American Sign Language Recognition using Deep learning

A P Purushotham[1], Jayanth A[2], Kiran Kumar S[3], Akilesh N S[4]
[1]*Student, Computer Science and Engineering Department, New Horizon College of Engineering, Bangalore*
[2]*Student, Computer Science and Engineering Department, CMRIT, Bangalore*
[3,4]*Student, Electronics and Communication Engineering Department, R V College of Engineering, Bangalore*

***Abstract*— ASL (American Sign Language) is a difficult language. It is determined by the unique gesture stander of marks on the hands. Hands convey these marks, with face expression and body position assisting. ASL is the primary language of deaf and hard-of-hearing persons in North America and other parts of the world. The use of Deep Learning to recognise static ASL gestures is proposed in this paper. The contribution consists of a problem-solving approach. Convolution Neural Network (CNN) and Deep learning has been used to classify the 24 alphabetic static letters of ASL. The classification accuracy is 99.68 percent, with a loss function error of 0.32. In compared to other comparable studies like CNN, SVM, and ANN for training, the training is quick and produces excellent results.**

***Index Terms:* ASL, American Sign Language, CNN, Deep Learning.**

I.INTRODUCTION

The deaf community communicates mostly through American Sign Language (ASL). However, there are few ASL speakers, which limits the number of people with whom they can easily communicate. Textual communication is inconvenient, impersonal, and even unfeasible in an emergency. To overcome this obstacle and enable dynamic communication, we describe an ASL recognition system that translates a video of a user's ASL signs into text.

1. Get a video of the user signing (input).
2. Assigning a letter to each frame of the movie
3.Using categorization scores to reconstruct and present the most likely term (output).

While Neural Networks have previously been used to recognize ASL letters with over 90% accuracy, many of them require a 3-D capture element, such as motion-tracking gloves or a Microsoft Kinect, and only one allows for real-time classifications. The scalability and viability of these systems are limited due to the additional requirements. A pipeline in our system feeds video of a user making a hand gesture into a web application. Then, using a CNN, we extract individual frames from the video and generate letter probabilities for each of them (letters a through y, excluding j and z since they require movement). We organize the frames using a number of heuristics depending on the character index that each frame is thought to match.

II. PROBLEM STATEMENT

Due to a multitude of factors, including environmental problems (e.g., lighting sensitivity, background, and camera angle), this topic poses a substantial difficulty in computer vision. Obstruction (e.g., some or all fingers, or an entire hand can be out of the field of view), Co-articulation (where a sign is impacted by the preceding or subsequent sign), Abbreviations and Acronyms are all examples of sign boundary detection.

III. RELATED WORK

After the static ASL gestures have been recognized, the classification procedure must be used to classify the static ASL characters. The edge approach is used to detect the hand gesture boundary. The static ASL is classified using the Localized Contour Sequence (LCS) approach. The accuracy of the classification is 97.4 percent [1]. For ASL, the shape hand algorithm is divided into 24 static letters. The shape approach

determines how gestures are recognized. Hardware design with the glove sensor utilizing a neural network is introduced, with an accuracy of 180 points of landmark equaling 79.9% [2]. Deep neural network learning can be characterized as a machine learning method that comprises neural networks with more than one hidden layer. It has a variety of uses, including facial recognition, speech processing, and language processing. Convolution Neural Networks (CNNs) and layered denoising of auto-encoders are used in deep learning to recognize the 24 static ASL letters. For data, the accuracy percentages are 91.33 percent and 92.83 percent [3]. Edge oriented histogram (EOH) and multi-class SVM algorithms were employed in study [7]. The average system precision attained a success rate of 93.75 percent. With 64 characteristics, precision is employed. To extract the digital image and reduce the image's noise, alphabet sign language recognition for Peru is offered. In addition, the dissimilarity must be processed under various lighting conditions. CNN is a trademarked hand gesture. . The initial score for CNN accuracy was 95.37 percent, while the second result was 96.20 percent [4]. Support Vector Machine (SVM) and Artificial Neural Network (ANN) were utilized for training in study [5]. SVM and ANN both identify the static ASL alphabet. Per alpha, the researcher gathered a 100-deep histogram of oriented gradient features. The level of precision is 94.7 percent. The static ASL alphabet is recognized as Edge Oriented Histogram (EOH) in [6]. Within 0.5 seconds, they had an 88.26% recognition rate.

IV. METHODOLOGY

4.1 SYSTEM ARCHITECTURE

A Convolutional Neural Network (CNN) is a Deep Learning system that can take an input image, give relevance (learnable weights and biases) to various objects in the image, and distinguish between them. In comparison to other classification algorithms, a CNN requires substantially less pre-processing. While filters are hand-engineered in basic approaches, CNN can learn these filters/characteristics with adequate training. This study created a real-time ASL fingerspelling recognition using CNNs algorithm with real-colored pictures. There was a total of 26 alphabets, including J and Z, as well as two space and delete classes. The system was divided into three parts, the first of which was data collection. Because the Hand-Gesture Recognition algorithms studied in this study required a large dataset for training, it was decided to construct additional datasets with a broader diversity of features, such as different lightings, skin tones, backgrounds, and scenarios. This technique allows the majority of hearing and deaf cultures to converse more readily. It's a camera-based computer input system. The writing system reflected the computer's link with the user, whereas the second phase involved CNN multi-class recognition. Figure 3 depicts the architecture of the proposed system.

Figure 1: Sample Dataset

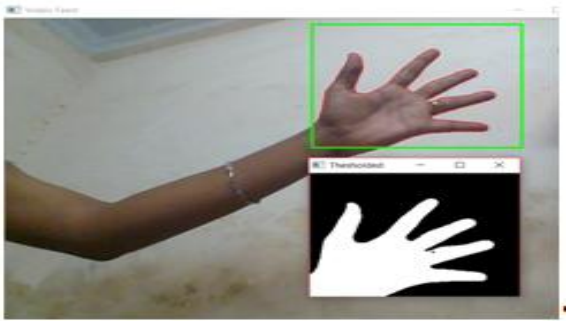

Figure 2: Hand Gesture Sample

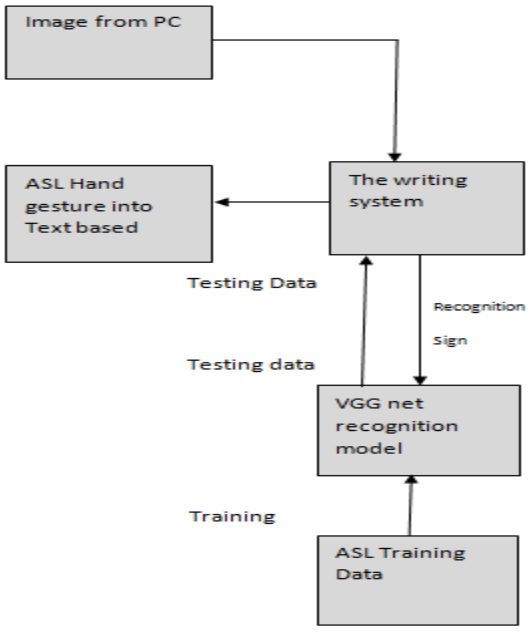

Figure 3: System Architecture

4.2 MULTI CLASS RECOGNITION WIH CNN

The majority of the previous works used feature-based techniques, but not this one. A multi-class recognition system was built using convolutional neural networks (CNNs) and a deep learning data structure. Each ASL sign was assigned to a distinct category. The output of the classifier would be classified into one of 28 categories ranging from 0 to 27. The CNN architecture used was the VGG Net, which is a very deep convolutional network design for high-scale image recognition. The proposed re-training methodology started by randomly initializing network weights, which were subsequently changed to perform the tasks with less errors. The weights of the network were saved and loaded as the starting weights for subsequent testing, a procedure known as fine-tuning.VGG blocks from the TFlearn (TFlearn Development Team GitHub, 2017) original site were used in this project.

4.3 UNITS TRAINING FOR THE MULTI CLASS RECOGNITION SYSTEM

The CNNs job is to compress the images into a format that is easier to process while preserving important attributes for a decent prediction. This is critical for designing an architecture that is capable of learning features while also being scalable to large datasets.In the below demonstration, Figure 4, the green section in the following sample mimics our 5x5x1 input image, I. The Kernel/Filter, K, is the element that performs the convolution operation in the initial half of a Convolutional Layer. It is illustrated in yellow. K has been chosen as a 3x3x1 matrix.Because Stride Length = 1 (Non-Strided), the Kernel shifts 9 times, each time conducting a matrix multiplication operation between K and the image part P over which the kernel is hovering. In American Sign Language, the photographs were utilized to produce 28 classes of static fingerspelling (ASL). All of the photographs were scaled to 224 by 224 pixels and then normalized to feed the VGG Net. The paths of the photographs, as well as their labels, were saved to a text file. The images were converted into NumPy array form (No. of Images, 32, 32, 3) and fed to the system using TFlearn data. The model was trained with a total of 61.614 training datasets, with at least 2200 for each class. The validation datasets were created using about 0.30 of the training datasets, resulting in a total of 43,120 training datasets and 18,480 validation datasets. Before isolating and training the validation dataset, the shuffle=True specifies that the training datasets will be mixed each time. In order to test the mode for roughly 100 images in each class, a total of 2816 photographs were recorded. It took more than seven days to train the model.

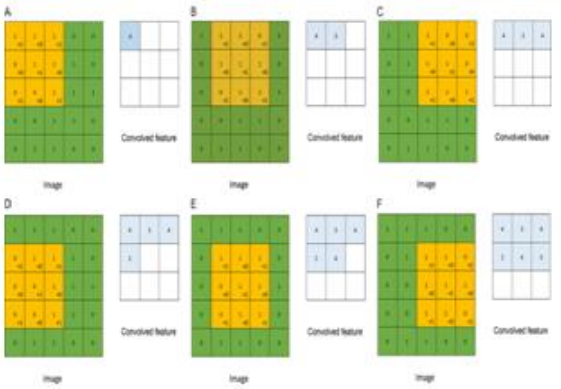

Figure4:Convoluting a 5x5x1 image with a 3x3x1 kernel to get a 3x3x1 convolved feature

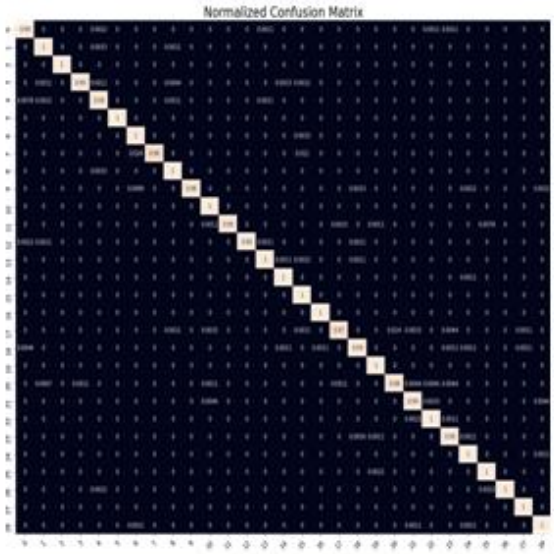

Figure5: Result – Normalized Confusion Matrix

V. RESULTS

An ASL recognition with a CNN algorithm was built using real coloring images from a PC camera. This work converts deaf signals into text statements, which aids in the development of a writing system that can be used as an input method for any computer camera. This system produced outstanding results

using a deep learning technique. There were a lot of different outcomes from the experiment. A multi-class recognition system was created using VGG Net. Using CNNs as the recognition mechanism, each ASL sign was represented as a distinct property. The classifier's output would be one of 28 categories ranging from 0 to 27.The system was trained with only 10 labels and around 20300 training data for each class to create less than 1995 images.The implemented CNN models have been our mentionable factors. We observe a validation accuracy of 99.70% (0.30% error rate) for our best model. Adding further, ReLUs' (Rectified Linear Unit) prove to be very effective with an improvement of 23.8% with respect to tanh units. The accuracy on the main test set is 99.68% and we observe a 0.32% false positive rate, caused by the noise movements. Here, the test result is higher than that of the validation result and the reason behind this is that the validation set doesn't contain users and backgrounds in the training set.

VI. SCOPE IN FUTURE

This approach can be applied to other sign languages, such as Indian Sign Language, however it is currently only applicable to American Sign Language.

- The model can be trained further using a dataset so that it can automatically segment the gesture from the collected frame by eliminating the backdrop.
- Tuning and enhancing the model to recognize common phrases and idioms.
- Furthermore, training the neural network model to recognize symbols in a well-organized manner requires two hands.
- Incorporate active hand motions in addition to the static finger spelling that is currently in use.
- Integration of the improved model with existing AI systems, such as Amazon Alexa, to improve visual recognition.

VII.CONCLUSION

The goal of this study is to demonstrate how convolutional neural networks can be used to reliably identify different signs in a sign language without include individuals or their environment in the training set. This generic capability of CNNs in spatiotemporal data can help advance the field of automatic sign language recognition research. When we consider all of the conceivable combinations of motions that a system like this must interpret and translate, sign language recognition is a difficult challenge. That stated, the most efficient method to handle this problem is to break it down into smaller chunks, with the system provided here serving as a possible solution to one of them. Although the system didn't perform particularly well, it did show that a first-person sign language translation system could be developed using only cameras and convolutional neural networks. The model was discovered to have a habit of mixing up various signs, such as U and W. However, after some thought, flawless performance may not be required because the use of an orthography corrector or a word predictor may increase translation accuracy. The next step is to evaluate the response and come up with methods to improve it. More high-quality data, more convolutional neural network topologies, and a redesigned vision system could all help.

REFERENCE

[1] A. Julka and S. Bhargava, "A static hand gesture recognition based on local contour sequence," International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 7, 2013.

[2] A.K. Gautam and A. Kaushik, "American sign language recognition system using image processing method," International Journal on Computer Science and Engineering (IJCSE), Vol. 9 No.07, 2017.

[3] O.K. Oyedotun and A. Khashman, "Deep learning in vision-based static hand gesture recognition," Springer, Neural Computing and Applications, 2016.

[4] J.L. Flores C.E. Gladys Cutipa, R.L. Enciso, "Application of convolutional neural networks for static hand gestures recognition under different invariant features," IEEE, 2017

[5] J. Bamwend and M. Özerdem, "Recognition of static hand gesture with using ANN and SVM"," Dicle University Journal of Engineering, 2019

[6] J. Pansare and M. Ingle, "Vision-based approach for American sign language recognition using

edge orientation histogram," International Conference on Image, Vision and Computing, 2016

[7] S. Nagarajan and T. Subashini," Static Hand Gesture Recognition for Sign Language Alphabets using Edge Oriented Histogram and Multi Class SVM", International Journal of Computer Applications, Vol. 82, 2013.