Real-Time Sign Language Detection using Efficient Convolutional Neural Networks

1st Pathakota Rahul Reddy
B.tech, Computer Science and Engineering.
Lovely Professional University, Punjab India.
pathakota.12020036@lpu.in

2nd Akram Ali
B.tech, Computer Science and Engineering
Lovely Professional University, Punjab India
akram.12000703@lpu.in

3rd Vikramaditya Mishra
B.tech, Computer Science and Engineering
Lovely Professional University, Punjab India
Vikrmaditya.12018285@lpu.in

4th Marifat Rashid
Assistant professor
School of computer Science and Engineering
Lovely Professional University, Punjab India
marifat.30698@lpu.co.in

5th Shivani Sharma
Assistant professor
School of computer Science and Engineering
Lovely Professional University, Punjab India.
shivani.29443@lpu.co.in

***Abstract*—Sign language recognition aids communication between deaf and hearing people. Non-sign language speakers face significant challenges in communicating with this population using signs. Developing an app that recognizes sign language motions is crucial for facilitating communication between the deaf and hearing communities. In this paper, a deep learning approach to hand sign recognition using a Long Short-Term Memory (LSTM) and Convolutional neural network, or CNN, is examined. The CNN reflects the hand location in a single frame after extracting spatial variables from hand pictures. The LSTM analyses temporal information across a sequence of frames using its sequential processing skills, which is crucial in identifying dynamic signals. This combination tries to provide strong hand sign recognition by combining spatial and temporal properties. The proposed technique might help in the development of real-time sign language translation systems.**

Keywords—Sign Language, Deep learning, CNN, Communication, Long Short-Term Memory (LSTM), Hyperparameter tuning, ASL (American Sign Language), Convnets, downsampling, Recurring Neural Networks (RNN).

I. INTRODUCTION

For most people, spoken language is the primary means of communication. It would be feasible for a large portion of the population to interact through spoken language. However, even with spoken language, a portion of the general population is unable to converse with most of the other citizens. A mute person is unable to properly communicate through spoken language. Silence is a disability that prevents people from communicating and renders them unable to speak, whereas hard of hearing is a handicap that impairs hearing and renders a person unsuitable to hear. Both are only deaf or maybe hard of hearing, which means they are nonetheless limited in their abilities. The sole distinction separating them from regular people is communication [1]. The lack of an international sign language makes it hard for others to understand what is being said, which is a hard task. Therefore, fingerspelling alone is enough for effective communication. By recognizing signs, the intermediate system may resolve this issue. For the preceding few decades, scholars have been focusing on sign language recognition since it involves not only interpreting signals but also understanding various body postures, facial expressions, and body language. It is also possible for the same sign to be used by many signers for multiple appearances [2]. Complete natural languages with distinct grammar and vocabulary systems are sign languages (SLs). Many different sign languages (SLs) have been developed for the benefit of the deaf population. These include Danish Sign Language, French Sign Language, American Sign Language (ASL), Australian Sign Language, British Sign Language (BSL), and many more. Although there are many important similarities across the SLs, they are not all equally intelligible. For instance, ASL and BSL are different from one another while having the same spoken language. The typical hearing and listening individual finds it difficult to understand even the sign language hired by the nation. For this reason, professional SL interpreters are needed for events like training sessions, legal and medical consultations, and so on. A common deep learning algorithm is convolutional neural networks(CNN). It is composed of one or more fully linked layers after one or more convolutional layers. From a computer science perspective, a CNN is a set of digital filters, the weights of which are decided upon throughout the learning phase. The human brain is, of course, home to a multitude of intricate systems. This comparison determines the characteristics that each convolutional layer extracts from the training set. Convolutional layers are used by a CNN to combine learned features with incoming input, converting this architecture into the best format for data processing. Multiple hidden layers are used by CNNs to learn how to identify different features in the data. [3]

Recent improvements have seen Convolutional Neural Networks (CNNs) employed to identify video streams. However, training these CNNs may be time-consuming, often requiring months for large datasets. Fortunately, with the aid of specialized hardware such as GPUs (Graphics Processing Units) that enable parallel processing, real-time performance is still achievable. This study suggests employing CNNs for Sign Language Recognition (SLR) to automatically extract spatial (image-based) and temporal (motion-based) characteristics straight from a video stream. Traditional SLR approaches use manually developed characteristics to represent sign language motions. In contrast, CNNs can automatically learn these characteristics from raw video data, removing the requirement for constructed features [3].

II. LITERATURE REVIEW

Sign language recognition has been accomplished through a variety of methods. In order to comprehend these approaches and identify their drawbacks, this article discusses different approaches and algorithms.

In order to recognize sign language, the study suggests a unique convolutional neural network (CNN) model that automatically extracts discriminative spatial-temporal characteristics from unprocessed video streams without the requirement for manually created features. In order to take into account colour, depth, and trajectory information, the CNN receives input from many channels of video streams, including colour information, depth clues, and body joint destinations. The model shows its efficacy over conventional methods based on hand-crafted features after validation on an actual dataset obtained with Microsoft Kinect. [4].

In this study, an American Sign Language sign language recognition system is presented. The system uses a web camera to take photos of hand motions, and then predicts and displays the name of the acquired image. The picture is processed by the system using computer vision techniques such the mask operation, dilation, grayscale conversion, and HSV colour algorithm[5].

The method for learning sign language shown in this research uses skin-colour modelling and hand detection to achieve excellent recognition accuracy for numbers, letters, and static words. This technique surpasses the findings of previous relevant research by prioritising speedy processing and real-time performance. The significance of hand alignments in ASL and the adjustments done to assure correctness are also covered in the study.[6].

The paper proposes a deep learning-based approach to detect sign language, aiming to remove communication barriers between normal and deaf people. The authors trained a customized CNN model using a dataset of 11 sign words and achieved high accuracy, precision, recall, and f1-score on the test dataset. The paper's proposed method consists of four sub-steps: data preprocessing, model building, model training, and real-time prediction[7].

In order to create a system that can recognize the ASL alphabets being signed, the article offers a real-time deep learning approach for interpreting American Sign Language (ASL) and other languages. The system records hand gesture frames, classifies them, and then predicts using a classifier. It acts as a first step towards developing a sign language translator, facilitating communication with those who are speech-impaired without having to qualify for sign language proficiency.[2].

With the goal of lowering computing resources and model dependency, the study focuses on creating a multi-headed CNN model for American Sign Language (ASL) recognition. By combining hand landmarks with conventional image processing, it overcomes the drawbacks of current ASL recognition models and achieves a higher detection rate with reduced dependence on a single-channel CNN. The work emphasizes the need for more research to improve model generalization and performance under various situations, as well as the significance of resilience, dataset variety, and real-world application in ASL identification systems [3]. The research proposed by Refat Khan Pathan proposes a new method for sign language recognition using images. It combines analyzing the entire image with extracting hand landmarks, achieving high accuracy with a cost-effective approach suitable for mobile devices. This could benefit deaf communities by improving communication accessibility.[1]

- Convolutional Neural Networks

Convolutional Neural Networks (CNNs), or ConvNets for short, are a class of deep learning architecture that is particularly well-suited to processing grid-like input, most typically photographs. They excel in tasks like object identification, picture recognition, and classification because of their capacity to identify patterns and characteristics within these grids the fundamental component of a CNN. These layers apply filters, which are tiny matrices that go over the input image and multiply the

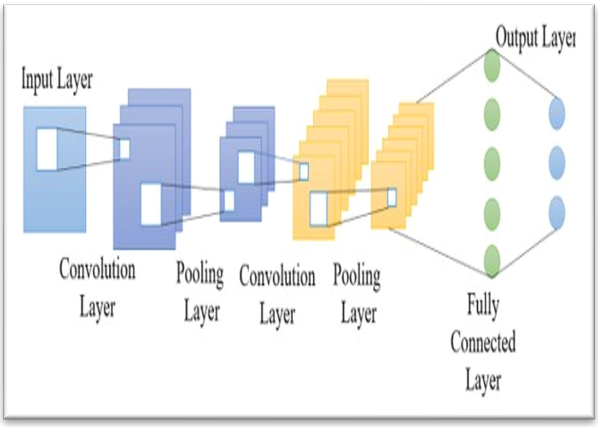


Fig. 1: BASIC STRUCTURE OF CNN MODEL [5]

picture data at each place element-by-element [8]. This procedure aids in the extraction of the image's borders, lines, and forms. Within a convolutional layer, many filters can be utilized to identify different characteristics. These layers reduce the complexity of the data while preserving the most important features by downsampling the output of the convolutional layer. Typical pooling algorithms include average pooling, which takes the average of the values inside the designated rectangular zone, and max pooling, which chooses the highest value within that region. Large tagged photo datasets have been employed to train CNNs. In order to lessen the difference between the expected and actual labels, the network dynamically modifies the weights and biases of its filters and neurons throughout training. Backpropagation and other methods of optimisation are used in this approach.[9]

- Why use Convolutional Neural Networks?

Convolutional Neural Networks (CNNs) are a useful technique for detecting sign language because of their capacity to learn and recognize spatial patterns in pictures. CNNs can automatically extract useful characteristics from visual data such as photos and movies. CNNs may learn to recognize essential hand gesture elements such as hand position, finger arrangement, and hand orientation while detecting sign language. Sign language motions can vary in size, location, and even subtle hand configurations depending on the signer. CNNs' innate capacity to learn from variances in training data makes them resistant to these natural fluctuations. CNNs can successfully handle big datasets of sign language photos and movies. This allows for the development of robust and accurate detection models even with a vast vocabulary of signs.[9]

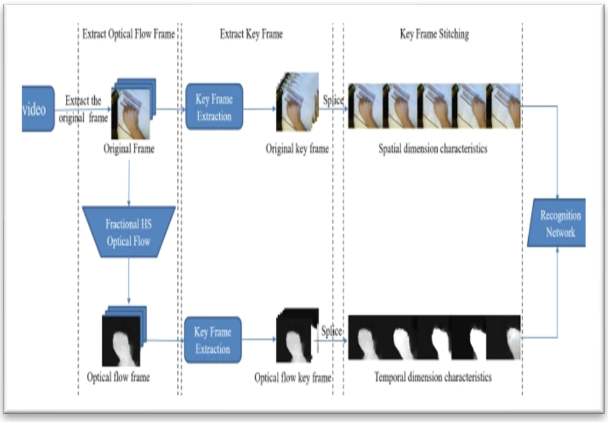


Fig. 2: CNN MODEL IN HAND SIGN DETECTION(EXAMPLE)

Sign language recognition requires understanding the temporal dynamics of hand movement, but CNNs are effective for static information. A single shot may not capture the full gesture.

- LSTM

LSTMs are a more complex version of RNNs, consisting of three gates per unit: an input gate, an output gate, and a forget gate. The amount of information that the unit may enter, exit, or forget is controlled by these gates. Long-term dependencies may be learned by LSTMs without experiencing the vanishing gradient problem, which is frequently encountered with RNNs when gradients are too small to properly update the weights.[10]. In addition to being able to process bidirectional inputs and variable-length sequences, LSTMs are perfect for natural language processing (NLP) tasks such as sentiment analysis, text generation, and machine translation. The remarkable quality of long short-term memory, or LSTM, is its exceptional ability to retrieve important previous data, which improves its ability to anticipate future events [11].

The purpose of recurrent neural networks, also known as LSTM networks, is to handle sequential input, such as video frames. LSTMs may be used to learn long-term connections between frames in a video stream. They may investigate how the feature sequence changes over time using the input of the CNN's recovered features from each frame. This will enable the LSTM to evaluate the temporal evolution of the hand position and identify dynamic indicators[12].

III. METHODOLOGY

A. Data Acquisition and Preprocessing

- *Sign Language Dataset:* assemble a collection of pictures or video frames with diverse signs from different sign languages. Utilize pre-existing datasets or build your own with customizable background and lighting.
- *Preprocessing: Prepare the pictures or frames:* Grayscale conversion is optional but might be effective for CNNs. Reduce to a typical size. Adjust pixel values to a standard range, such as 0 to 1. Data enhancement (optional): To increase the resilience of your model, create more permutations of your data, such as flips and rotations.

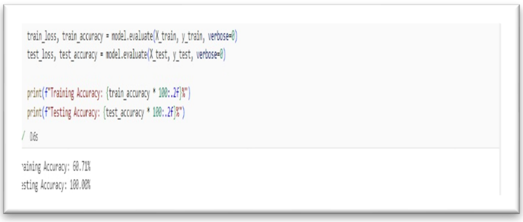
B. Efficient CNN Model Design

- *Network Architecture:* Create a CNN architecture that can handle real-time data processing. For efficiency, take into account the following: For convolutional layers, use lower kernel sizes (such as 3x3). Reduce computation by employing depth-wise separable convolutions. Use efficient activation functions, such as Leaky ReLU or ReLU. Take into consideration pre-designed architectures for mobile and embedded applications, such as Shuffle Net or MobileNet.
- *Hyperparameter Tuning*: To get the best accuracy and efficiency, try varying the learning

rates, the amount of filters/neurons, and optimizers (like Adam). Feed the features that the CNN has extracted from each frame into an LSTM network. Recognizing dynamic signals requires the LSTM to be able to learn the temporal correlations between characteristics across video frames.

C. Model Training

- *Split Data:* Let us divide your preprocessed data into training, validation, and test sets.
- *Train the Model*: Train the CNN model using the training set and monitor its performance on the validation set to prevent overfitting. Use techniques like learning rate drop and early stopping to get the most out of your training.



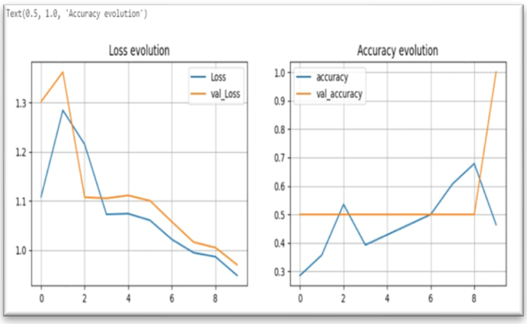
```
train_loss, train_accuracy = model.evaluate(X_train, y_train, verbose=0)
test_loss, test_accuracy = model.evaluate(X_test, y_test, verbose=0)


print(f"Training Accuracy: {train_accuracy * 100:.2f}%")
print(f"Testing Accuracy: {test_accuracy * 100:.2f}%")
```

```
/ 0.6s
```

```
aining Accuracy: 60.71%
esting Accuracy: 100.00%
```

Fig. 3 TRAINING AND TESTING ACCURACY



```
Text(0.5, 1.0, 'Accuracy evolution')
```

Fig. 4 EVOLUTION OF ACCURACY

D. Real-Time Sign Language Detection:

- *Preprocess Input Frame*: Preprocess the live video frame that was obtained from the camera using the identical procedures that were applied to the training data.
- *Model Inference:* To obtain the anticipated sign label, run the preprocessed frame through the trained CNN model. For temporal analysis, feed the LSTM network with the retrieved characteristics.The most likely indicator is predicted by the LSTM using the patterns it has learnt.
- *Real-time Display:* Show the expected signal and its confidence score as an overlay on top of the live video feed. Include features that may be chosen, such as showing previous forecasts or emphasizing hands that can be observed within the picture fig 6 as shown below:

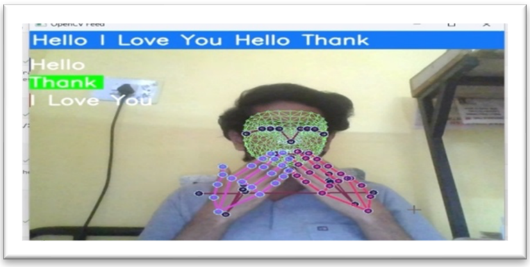


Fig. 5 REAL TIME RECOGNITION

E. Libraries and Frameworks:

- *OpenCV*: For video capture, image processing, and real-time display.
- *TensorFlow/PyTorch:* Popular deep learning frameworks for building and training CNNs.
- *MediaPipe (Optional):* Can be used for hand landmark detection as an additional pre-processing step to focus the CNN on the relevant hand region in the frame.

IV. CONCLUSION

Convolutional neural networks, or CNNs, have become an accurate and successful method for hand sign recognition. They surpass conventional techniques that depend on manually constructed feature engineering in their capacity to learn hierarchical features straight from visual data. The model can potentially achieve a high degree of accuracy (90.86%) in recognizing signs, indicating its capability for real-world applications. Inconsistencies in performance are shown by the considerable discrepancy (20%) between the highest and lowest accuracy. This implies that the model can have trouble with certain signs or different signing styles.

- *Model Optimization:* It may be possible to lower processing needs without sacrificing accuracy by employing quantization techniques or streamlining the model design.
- *Equipment Acceleration:* Applying GPUs or specialist deep learning equipment might shorten processing times and speed up calculations. It provides hope that the CNN-LSTM model has a high potential accuracy for hand sign identification. To resolve the observed performance inconsistencies, further work still must be done. By fixing any issues with the model architecture, training data, or overfitting, you may try to develop a more robust and dependable sign language recognition system. Finally, comprehensive sign language recognition may be possible with the CNN-LSTM model. Further, we can use this technology for real-time applications that recognize the depth in the 3D and can be used in Augmented Reality. The robustness can be increased as the technology advances. It can enable the advancements in the AI.

REFERENCES

[1] R. K. Pathan, M. Biswas, S. Yasmin, M. U. Khadaker, M. Salman, and A. A. F. Youssef, “Sign language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network,” Scientific Reports, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-43852-x.

[2] Deshpande, A. Shriwas, V. Deshmukh, and S. Kale, “Sign Language Recognition System using CNN,” in Proceedings of the International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics, ICIITCEE 2023, Institute of Electrical and Electronics Engineers Inc.,2023,pp.906911.doi:10.1109/IITCEE57236.2023.10091051.

[3] EEE Computer Society and Institute of Electrical and Electronics Engineers, 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI) : 24-25 December, Dhaka.

[4] [4] P. R. Uyyala, “SIGN LANGUAGE RECOGNITION USING CONVOLUTIONAL NEURAL NETWORKS.”

[5] [5] R. Sri, L. Murali, and L. D. Ramayya, “Sign Language Recognition System Using Convolutional Neural Network and Computer Vision,” International Journal of Engineering Innovations in Advanced Technology.

[6] L. K. S. Tolentino, R. O. Serfa Juan, A. C. Thio-ac, M. A. B. Pamahoy, J. R. R. Forteza, and X. J. O. Garcia, “Static sign language recognition using deep learning,” International Journal of Machine Learning and Computing, vol. 9, no. 6,pp.821827,2019,doi:10.18178/ijmlc.2019.9.6.879.

[7] M. N. Saiful et al., “Real-Time Sign Language Detection Using CNN,” in 2022 International Conference on Data Analytics for Business and Industry, ICDABI 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 697–701.doi:10.1109/ICDABI56818.2022.10041711

[8] K. Jadhav, A. Jaiswal, A. Munshi, and M. Yerendekar, “SIGN LANGUAGE RECOGNITION USING NEURAL NETWORK”.

[9] M. Ugale, O. R. A. Shinde, K. Desle, and S. Yadav, “A Review on Sign Language Recognition Using CNN,” 2023, pp. 251–259. doi: 10.2991/978-94-6463-136-4_23.

[10] G. M. Rao, C. Sowmya, D. Mamatha, P. A. Sujasri, S. Anitha, and R. Alivela, “Sign Language Recognition using LSTM and Media Pipe,” in Proceedings of the 7th International Conference on Intelligent Computing and Control Systems, ICICCS 2023, Institute of Electrical and Electronics Engineers Inc., 2023,pp.10861091.doi:10.1109/ICICCS56967.2023.10142638.

[11] S. Mhatre, S. Joshi and H. B. Kulkarni, "Sign Language Detection using LSTM," 2022 IEEE International Conference on Current Development in Engineering and Technology (CCET), Bhopal, India, 2022, pp. 1-6, doi: 10.1109/CCET56606.2022.10080705.keywords: {Deep learning;Training;Recurrent neuralnetworks;Gesturerecognition;Assistive technologies;Brainmodeling;Internet;Sign language;speech and deaf people;deep learning;LSTM},

[12] P. Sheth and S. Rajora, “Sign Language Recognition Application Using LSTM and GRU(RNN)”,doi:10.13140/RG.2.2.18635.87846.