converting sign language into text or voice. The first publication presents an automatic sign language recognition system that uses a dataset from UCI and recognizes the sign language and converts them into text. The used dataset contains 2,00,000 points for training using algorithms like SVM and Random Forest. [29]70% of the data in the dataset is used for training and the remaining 30% of the [29]data is used for testing. SIFT features are used to find key points in the image. Approaches for extracting feature vectors are SVM and Random Forest. It is a 26-class problem and SVM got 4.76% accuracy on the HOG feature, and Random Forest got 46.45% accuracy on the HOG feature. In classification linear kernel SVM is used to classify alphabets into one and two-handed, and it has 95% accuracy. Then for classifying these ones, two-handed alphabets use linear kernel multi-class SVM. The image is first captured in the specified manner and image preprocessing is done, image cropping, segmentation, and edge detection are used here. The last step is feature extraction.

The first paper uses the algorithm SVM, but the second paper uses the CNN algorithm and computer vision developments to convert sign language to text. A collection of handwritten digits dataset that is MNIST dataset is [28]used in this paper. For both [28]training and testing the same dataset [28]is used. With 24 classes of letters, the American sign language alphabet and a number of collections of hand gesture images produce multi-class issues. J and Z are excluded. In the first paper, segmented images are taken using the YUV-YIQ model but in the second paper, 11 layers of CNN are used to segment the image from which video input that is: [30]4 convolutional layers, 3 pooling layers, 2 dense connected, [30]one flatten and one dropout layer. [30]The images are converted into grayscale format. Then prediction, classification, labelling, and display of text are performed. In CNN architecture After the flattening layer, a fully connected layer boards a pool with ReLu activation to prevent negative effects. The dropout layer provides a 20% probability of avoiding overfitting the model. At the end of training the output is taken from the SoftMax layer. Actually, the CNN model is trained with 27455 training samples of 784 features of the MNIST ASL dataset. The second paper achieves accuracy close to that of the first paper. In the third work, a method for recognizing signs from a camera and processing them for recognition is indicated. The existing dataset was used in the first paper and second papers, but in this paper, a user-defined dataset of more than 2000 images was used, which includes 5 symbols: [8]Hello, Yes, No, I Love You, and Thank You. The sign images taken at different angles are given for segmentation and then labelling is done using the labelling tool. Selecting images for training and testing. TF records are made and classification is done. The algorithm adopted by this is CNN. CNN gives importance to different objects of the image. [8]The amount of preprocessing required by CNN [8]is less.

In the second study, the CNN model is used for post-scaling transformation, and in the third study, the CNN model is used for ease of image processing. While 11 layers of CNN were used in the second paper, only 3 layers of CNN are used in this paper that includes: convolutional, pooling, and fully connected layers. Unlike previous papers, TensorFlow tools are used for classification, discovery, and prediction. Uses Object Detection API to detect objects in images. OpenCV is used for tackling computer vision issues. Algorithms are used for training in the first paper, but in this paper 2 methods namely SSD Mobile net v2 and Transfer Learning are used. The result of the training is as follows: 46% accuracy is obtained when 50 images are trained. 52% accuracy for 100 images, 72.5% for 200 images, and 86.4% accuracy for 500 images. In sign recognition training, Yes achieves [8]88.7%, No 88.6%, Thank You 84.1%, Hello 91.0% and [8]I Love You 82.4% accuracy. Although the same algorithms are used in the second paper and this paper, this study does not achieve the same accuracy as the other one. The fourth study suggests a practical system to detect sign language gestures and translate them to text. ASL Alphabets, a 29-class dataset consisting of A- Z English alphabets and 3 classes such as SPACE, DELETE, and NOTHING, and [31]Sign Language Gesture Images dataset with 37 classes consisting of [31]A- Z alphabet gestures, 0- 9 numbers, and space are used here. Image pre-processing is performed operations on images as means of algorithms. Although the same algorithms were used in the second, third, and also this study, they were all for different purposes. In this, feature extraction and classification is done. CNN contains steps like convolutional, pooling, flattening, and full connection. 3 models were trained in this study. The first

LeNet-5 [32]consists of 2 convolutional, average [32]pooling layers, flattening convolutional layers, 2 [32]fully connected layers, and a SoftMax classifier [32]. The second Mobile Net V2 works best on mobile devices also it is efficient. The final model is its own architecture. Horizontal voting is used to make predictions and train models. During training, models achieve prediction accuracy as follows: Mobile Net V2 [24]98.9%, LeNet-5 97%, own model 98%, and [24]ensemble 99.8%.

IV. CONCLUSION

[33]Any kind of nonverbal communication, particularly using the hands and arms, known as sign language is used when spoken language cannot or should not be utilised. Sign language recognition systems are designed to give a quick and accurate means to translate sign language into text or voice. There are a variety of techniques that use [4]advances in computer vision and deep [4]learning to recognize [4]sign language. Here, we may compare a few of those research techniques. The first paper employs a machine-learning approach. Got almost 100% of accuracy from this approach as a result. SVM used for feature extraction achieves 4.76% accuracy with the HOG feature and 46.45% accuracy with the Robert Forest HOG feature. The technology will provide a 100% recognition rate if the intended user has already contributed data to our data sets. A custom CNN model with 11 layers is used to segment the images in the second paper. 99% accuracy at the end of training. The validation accuracy of the model is more than 93%. This method yields a low false positive. The third study proposes a system that gives 70%- 80% accuracy using CNN and for training uses [19]pre- trained model SSD Mobile Net V2. This system is able to give [19]accurate results in [19]controlled light and intensity. As the dataset gets bigger, it can be taken to a large scale. A meaningful nearly 100% system is introduced in the fourth paper. It uses computer vision to recognize gestures by creating CNN architecture. In the future, these systems could also be modified to recognize expressions, body gestures, and other dynamic gestures.

ACKNOWLEDGEMENT

We extend our sincere gratitude to all the teaching staff, Department of Computer Science, Vimala College(Autonomous), Thrissur. for their valuable guidance and support at each stage of the project. We are also thankful to our parents for the support given in connection with the project.

V. REFERENCES

[1] T. Hassya, M. F. Hanif, Alvian, F. L. Gaol, and T. Matsuo, "Investing in products with the greatest demand on online stores during the pandemic," in Inventive Systems and Control (V. Suma, P. Lorenz, and Z. Baig, eds.), (Singapore), pp. 809–815, Springer Nature Singapore, 2023.

[2] W.-Y. Yeh, T.-H. Tseng, J.-W. Hsieh, and C.-M. Tsai, "Sign language recognition system via kinect: Number and english alphabet," pp. 660– 665, 07 2016.

[3] S. Sharma, S. Goyal, and i. sharma, "Sign language recognition system for deaf and dumb people," International Journal of Engineering Research and Technology, vol. 2, pp. 382–387, 01 2013.

[4] I. R. J. O. M. I. E. TECHNOLOGY and SCIENCE, "Irjmets." https://www.irjmets.com/editor.php .

[5] S. Tyagi, P. Upadhyay, H. Fatima, S. Jain, and A. Sharma, "American sign language detection using yolov5 and yolov8," 06 2023.

[6] S. Tamura and S. Kawasaki, "Recognition of sign language motion images," Pattern Recognition, vol. 21, no. 4, pp. 343–353, 1988.

[7] R. S. Shirbhate, M. V. D. Shinde, M. S. A. Metkari, M. P. U. Borkar, and M. M. A. Khandge, "Sign language recognition using machine learning algorithm," 2020.

[8] ResearchGate, "Researchgate." https://www.researchgate.net/ , 2023.

[9] ijraset, "ijraset." https://www.ijraset.com/ .

[10] U. Patel and A. Ambekar, "Moment based sign language recognition for indian languages," pp. 1–6, 08 2017.

[11] Y. Quan and P. Jinye, "Chinese sign language recognition for a visionbased multi-features classifier," vol. 2, pp. 194–197, 01 2008.

[12] T. on APPLIED ELECTRONICS, "Communications on applied electronics." https://caeaccess.org/ .

[13] D. R.S, S. Bharathwaj, and M. Aadhil, "Sign language recognition using deep learning and computer vision," Journal of Advanced Research in Dynamical and Control Systems, vol. 12, pp. 964–968, 05

2020.

[14] R. M. Ruben, P. Kambli, and S. K. K. R, "Sign language dynamic gesture recognition system leveraging deep learning and computer vision," in 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), pp. 1–9, 2022.

[15] T. D. Science, "Towards data science." https://towardsdatascience.com/ .

[16] P. Asia Pacific Institute Of Information Technology [APIIT], "Apiit." https://collegedunia.com/college/12894-asia-pacific-institute-ofinformation-technology-apiit-panipat

[17] A. Pathak, A. Kumar—priyam—priyanshu, G. Chugh, and E. Ijmtst, "Real time sign language detection," International Journal for Modern Trends in Science and Technology, vol. 8, pp. 32–37, 01 2022.

[18] S. V. University, "Somaiya vidyavihar university." https://www.somaiya.edu/en .

[19] S. Sen, S. Narang, and P. Gouthaman, "Real-time sign language recognition system," in 2023 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), pp. 1–6, 2023.

[20] A. Alsaffar, H. Tao, and M. Talab, "Review of deep convolution neural network in image classification," pp. 26–31, 10 2017.

[21] U. of Bath, "University of bath." https://www.bath.ac.uk/ .

[22] S. R. Kodandaram, N. Kumar, and S. Gl, "Sign language recognition," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, pp. 994–1009, 08 2021.

[23] S. Chavan, X. Yu, and J. Saniie, "Convolutional neural network hand gesture recognition for american sign language," in 2021 IEEE International Conference on Electro Information Technology (EIT), pp. 188– 192, 2021.

[24] T. J. of Computer and M. E. (TURCOMAT), "(turcomat)." https://turcomat.org/index.php/turkbilmat , 2009.

[25] M. U. London, "Middlesex university londo." https://www.mdx.ac.uk/study-with-us/international/ .

[26] I. I. U. Malaysia, "International islamic university malaysia." https://www.iium.edu.my/v2/ .

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4510– 4520, 2018.

[28] Document.PUB, "Digitale Bibliothek, Erkenntnisse finden und austauschen." https://dokumen.pub/ .

[29] U. of Nottingham, "University of nottingham." https://www.nottingham.ac.uk/

[30] T. S. Press, "Tsp." https://www.techscience.com/ , jan 1997.

[31] Kaggle, "Kaggle." https://www.kaggle.com/

[32] U. of Wolverhampton, "University of wolverhampton." https://www.wlv.ac.uk/

[33] A. T. University, "Atu." https://www.lyit.ie