Implementation of CNN-based Word Recognition System

Bittu Kumar
Dept. of ECE
MLR Institute of Technology
Hyderabad, India
Bittu.mlrit@gmail.com

Satya Prakash Dash
Dept. of ECE
MLR Institute of Technology
Hyderabad, India
Satyaprakashdash879@gmail.com

Sai Kiran Bagli
Dept. of ECE
MLR Institute of Technology
Hyderabad, India
kiransai2589@gmail.com

Shashanth Bhaidishetty
Dept. of ECE
MLR Institute of Technology
Hyderabad, India
Bhaidishettyshashanth@gmail.com

Pradeep Oalati
Dept. of ECE
MLR Institute of Technology
Hyderabad, India
19R21A0497@mlrinstitutions.ac.in

Abstract- For those who are deaf or hard of hearing, sign language serves as a crucial communication tool. Unfortunately, not everyone is conversant in sign language, making it difficult for deaf people to interact with the larger hearing population. We provide a sign language to text conversion system that instantly translates text from American Sign Language (ASL) to English in order to solve this problem. The system recognises hand movements and converts them into text using a deep learning method that blends convolutional and recurrent neural networks. The method has the potential to enhance deaf people's accessibility and communication while achieving high levels of accuracy. By deriving hand movement parts, hand movement modelling is also conducted out here using target sign language independent data. This may be developed using TDNN, CNN, and DNN. Although it has made progress, sign language recognition technology is still in its relative infancy. We analyse the performance of proposed methods through the investigation on own generated/standard data set.

Keywords— American Sign Language (ASL), Convolutional Neural Network (CNN), Deep Neural Network (DNN), Time Delay Neural Network (TDNN).

I. INTRODUCTION

Everyone needs to communicate, but those who are deaf or have hearing impairments find it difficult to interact with the larger hearing community. Deaf people require sign language as a vital communication tool, and several nations recognise it as an official language. Nonetheless, there may be communication hurdles because not everyone is conversant in sign language. We provide a sign language to text conversion system that can instantly translate American Sign Language (ASL) to English text in order to solve this problem. Hand gesture detection for HCI is a significant field to investigate for computer vision and machine learning researchers.

Making systems that recognise specific signs and use them to communicate information or control things is one of its main objectives. However, although gestures are the dynamic movement of the hand, hand postures are the static structure of the hand, and gestures must be defined in both the spatial and temporal domains. Vision-based addresses and data glove methods are the two basic techniques for recognising hand motions. The main goal of this effort is to create a vision-based methods and data glove methods. The primary objective of this endeavour is to develop a vision-based system capable of real-time sign language recognition. A system based on vision is preferred because it offers a more straightforward and natural means of communication between a human and a machine. The Figure 1 shows standard hand gestures for alphabets and numbers.

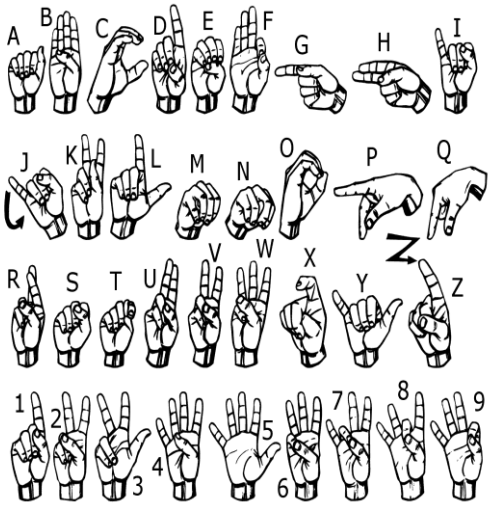

Fig.1. Sign Language chart for alphabets and numbers

A field called "digital image processing" uses computer algorithms to manipulate digital photographs in order to extract information about the scene being captured. Images are produced by numerous physical tools including cameras, x-ray machines, and microscopes and utilised for a variety of reasons including entertainment, medical care, business, industrial, military, civil, security, and research. Digital image processing includes two essential steps: picture enhancement and analysis. Image enhancement aims to highlight important image details while minimising unimportant ones, whilst analysis aims to glean information about the scene.

Digital image processing involves transforming images into arrays of pixels that reflect a physical amount stored in digital memory, and has benefits over analogue image processing, including better speed, cost-effectiveness, and versatility, which are then processed by computer or other digital hardware.

A two-dimensional function of spatial coordinates with an amplitude known as intensity can be used to depict a picture. Whereas digital image processing uses finite sets of data values called pixels, analogue image processing manipulates images by altering electrical signals. Any quantifiable quantity that carries information through time or space is referred to as a signal. In order to create a digital image, digital cameras use sunlight to generate a continuous voltage signal that is then sampled and quantized into a two-dimensional array or matrix of values.

As one of the most crucial instruments for human communication in daily life, research on human-machine interaction has become more and more relevant as image and video processing methods continue to progress.

The remainder of the paper is organised as follows. A brief literature review of the techniques used in sign language recognition is included in section II. Section III discusses technique. Results and debates are discussed in section IV. Section V concludes by outlining the research's future direction and scope.

II. LITERATURE SURVEY

In [1], authors explain a system that can recognize sign words based on hand gestures using a convolutional neural network. The system involves various stages such as pre-processing, feature extraction, and classification to recognize the sign words and detect hand movements. Pre-processing involves converting the hand gesture images into YCbCr, selecting grayscale images, binarizing, erosion, and gap filling. Features are extracted from the images and provided to the classifier for classification using a deep learning technique called CNN. The gestures are classified using a SoftMax classifier, and the system is implemented in real-time using a webcam. The system has an average acceptance rate of 96.96% and outperforms the state-of-the-art systems. The goal of future research is to develop an algorithm that can interpret innovative sentences and sign words based on hand gestures for the sign language recognition system.

In [2], researchers proposed two models for recognizing 3D handwritten characters and gestures using smartphones. Smartphones have sensors such as an accelerometer, gyroscope, and gravity sensor that can track the phone's movement in three-dimensional space. The systems are made more expressive and harder to imitate by the model's using data from the accelerometer and gyroscope sensors. The accelerometer's 3-axis readings are used to calculate the hand movements' speed. The gyroscope data is used to derive a quaternion rotation matrix, which reduces the inclination offset and tilting. Each character or gesture in a series is determined using an automatic segmentation method. Both models have the ability of both user-dependent and user-independent recognition of 3D characters and motions. According to the results, Model I have a recognition rate of 90% to 94% with little instruction, whereas Model II has an 88% recognition rate. The first model requires more computation for training and evaluation, while the second model requires less computation.

In [3], authors examined how to use hand movement subunits derived from HMMs to represent hand movement data in a language-independent manner. The research demonstrated a distinction in performance between modelling hand movement data in a language-independent and language-dependent way. Yet, when hand shape information is combined with this gap, competitive systems can be created. Sharing multiple sign language resources can facilitate the development of sign language processing systems, which is a positive outcome from these findings. The research results serve as the foundation for future work to the processing of sign language must take account of limitations in resources by creating systems with fewer signers and examples. Additionally, the study was investigated whether this multilingual approach can be used to evaluate sign language.

In [4]. To handle all the different sorts of data generated via the two devices, various feature sets have been used. While the Kinect presents the complete depth map, the Leap Motion only delivers a higher-level, more constrained data description. The proposed set of elements and order calculation allows for a good overall accuracy, despite the possibility of certain fingers not being recognized. A very high level of accuracy may be achieved by integrating the two devices, and the depth map of the Kinect, that offers an additional description, enables the acquisition of more attributes that are missing in the Leap Motion output.

The results of the tests revealed that performance is significantly enhanced with allocating each finger to a specific angular zone. The two sensors were to be used to detect dynamic motions, and future research would focus on calculating new features based on the combination of the two devices' 3D positions.

In [5], The authors demonstrate the use of the Leap Motion and depth sensor-based gesture recognition pipelines. Leap Motion gives a small amount of information, whereas depth cameras describe the hand shape in more detail. The proposed calibration method enables the combined use of data form both of the devices. Leap Motion data has been used to speed up computation and increase the precision of the depth traits, and different feature sets have been given for the two sensors. The proposed set of features and classification algorithms, in accordance with experimental results, show good accuracy, especially for depth data. Future research will focus on the recognition of dynamic gestures as well as enhanced methods for translating and tracking fingertip motion.

III. METHODOLOGY

The system described in the input is a vision-based approach for sign language recognition. It allows for interaction without using any artificial devices, as all signs are represented with bare hands. The system includes dataset generation, gesture classification, a CNN model, activation

function, pooling layer, finger spelling sentence formation implementation, and training and testing.

A. *Dataset Generation*:

As there were no pre-made datasets in the form of raw pictures, the first stage is dataset manufacture, which entails establishing a data set. Each ASL symbol is photographed for use in training and testing using OpenCV programming. A blue bordered rectangle identifies the region of interest (ROI) in each frame. The ROI is retrieved from the entire image, converted to grayscale, and then subjected to a Gaussian blur filter to extract various features. Figure 2 shows the new image's ROI. Figure 3 shows the image that was taken from the ROI grayscale image. Figure 4 displays many attributes for a picture in a Gaussian blur filter.

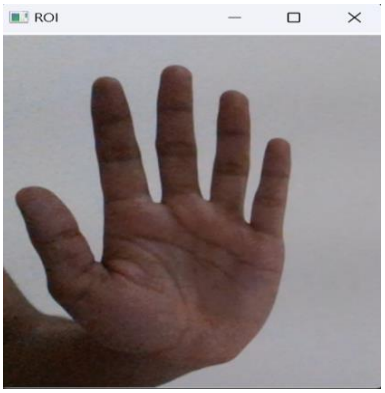

Fig. 2 ROI Image

Fig. 3. Grey scale image

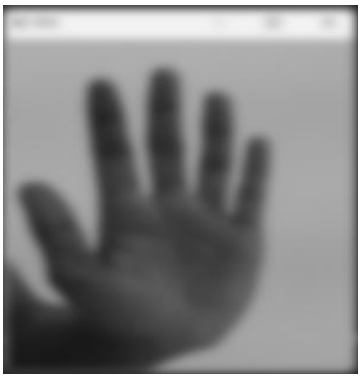

Fig. 4. Gaussian blur image

B. *Gesture Classification*:

The next step is gesture classification, which analyses the user's final symbol using two levels of algorithms. Algorithm Layer 1 involves applying a Gaussian blur filter and threshold to the OpenCV frame and sending the processed image for prediction, to the CNN model. The letter is printed and used to create the word if it occurs in more than 50 frames. blank symbol is used to consider the space between the words. Algorithm Layer 2 involves finding various sets of symbols that exhibit similar results and using classifiers designed just for those sets to differentiate between those sets.

C. *CNN Model*:

The CNN model is a crucial component of the system. The input image has a 128 × 128-pixel resolution in the first layer of convolution, with the first convolutional layer processing it first using 32 filter weights. The images are down sampled to 63 by 63 pixels using maximum pooling of 2x2 in the initial layer of pooling. The second convolutional layer processes the output of the first pooling layer by using 32 filter weights to create an image which measures 60 × 60 pixels. After being out sampled once more with a maximum pool size of 2x2, the final photos are scaled down to a resolution of 30 x 30.

There are 128 neurons in the top densely connected layer as well as 96 neurons in the bottom densely connected layer. The output of the CNN model has as many neurons as classes we are grouping (letters in sequence plus a clear image). The figure 5 represents matrix convolution of image. First, we have input matrix. From input we taken blue colour layered image patch (Local respective field) and multiplied by Kernel (Filter) to get output.

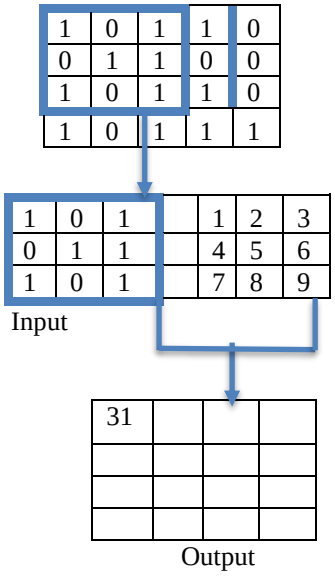

Fig. 5 matrix convolution of image

D. *Activation Function*:

Rectified Linear Unit (RELU) is the activation function that is employed in each layer (convolutional and fully connected neurons), generating Max (x, 0) for each input pixel to provide nonlinearity to the recipe and aid in learning more muddled highlights.

In this case, the negative matrix value in the input signals that the image is blurry. To eliminate the blur, we apply ReLu, in which the negative values are changed to zeroes while the positive values stay the same. Finally, there is no fuzz in the image. The Table 1 and 2 represent the Output of matrix

convolution of Input image and Output when we apply ReLu respectively.

Table 1: Output of matrix convolution of Input image		
-249	-91	-37
12	-133	137
27	61	-153

Table 2: Output when we apply ReLu		
0	0	0
12	0	137
27	61	0

E. *Pooling Layer*:

A maximum join with pool size (2, 2) is applied to the input image using the RELU activation function in the pooling layer to reduce the number of parameters and minimize overfitting and the cost of computation.

Here, we have a 4 by 4 matrix that may be divided into four 2 by 2 matrixes. We obtain the maximum value of each 2x2 matrix by performing pooling. Finally, given a 4 x 4 matrix input, we have a 2 x 2 matrix output. The figure 6 represents the whole process.

2	2	7	3
9	4	6	1
8	5	2	4
3	1	2	6

Max Pool

Filter - (2 x 2)

Stride - (2, 2)

9	7
8	6

Fig. 6 Pooling

F. *Finger spelling sentence formation Implementation*:

Generating fingerprint statements involves printing a letter and adding it to the current string if its count exceeds a certain value and no other letter is close to it by a threshold. If an incorrect letter is expected, the current word reference containing the number of hits in the current image is removed to avoid anticipating a non-prime letter. If the number of detected blanks (array backgrounds) exceeds a predetermined value, no blanks are detected in the current buffer. Otherwise, it prints a space to indicate when the word has ended, and the current word is appended to the sentence below.

G. *Training and Testing*:

Training and testing involve pre-processing the input images by grey scaling them and applying Gaussian blurring reduces unneeded noise. The prediction layer then evaluates the likelihood that an image will belong to any of the classes after the images have been downsized to 128x128 and given to the CNN model. Since each output is normalised within 0 and 1, their aggregate is 1.

IV. . RESULTS AND DISCUSSIONS

A. *Collection*:

Raw image is collected through web cam of ROI detecting the part where only image of hand is required. Here hand gesture for A is collected.

Fig. 7 Data collection alphabet A

Then the image is functioned to various methods like grey scale image, blur image. This section is used to pinpoint the locations of the temporal restricts of this information after that segmentation. A webcam-captured hand image has been separated to get the hand region. The shapes, actions, and materials used in the various hand gestures vary & also, binary information of image.

B. *Conversion*:

In this step the .png image is changed into a.npy file (matrix form), which includes a variety of properties to function under many circumstances. Here the data frames are collected.

Fig. 8 Conversion of frames alphabet A

C. Training and Testing:

Since they operate well with image data, CNNs, or artificial neural networks made up of convolutional layers, are the approach of choice for challenges involving image categorization. The categories training photographs were

saved in the database. The image with the highest accuracy was generated as text after comparisons between the tested image and training dataset models. In output image orange arrow represents alphabet and green arrow represents accuracy of image.

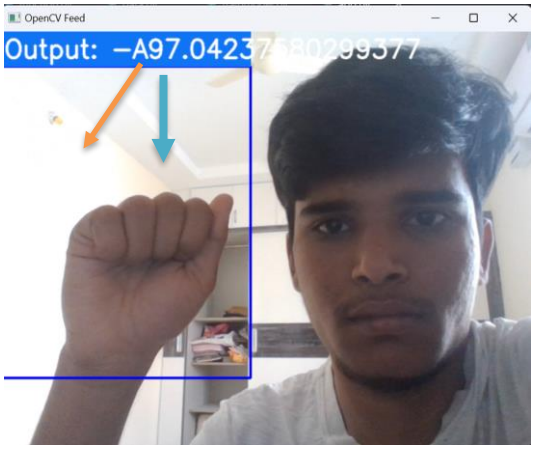

Fig. 9 Output of A

The laptop's embedded webcam gets utilised in our report, so it is a huge plus. Following is the confusion matrix of our findings.

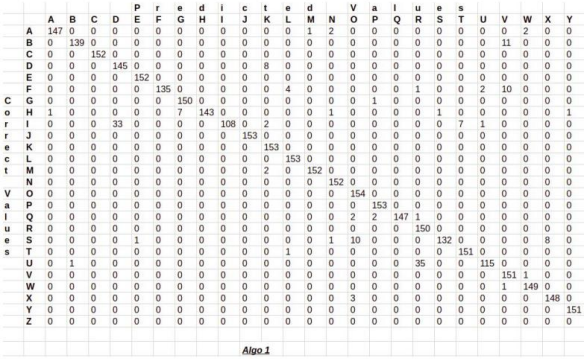

Fig. 10 Confusion Matrix

V. CONCLUSION

In this paper, we introduced a real-time system for transcribing text from American Sign Language (ASL) to English for sign language. The system recognises hand movements and converts them into text using a deep learning method that blends convolutional and recurrent neural networks. In both hand gesture recognition and phrase reconstruction, the system produced highly accurate results. The method, which may be expanded to other sign languages, holds the possibility to improve deaf people's accessibility and conversation. The classification of gestures and gesture recognition in this method has been discussed using photographs. The framework that has been put in place works effectively for altering both lighting and gesture orientation. By using a specific dataset, we were able to achieve a high accuracy of 97.0% in recognizing symbols.

This technique can almost always correctly identify the symbols as long as they are discernible, there is no background noise to interfere with them, and there is enough light.

VI. . REFERENCES

[1]. H. Cooper, B. Holt, and R. Bowden, “Sign language recognition,” in Visual Analysis of Humans, 2011.

[2]. F.S. Chen, C.M. Fu and C.L. Huang. "Hand gesture recognition using a real-time tracking method and hidden Markov models." Image and vision computing, vol. 21, no. 8, pp. 745-758, Aug. 2003.

[3]. G. Marin, F. Dominio, and P. Zanuttigh. "Hand gesture recognition with jointly calibrated leap motion and depth sensor." Multimedia Tools and Applications, vol. 75, no. 22, pp. 14991-15015, Nov. 2016.

[4]. P. Kumar, H. Gauba, P.P. Roy, and D.P. Dogra. "Coupled HMM-based multi-sensor data fusion for sign language recognition." Pattern Recognition Letters, vol. 86, pp. 1-8, Jan. 2017.

[5]. D. Lifeng, R. Jun, M. Qiushi, W. Lei, “The gesture identification based on invariant moments and SVM[J].” Microcomputer and Its Applications, vol. 31, no. 6, pp. 32-35, 2012.

[6]. M.A. Rahim, J. Shin, and M.R. Islam, “Human-Machine Interaction based on Hand Gesture Recognition using Skeleton Information of Kinect Sensor.” In Proceedings of the 3rd International Conference on Applications in Information Technology, ACM, pp. 75-79, Nov. 2018.

[7]. T. Yamashita, T. Watashe, “Hand posture recognition based on bottom-up structured deep convolutional neural network with curriculum learning.” IEEE international conference on image processing (ICIP), pp 853–857, Oct. 2014.

[8]. Y. Liao, P. Xiong, W. Min, W. Min, and J. Lu, “Dynamic Sign Language Recognition Based on Video Sequence with BLSTM-3D Residual Networks.” IEEE Access, 2019.

[9]. [Oscar Koller, Necati Camgoz, Hermann Ney, and Richard Bowden, “Weakly supervised learning with multi-stream cnn lstm-hmms to discover sequential parallelism in sign language videos,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 04 2019.

[10]. S. Tornay, M. Razavi, N. C. Camgoz, R. Bowden, and M. Magimai.-Doss, “HMM-based approaches to model multichannel information in sign language inspired from articulatory features-based speech processing,” in Proc. of ICASSP,2019.

[11]. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition by Aurélien Géron

[12]. Practical Python and OpenCV+ Case Studies- An Introductory, Example Driven Guide to Image Processing and Computer Vision 4thEdition

[13]. T. Yang, Y. Xu, and “A. , Hidden Markov Model for Gesture Recognition”, CMU-RI-TR-94 10, Robotics Institute, Carnegie Mellon Univ.,Pittsburgh,PA, May 1994.

[14]. Pujan Ziaie, Thomas M¨uller , Mary Ellen Foster , and Alois Knoll“A Na¨ıve Bayes Munich,Dept. of Informatics VI, Robotics and Embedded Systems,Boltzmannstr. 3, DE-85748 Garching, Germany.

[15]. Lars Bretzner, Ivan Laptev, and Tony Lindeberg. 2002. Hand gesture recognitionusing multiscale color features, hieracrchichal models and particle in Ąltering. In proceedings the fifth IEEE International Conference on Automatic Face and Gesture Recognition, 423-428.

[16]. Nasser H. Dardas, and Nicolas D. Georganas. 2011. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. IEEE Transactions on Instrumentation and measurement, 60,11 (November 2011), 3592-3607.

[17]. Sebastian Nowozin, Pushmeet Kohli, and J. D. J. Shotton. 2017. Gesture detection and recognition. U.S. Patent 9,619,035, issued April 11.

[18]. Trygve Thomessen, Mihoko Niitsuma, Keita Suzuki, Takashi Hatano, and Hideki Hashimoto. 2015. Towards virtual presence based on multimodal man-machine communication: A remote operation support system for industrial robots. IFACPapersOnLine, 48,19 (January 2015), 172-177.

[19]. C. Gulcehre and Y. Bengio, “Knowledge matters: Importance of prior information for optimization”, arXivpreprint arXiv:1301.4083, 2013.
[20]. Mohammed Waleed Kalous, Machine recognition of Auslan signs using PowerGloves: Towards large-lexicon recognition of sign language.