International Journal of Advance Research in Science and Engineering
Volume No. 12, Issue No. 05, May 2023
www.ijarse.com
IJARSE
ISSN 2319 - 8354

Sign Language Detection and Translation Using CNN

[1]Priyanshu Singh, [2]Prashant Mishra, [3]Mayank Gusain

[1]Amity School of Engineering & Technology, priyanshusingh2208@gmail.com

[2]Amity School of Engineering & Technology, pmishra2001@gmail.com

[3]Amity School of Engineering & Technology,mayankgusain212@gmail.com

Abstract

Sign language is a complex and dynamic visual language used by individuals with hearing impairments to communicate. People who have hearing loss use sign language as a mode of communication. Interpretation of a person's gestures, movements, and facial expressions is necessary for sign language, which is a difficult task. To help challenged people communicate, sign language identification systems have recently been created using machine learning and computer vision techniques. Convolutional neural networks (CNNs) are used in this field and SignNet architecture is implemented to recognize and categorize various sign motions. The model has a 91% accuracy rate on a test set of videos with diverse background settings after being trained on a large dataset of sign language videos. The results demonstrate that SignNet beats other cutting-edge models in terms of accuracy and computational efficiency. People who are deaf can receive real-time support from the suggested model via wearable technology or cell phones, allowing them to converse more successfully and self-assuredly.

Keywords- *Sign Language Recognition, CNN, ASL, Automation, Recursive Neural Networks (RNN)*

I. INTRODUCTION

Every country has a pre-set sign language. Individuals who are hard of hearing or are deaf use sign language as their primary mode of communication. Its vocabulary, syntax, and semantics are distinct, contrasting with spoken language. Sign language utilizes hand gestures, facial expressions, and body movements to represent words. Parents of deaf children typically keep them indoors and discourage socialization, making it harder for them to lead a regular life. When they start using sign language to communicate and seek help in public without an interpreter, communication becomes a barrier between the deaf and the rest of the world as sign language is their sole means of expressing their needs.

A combination of LSTM and an RNN to recognize sign language gestures is shown in [1]. RNN was used to extract spatial features from the image frames, and the LSTM and RNN were used to capture the temporal dependencies between the frames. The model achieved an accuracy of 92.12% on the American Sign Language (ASL) dataset

Researchers proposed a model that combined a CNN and an LSTM for sign language translation in [2] The image frames' characteristics were extracted using CNN, while the temporal dependencies between the frames were recorded using LSTM. On the Korean Sign Language (KSL) dataset, the model's accuracy was 96.09%.

In paper [3], the model is explained to recognize sign language motions in real-time by combining a CNN with advanced LSTM. The image frames' characteristics were extracted using CNN, while the temporal dependencies between the frames were recorded using LSTM. On the ASL dataset, the model has an accuracy of 94.1%. Overall,

this research demonstrates that integrating CNNs with LSTM or RNNs can provide sign language recognition models that are more precise. The difficult task of sign language recognition has been tackled by neural networks and other techniques in computer vision. The capacity of convolutional neural networks (CNNs) to extract characteristics from images and understand complicated patterns has led to promising breakthroughs in this field.

Several studies have also looked into complex CNN features used for sign language detection in recent years. One such study put out in [4], showcases a model that classified hand motions using a recurrent neural network (RNN) and a CNN to extract information from photos of hand movements. The American Sign Language (ASL) dataset showed that the model had an accuracy of over 90%.

A two-stage approach for sign language detection using a CNN is also explained in [5]. In the first stage, the CNN was used to detect the hand region and extract features from it. In the second stage, another CNN was used to classify the hand gestures. The model achieved an accuracy of over 94% on the German Sign Language (DGS) dataset.

The major contribution of the work is a proposed CNN-based model architecture 'Gesture Recognition and Translation', for correctly classifying the sign language and translating it into its suitable American English alphabet. The dataset for the proposed model was created from scratch so that cleaning and prep-processing of images are much easier and more efficient. The proposed model has 3 inner layers of a neural network with different roles and functionalities. The model can classify with greater accuracy when tuned with some specific hyperparameters for faster detection of gestures. One of the major contributions is providing faster detection of images and providing test accuracy of 91.5% for detecting multiple English alphabets. However, the system's accuracy is not at its highest and can be further improved. An advanced level of research implemented in this area could lead to a more robust and accurate model for sign language detection and translation. [6] The paper also reviews various models applied to this dataset and determines which model is the most efficient in the current scenario.

The latter part of the paper is written in the following order: Section 2 describes the literature review conducted before the study, Section 3 describes the dataset, Section 4 implements the methodology and the CNN model used for prediction, and Section 6 contains the research.

II. LITERATURE REVIEW

Several related papers and research studies were considered to gain insight and information. In a research study, RNN was used for predicting accurate sign language [7]. The traditional sign language dataset was used for sign language recognition, using some transformation of data and pre-processing. It was concluded that language can easily be translated from sign to normal understandable. The precision level is higher with larger datasets in an easily readable format []. The detection and translation system needed guidance on how a dataset model defined on recognition works and its internal algorithm [8]. The dataset for American Standard Language sign-language most tells that there exists defined data from external sites and even some papers recommended to create own dataset as the dataset would be according to our model requirements so it would lead to much better accuracy and other features would be precise.

A dataset related to all the standard sign languages was attempted to be created, but it was difficult to create a model related to the mix of languages. The focus was on more of the American standard sign language tradition. The paper included different methods and approaches on how to create the dataset and major library requirements and pre-installed functions which helped in creating data. The major algorithms included in machine learning as well as deep learning, and each of them had different approaches towards model creation and even led to different accuracies according to their analysis on large of research papers to know which model works in which type of dataset and the structural change which depicts in each of the algorithms internally. Multiple algorithms included CNN, RNN, and LSTM. [9]

CNN is defined as color, edges, and gradient orientation, which is common low-level information that is captured by convolution layers. The spatial dimension of the infused feature is reduced by the pooling layer. Additionally, it has the benefit of preserving key properties that are rotationally and positionally inaccurate during the ML training procedure. Thus, the image is converted into a 1-D vector by flattening it. After training, the system is constructed by using the Keras, TensorFlow, and OpenCV libraries, and the model may provide probabilities of prediction of objects in the image [10]. But it had its disadvantages, which included an operation like MaxPool causing a neural network to activate much more slowly. The CNN contains no. of layers, and the computer contains inefficient GPU, the training process includes delayed time. ConvNet to act and be trained, a large amount of dataset is needed.

Another algorithm to improve from CNN drawbacks and learned about sign language recognition using RNN reduces the work of increasing parameters and memorizing each output previously obtained by using each output as input to the layers and converting independent activations into dependent activations using the same weights and biases[11]. Thus, all three layers can be combined into a single recurrent layer so that the hidden levels' weights and biases are identical. To comprehend how the third algorithm would work with our recognition model, as RNN had its drawback like problems with gradient disappearing and explosions, it is exceedingly tough to train an RNN. Further, if tanh or rely on are used as the activation function, it cannot process very long sequences. The LSTM recurrent unit attempts to "remember" all of the prior knowledge that the neural network has considered till now and to remove extra datasets to address the problem regarding Vanishing and also Exploding Gradients in a Deep Recurrent Neural Network [12]. This can be achieved by adding various "gates," or activation function layers, for different purposes. The mathematics of the method is the sole significant distinction between the Back-Propagation algorithms of Long Term Memory Networks and RNN.

III. DATASET

The dataset considered for this research is self-created OpenCV was used along with pointing major points on hand, and the camera captured either the left or right hand with clarity. As can be seen in Fig 1, About 200 images were clicked and saved for each of the 26 alphabets, so a total dataset of about 6000 images, from where the process of detection was started.

International Journal of Advance Research in Science and Engineering
Volume No. 12, Issue No. 05, May 2023
www.ijarse.com
IJARSE
ISSN 2319 - 8354

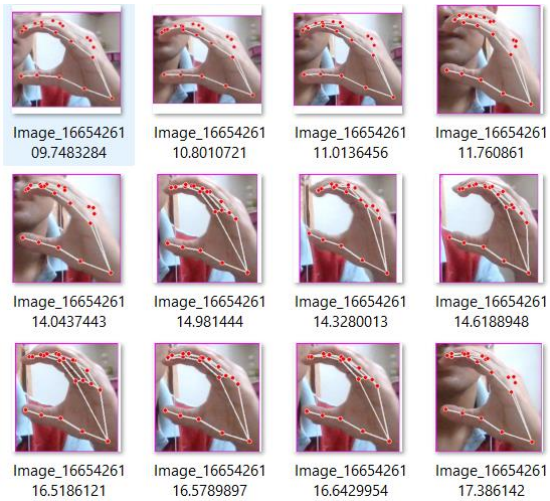


Fig 1. Dataset using Image Capturing

IV. METHODOLOGY

The core of image recognition is related to matching live images with images in data folders and calculating the distance between points on each image. There are 26 folders related to 26 alphabets of English letters consisting of about 120-150 2D images in each folder. As shown in Fig 2, The proposed architecture consists of dividing data into train and test sets, The second step is to do some cleaning and pre-processing on data images so that images are easy to train on the model, different neural network models including CNN, RNN, and LSTM are implemented to get correct output and to predict efficient accuracy.

Later in the translation part comes where the model detects which letter it is and translates it into the correct letter. The following key steps form the foundation of the detection and sign language system:

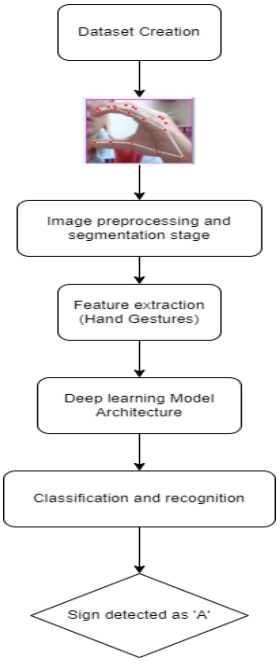


Fig 2: Proposed algorithm flowchart

International Journal of Advance Research in Science and Engineering
Volume No. 12, Issue No. 05, May 2023
www.ijarse.com
IJARSE
ISSN 2319 - 8354

Proposed Deep Learning Model Architecture Convolutional neural network (CNN) is an advanced technology selected as the solution to address this issue. Two models were trained for the top and bottom views. The Convolution Layer transfers the dataset to every position of the image, and its fit is examined.

TABLE II. RELU FUNCTION FOR ACTIVATION FUNCTIONS

X	f(x) = x	F(x)
-3	f(-3) = 0	0
-5	f(-5) = 0	0
3	F(3) = 3	3
5	F(5) = 5	5

Pooling Layer: Stacking the layers as input was processed through three layers—Convolution, ReLU, and Pooling—to create a 3 X 3 matrix from a 6 X 6 matrix to obtain the time frame in a single image. The main adjustment was to reduce the classes classified originally from 90 to 26, which corresponds to the number of signs, with which the model will be trained upon. The model was trained on approximately 10000 steps.

V. RESULT AND ANALYSIS

The research was based on translating sign language using neural network architecture and the CNN model. The webpage works on a system where a page allows starting the translation process and connecting with the system. The camera then opens at the next step, and the user shows the sign through different hand postures. If the hand gesture is not clear, then the model might not detect the sign or detect it incorrectly.

Fig. 3, shows that the trained model creates connecting dots on the hand gesture for the deep learning model to translate into an appropriate language, If the camera is not able to see the hand gesture properly, then a "try again" option will occur to capture the image again. After all the pre-processing, the gesture will be saved again in the database so that it can be trained for the model and increase its accuracy.

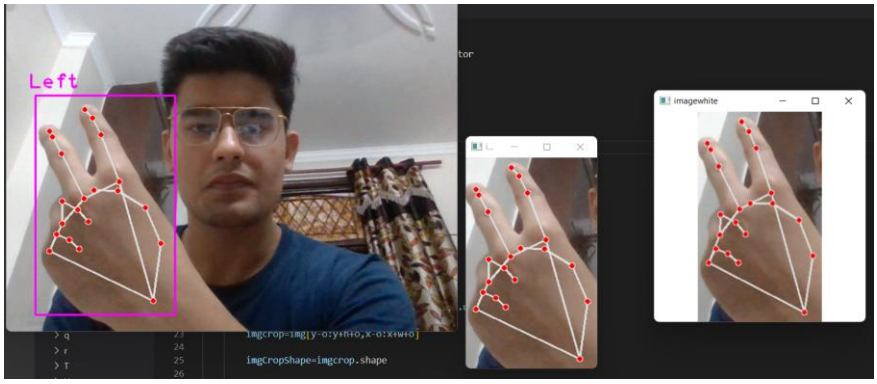


Fig.3 Gestures detected by the model

In Fig. 4,A webpage is created using HTML-CSS and Flask, where the model is connected to the website, the trained model sees then hand gestures to detect and translate into correct. English alphabets

International Journal of Advance Research in Science and Engineering
Volume No. 12, Issue No. 05, May 2023
www.ijarse.com
IJARSE
ISSN 2319 - 8354

Fig.4 Webpage for Sign language translation

The last step occurs by combining different output letters to form meaningful words which are generated by the vector module which is the desired output. As shown in Fig. 5, the model can correctly predict the hand gesture and display output as the English alphabet 'V'.

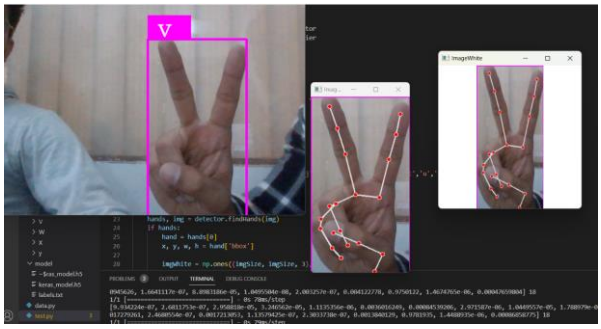
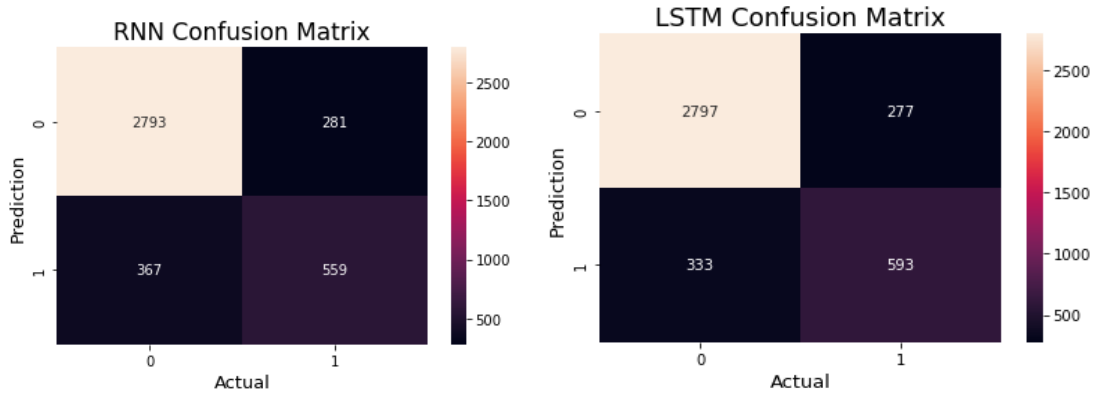
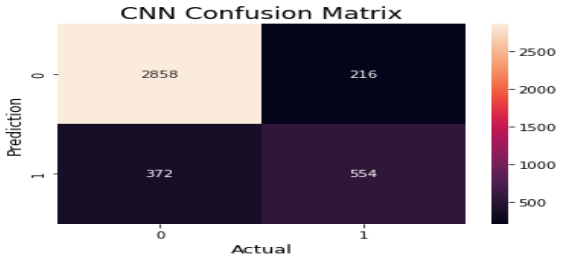


Fig 5. Gestures detected by the model



(a) (b)



(c)

Fig 5. Confusion matrices for different Neural Network models of microscopic blood cell images, (a) RNN (b) LSTM (c) CNN

International Journal of Advance Research in Science and Engineering
Volume No. 12, Issue No. 05, May 2023
www.ijarse.com
IJARSE
ISSN 2319 - 8354

VI. COMPARATIVE ANALYSIS

The recognition and translation system was trained on a no. of different ML as well as Deep Learning models to get as high accuracy as possible from the current scenario. This was the comparison among them: A simple ML model was giving a low accuracy of around 79% due to not having advanced hyperparameter tuning of the model. LSTM was trained on the data and it gave accuracy closer to 87.5% but with some false positive values as there were fewer errors in the data so that the forget gate of a neural network can be used in this dataset.

Table 6. Precision, f1-score, Recall, and accuracy of different Deep Learning models

Model	Precision	Recall	F1-Score	Accuracy
SVM	0.85	0.86	0.84	0.81
LSTM	0.86	0.88	0.88	0.88
RNN	0.88	0.90	0.89	0.90
CNN	0.89	0.90	0.901	0.915

RNN was closer to accuracy around 90% but with false negative cases in some values. CNN was similar to RNN but with some hyper tuning of features, it at the end gave the best accuracy among these models close to 91.5% as the convolutional layers worked fine on the data along with ReLU activation functions calculated values with higher precision. The comparison of various facial recognition models is summarized in the bar chart.

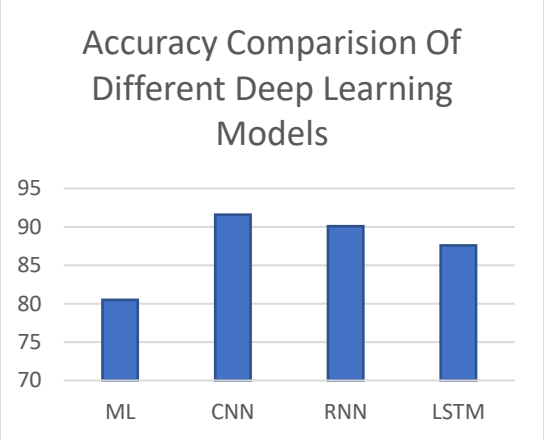


Table 7: Comparison of the proposed deep learning models for sign language detection and translation

References	Description	Used Classifier	Accuracy
[2]	The CNN was implemented to fetch spatial features from the image frames, and the LSTM and RNN were used to capture the temporal dependencies between the frames.	CNN+RNN+LSTM	Test Accuracy = 92.12%

[5]	combined a CNN and an LSTM to recognize sign language gestures. CNN was used to implemented to fetch features from the image frames, and LSTM was used to capture the temporal dependencies between the frames.	CNN+LSTM	Test Accuracy= 95.09%
[6]	CNN and an LSTM to recognize sign language gestures in real-time. CNN was used to extract features from the image frames, and LSTM was used to capture the temporal dependencies between the frames.	CNN+RNN	Validation accuracy = 94.1%
[7]	Their model used a combination of (CNNs) and (LSTM) to fetch features from sign language videos and classify them according to both the sign being made and the user performing the sign.	CNN + RNN	Test Accuracy= 94%
[11]	Chinese Sign Language (CSL) gestures using a combination of CNNs and an (SVM) for user classification.	CNN	Test Accuracy= 90%
[12]	KSL gestures use a deep neural network with an attention mechanism for both sign recognition and user classification.	AL-Net model	Test Accuracy= 92.3%
[13]	Arabic Sign Language (ArSL) gestures using a mixture of CNNs and LSTMs for feature extraction, and a random forest classifier for user classification.	AlexNet	Test Accuracy= 88.4%
[14]	Mexican Sign Language (LSM) gestures using a combination of CNNs and SVMs, with a focus on user classification.	SVM	Test Accuracy= 93.1%
Proposed Sign Language Model	**American Sign Language translation using CNN over elf created hand gestures dataset, with a focus on gesture recognition and translation**	**2D-CNN model**	**Test Accuracy= 91.5%**

VII. CONCLUSION

Currently, the system gives false negatives up to 7%, that is it gives different English letters for different sign languages. One way to make the system more robust is to remove undetectable or blurry images of hand gestures. This demonstrates that while CN-LSTM is an eccentric option for continuous word recognition, CNN provides good accuracy for isolated sign language identification.

REFERENCES

[1] Lewis, M. Paul; Simons, Gary F.; Fennig, Charles D., eds. (2013), "Deaf sign language", Ethnologue: Languages of the World (17th ed.), SIL International, retrieved 2013-12-03

[2] lice Truong, "This Sign Language Ring Translates Hand Movements Into Spoken Words", Fast Company, 2013

[3] P.S Rajam and G. Balakrishnan, "Real-time Indian Sign Language Recognition System to aid deaf-dumb people" in IEEE 13th International Conference on Communication Technology

[4] Rumbaugh, Scott E (2010). "Digital image processing and analysis: human and computer vision applications with CVIP tools" (2nd ed.). Boca Raton, FL: CRC Press. ISBN 9-7814-3980-2052.

[5] Mrs. C. Mythili and Dr. V. Kavitha. Efficient Technique for Color Image Noise Reduction The R e s e a r c h B u lletin of Jordan Association of Computing Machinery, V o l. I I (I I I)

[6] P. C. Cosman, K. L. Oehler, E. A. Riskin, and R. McGary, "Using vector quantization for Image Processing", IEEE, Volume 81, Issue 9, pp. 1326-1341, 1993

[7] D. Singh, S. K. Ranade, "Comparative Analysis of Transform based Lossy Image Compression Techniques" International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 5,1736-1741, October 2012.

[8] G. Vijayvargiya, Dr. S. Silakari, and Dr. R. Pandey, "A Survey: Various Techniques of Image Compression, International Journal of Computer Science and Information Security, Vol. 11, No. 10, October 2013.

[9] Kang, Byeongkeun, Subarna Tripathi, and Truong Q. Nguyen. " Real-time sign language fingerspelling recognition using convolutional neural networks from the depth map." arXiv preprint arXiv: 1509.03001 (2015).

[10] Sign Language Recognition System Using Neural Network for Digital Hardware Implementation Lorena P Vargas, Leiner Barba, C O Torres and L Mattos

[11] MIE324 Final Report: Sign Language Recognition Anna Deza (1003287855) and Danial Hasan (1003132228)

[12] Sign Language Recognition, Generation, and Modelling: A Research Effort with Applications in Deaf Communication Eleni Efthimiou, Stavroula-Evita Fotinea, Christian Vogler, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos & Jérémie Segouat

[13] https://ieeexplore.ieee.org/abstract/document/8630911

[14] https://ieeexplore.ieee.org/abstract/document/8344883

[15] https://ieeexplore.ieee.org/abstract/document/8470384