Action Recognition in Sign Language Using Neural Networks

Heet M. Shah
B.E.(CS) Student
Birla Institute of Technology and Science Pilani
Dubai, UAE
f20210125@dubai.bits-pilani.ac.in

Suraj Saini
B.E.(CS) Student
Birla Institute of Technology and Science Pilani
Dubai, U.A.E
f20210235@dubai.bits-pilani.ac.in

Arjun Nadar
B.E.(CS) Student
Birla Institute of Technology and Science Pilani
Dubai, U.A.E
f20210062@dubai.bits-pilani.ac.in

***Abstract*—Sign language is a natural method of communication that uses gestures, hand motions, and facial expressions to transmit messages. It is predominantly utilized by the deaf and mute communities. In recent advances, researchers have used neural networks to improve sign language recognition systems. These devices are intended to improve the required accuracy and efficiency of interpreting sign language gestures. Researchers are employing the deep learning algorithms which consist of convolutional neural network (CNN) and the recurrent neural network (RNN) to develop new approaches for detecting and understanding sign language. This study article intends to not only examine the use of various neural networks but also give a detailed analysis of their accuracy levels, demonstrating the potential of deep learning algorithms in considerably enhancing the detection and interpretation of sign language gestures with high precision and efficiency.**

Keywords—Sign language, Neural Networks, Deep Learning Algorithms

I. Introduction (*Heading 1*)

Sign language is a very important means of communication for the deaf people and hard of hearing population, and its acknowledgment is critical for improving communication between the community and the broader public. The development of sign language recognition systems has piqued the interest of academics and developers due to the potential benefits for the deaf and hearing-impaired population. However, detecting activities in sign language presents various obstacles, including the heterogeneity in hand form, motion profile, and location of the hand, face, and body components that contribute to each sign.

To solve these issues, researchers used a variety of methods, including deep learning approaches. The Convolutional Neural Networks (CNNs) and the Long Short Term Memory (LSTM) networks are commonly utilized for sign language recognition applications. These models excel in extracting characteristics from visual data and capturing temporal correlations, both of which are critical for understanding the dynamic nature of sign language motions.

CNN (Convolutional Neural Network) is thought to be the best for action recognition because of its ability to successfully extract 2D spatial characteristics from still photos. In the domain of action recognition, CNNs excel at collecting spatial patterns and characteristics inside pictures, making them ideal for jobs that require evaluating and detecting actions based on visual input. Furthermore, CNNs have exhibited improved performance in picture classification tasks, which applies to action identification, where precisely detecting and categorizing individual actions is critical. CNNs' hierarchical feature learning method using convolutional operations enables them to learn complicated patterns and representations, making them an effective tool for identifying actions in photos or movies.

LSTM (Long Short Term Memory) networks are regarded as among the finest for sign language recognition due to their capacity to successfully capture long-term dependencies and temporal correlations in sequential data. In the domain of sign language recognition, where movements occur over time, LSTM networks excel at comprehending and identifying continuous sequences of signs or phrases. This capacity makes them particularly ideal for jobs like converting sign language to text in real-time and properly identifying dynamic sign language motions.

Furthermore, research has demonstrated that, while CNNs perform well for isolated sign language identification, LSTM models outperform them in continuous word recognition tasks, demonstrating LSTM networks' ability to handle sequential data and continuous motions inherent in sign language communication.

LSTM networks' unique architecture, combined with their ability to retain information over long sequences and mitigate vanishing gradient problems, makes them an effective tool for capturing the nuances and complexities of sign language gestures, resulting in more precise and efficient recognition systems.

The study paper conducts a thorough investigation of several neural networks for the process in sign language recognition, to identify the best successful models for this job.

The following data formats are utilized for sign language recognition using neural networks:

Image Data: Images of sign language motions are frequently used to train neural networks in sign language recognition systems. These photos record the hand and body motions associated with distinct signals, allowing the neural network to learn and detect the gestures correctly.

Feature Vectors: Feature vectors collected from photos are critical in training neural networks for sign recognition. These vectors reflect fundamental properties of the sign language motions, allowing the neural network to comprehend and categorize the signs efficiently.

Pixel information from photos is used to generate datasets for training neural networks in sign language recognition. By processing and molding pixel data into visuals, the neural network may learn to detect and discriminate between distinct sign language signals

Stored Images: Systems that save particular sign language symbols in picture form are used to train multilayer neural networks with backpropagation methods. These saved pictures serve as training material for the neural network to learn and reliably recognize distinct signals.

Gesture Images: Images of static and dynamic gestures, such as the sign alphabet, are used to train neural networks in sign language recognition. These gesture representations help the network learn to identify and categorize numerous signs successfully.

The study article investigates numerous neural networks for sign language recognition in order to determine the most successful models for the job. The study seeks to determine the optimal method for accurate and efficient sign language recognition by comparing the performance of several designs, such as a Convolutional Neural Network (CNNs) and Long Short Term Memory (LSTM) network.

II. Background Theory

This section introduces convolutional neural networks and compares various network topologies.

A. Concepts of Neural Network Architecture

Neural networks are a key component of artificial intelligence, inspired by the structure and function of the human brain. They are made up of linked nodes or neurons structured in layers that process data like the neural connections seen in the brain. In neural networks, whose information passes across these layers, with each neuron receiving input, processing it, and sending the output to the next layer. This method enables neural networks to learn from data, make judgments, and improve their performance found over time via a training procedure.

The Long short term memory (LSTM) networks are a type of recurrent neural network (RNN) that is commonly employed for sequential data processing applications such as natural language processing, time series analysis, and speech recognition.

To understand how LSTM networks operate, consider the LSTM cell, a fundamental unit of LSTMs. The LSTM cell maintains a memory state that can selectively remember or forget information over time, allowing it to capture long-range dependencies in sequential data.

Equation for LSTM Gate Operations:

The LSTM gates, which include some initial input gates, forget gate and finally the output gates, control the flow of information within the LSTM cellular structure. These gates are regulated by sigmoid activation functions, which influence cell state and hidden state dynamics.

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$$
$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$$
$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$$

Equation 1: Equation for Gates

where,

$i_t \rightarrow represents\ input\ gate.$
$f_t \rightarrow represents\ forget\ gate.$
$o_t \rightarrow represents\ output\ gate.$
$\sigma \rightarrow represents\ sigmoid\ function.$
$w_x \rightarrow weight\ for\ the\ respective\ gate(x)\ neurons.$
$h_{t-1} \rightarrow output\ of\ the\ previous\ lstm\ block(at\ timestamp\ t-1).$
$x_t \rightarrow input\ at\ current\ timestamp.$
$b_x \rightarrow biases\ for\ the\ respective\ gates(x).$

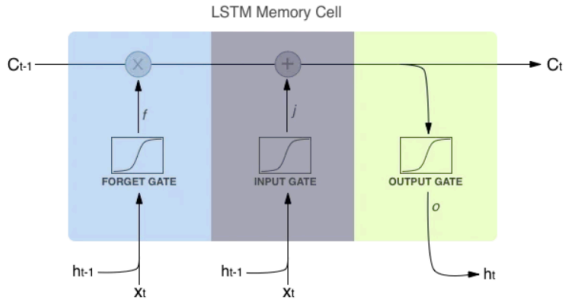

Figure 1: LSTM Memory Cell

.

The first equation is used to find the Input Gate, which can inform us what new information we will store in the cell states (as seen below).

The second is for the important forget gates, which indicates which information should be removed from the cell states.

The third is used for the output gates, that's used to set off the LSTM block's calculated very last output at the timestamp 't'.

B. Comparison of several neural network architectures

Neural Network	Type of data	Approach	Performance
Artificial Neural Network (ANN)	Tabular data	Feed-forward architecture	High accuracy in tabular data problems
Convolutional Neural Network (CNN)	Image Data	Convolutional layers, pooling layers, and fully connected layers	High accuracy in image recognition problems
Recurrent Neural Network (RNN)	Sequence data	Recurrent connections, backpropagation through time (BPTT)	Ability to work with incomplete knowledge, useful in time series prediction
Long Short-Term Memory (LSTM)	Sequence data	Memory cells, self-learns during backpropagation	Improved accuracy in handling long-term dependencies in sequence data
Hybrid Network	Various data	Combination of multiple neural network architectures	Enhanced performance by leveraging the strengths of multiple network types

Table 1: Comparison of different NN's

When comparing the Convolutional Neural (CNNs) and Long Short Term Memory (LSTM) type network to other neural networks for action detection, CNNs perform better in image-related tasks such as object detection and video analysis, whereas LSTMs excel at modeling dynamic time series data, making them ideal for recognizing complex sequences of actions in videos.

III. Limitations in Action detection

The limitations of action detection in sign language using neural networks include-

Limited datasets: The availability of extensive datasets for sign language recognition is restricted, which may impair model accuracy and generalization.

Variability in Sign Language: Sign language is a complex and dynamic system that includes a variety of elements such as hand forms, gestures, and facial expressions. The variety of these characteristics might make it difficult to recognize and understand sign language motions effectively.

Lighting Conditions and Camera Angles: Variations in lighting and camera angles can have an impact on the accuracy of sign language recognition systems because they alter the input data.

Real-time processing: Real-time sign language recognition is a demanding task that requires efficient processing of large amounts of information. The current hardware and software platforms may limit the development of real-time sign-language recognition systems.

Regarding the limitations of using LSTM for action detection:

LSTM Limitations: LSTMs are susceptible to vanishing gradient difficulties, which might impair their capacity to record long-term relationships in sequential data, potentially affecting the accuracy of sign language recognition systems.

The usage of normalization layers in LSTM models can influence prediction results. Normalization parameters, such as rescaling and symmetric normalization, can affect LSTM networks' accuracy and performance in forecasting time series data.

In time series forecasting applications, LSTM networks may have a lag between forecasts and real values. This lag can be modified by factors such as the model's structure, input data, and prediction horizon, resulting in differences between anticipated and actual values.

To summarize, while LSTM networks have demonstrated promising results in action detection for sign language, problems, and limits must be addressed to increase their performance and accuracy. More research is needed to address these constraints and create more robust and efficient sign language recognition systems.

IV. RESULT AND DISCUSSION

1) Training and Validation Based Graphs

A) Training & Validation - Model loss graph

A machine learning model's performance during training may be shown visually via a training loss graph. The overall trend of the model's loss (or mistake) over epochs is shown in this graph. Loss is a metric that expresses how well the model performs; lower numbers signify better performance.

Loss - 0.0662

Validation Loss - 0.003

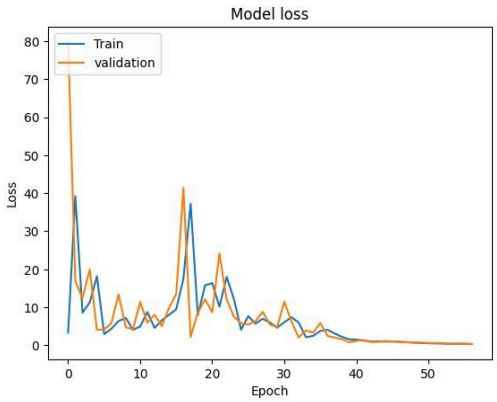

Figure 2: Training - Model Loss Graph

B) Training & Validation - Model Accuracy Graph

It is a visual representation of how the accuracy of a machine learning model changes over epochs during both the training. A popular indicator for assessing classification models is accuracy, which shows the percentage of examples that are properly identified out of all the occurrences.

Accuracy - 97.62%

Validation Accuracy - 97.8%

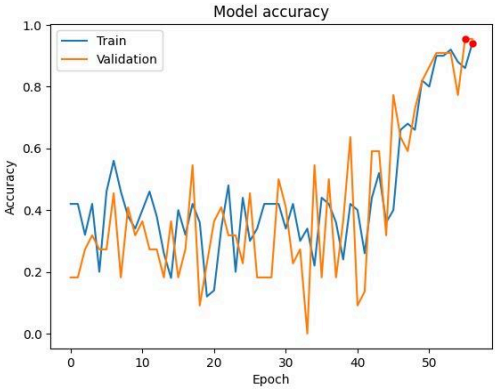

Figure 3: Training - Model Accuracy Graph

2) Precision - Recall graph

A precision-recall graph is a visual representation of the trade-off between precision and recall for different thresholds in a binary classification model. Two crucial variables, recall and precision, are used to assess how well classification models perform, particularly when dealing with unbalanced datasets.

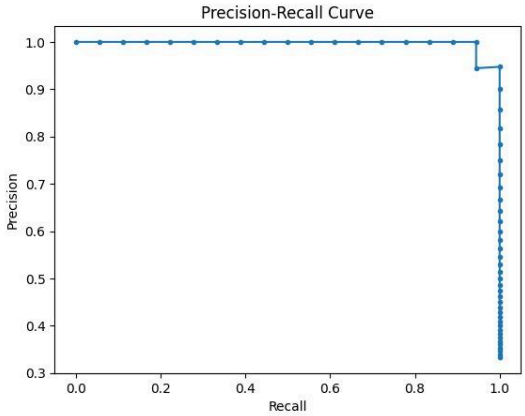

Figure 4: Precision Recall Graph

3) Receiver Operating Characteristic(ROC) Curve

It is a diagnostic diagram of the binary classification model of various distinctions. Compare the variable's true positive (sensitivity or regression) to its false positive (1 - specificity) using the ROC curve.

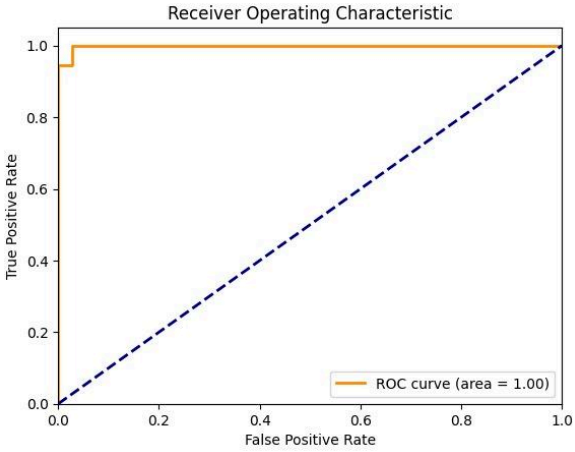

Figure 5: ROC Graph

4) Confusion Matrix

A table used in classification to evaluate a machine learning model's efficacy is called a confusion matrix. It presents the number of accurate and incorrect predictions the algorithm produced, as well as a summary of the results of a classification exercise. The confusion matrix is particularly effective for binary and multiclass classification issues. Here's how a confusion matrix is usually structured:

True positive (TP) : 6

False positive (FP) : 1

False negative (FN) : 0

True negative (TN): 3

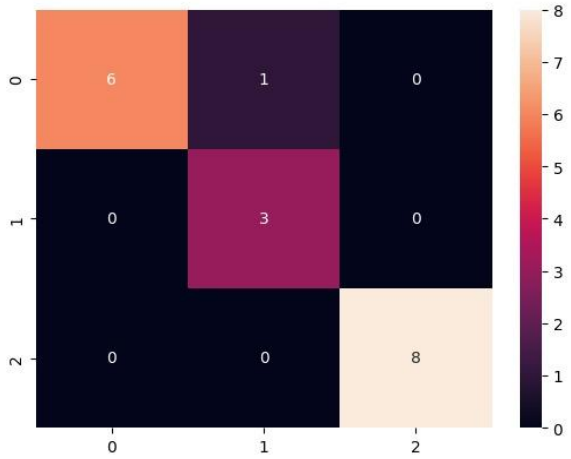

Figure 6: Confusion Matrix

5) Bar Graph-Classification Report

1. Recall rate = TP/(TP+FN) {Equation 2}

2. Specificity= TN/(TN+FP) {Equation 3}

3. Precision rate= TP/(TP+FP) {Equation 4}

4.F1-score=2*(Precision*Recall)/(Precision+ Recall)

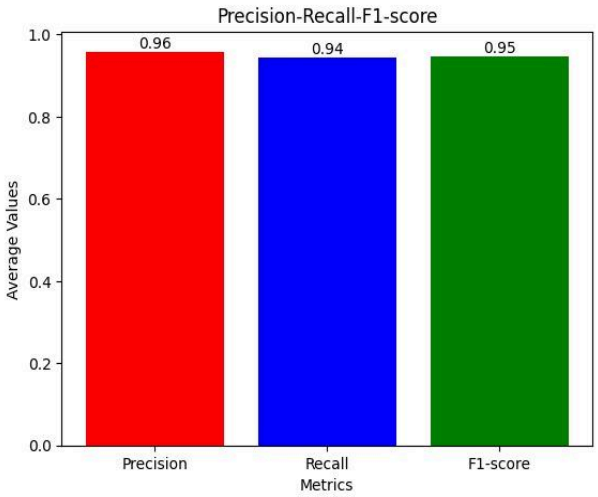

Figure 7: Precision-Recall-F1-score bar graph

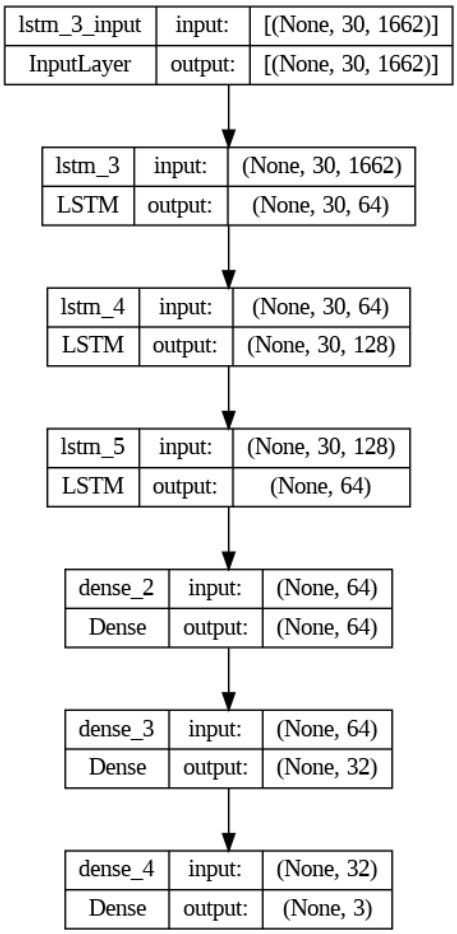

Figure 8: LSTM model we used

The above image represents the LSTM model, which is a sequential model and has 3 LSTM layers to it having different number of neurons 64, 128, 64 respectively and all of them uses ‘relu’ as there activation function then we add 2 more fully connected dense layers having 64, 32 neurons respectively and again ‘relu’ as there activation function and Finally, we have a dense output layer with three neurons, which represents the number of classes in the classification job, and the activation function is 'softmax', which is widely used for multi-class classification issues since it produces a probability distribution across the classes.

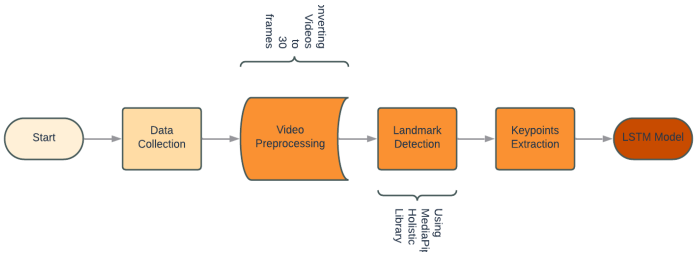

Figure 9: Flow chart of the process of our Model until it reaches the LSTM Model

Paper title	Author	Focus	Network Type	Dataset	Accuracy	Key Findings
Human Action Recognition using Hybrid Deep Evolving Neural Networks	P. Dasari, L. Zhang, Y. and R. Gao(2022)	For the purpose of classifying video actions, hybrid deep neural networks—that is, Convolutional Long Short-Term Memory (ConvLSTM) Networks and Long-term Recurrent Convolutional Networks (LRCN)—are used.	Long-term Recurrent Convolutional Networks (LRCN) and Convolutional Long Short-Term Memory (ConvLSTM) networks)	UCF50 dataset	Accuracy rate of 88.6%.	For the categorical analysis of 50 action classes in UCF50, the LRCN model with MobileNet as the encoder and a BiLSTM network as the decoder achieves the best accuracy rate, or 87%.
Sign language recognition based on spatiotemporal convolutional neural network and attention mechanism	Y. Shao and G. Chen	The problem with the isolated words of sign language in video recognition is studied through the fusion of the r(2+1)D network and attention mechanism.	R(2+1)D network	CSL dataset	92.2% and the lowest loss is 0.274938	Can successfully finish the sign language recognition task.
Sign Language Identification using Skeletal Point-based Spatio-Temporal Recurrent Neural Network	J. Johnson, J. Joseph, M. Reji, M. E. George and T. D. Sajanraj	Building a model for hand gesture recognition by using skeletal-point feature extraction framework to extract hand landmarks from sequences including unique signs.	Long Short-Term Memory (LSTM) Networks	ISL datasets	Success rate of 99%	The skeleton data is a faster learning candidate, and the framework performs better than CNN.

Dynamic Two Hand Gesture Recognition using CNN-LSTM based networks	V. Sharma, M. Jaiswal, A. Sharma, S. Saini and R. Tomar	Process for deploying the specific neural network on embedded hardware	CNN-LSTM architecture	20,000 photos that were taken with a Digital Single-Lens Reflex (DSLR) camera in quick succession	Accuracy of 99.14%	With an accuracy of 99.14%, CNN-LSTM outperforms other models that cause overfitting.
Recognition of Visually Perceived Compositional Human Actions by Multiple Spatio-Temporal Scales Recurrent Neural Networks	H. Lee, M. Jung and J. Tani	A deep learning model for action recognition that concurrently takes raw RGB input data and extracts spatiotemporal information	Multiple spatio-temporal scales recurrent neural network (MSTRNN)	The 900 movies in CL1AD are from 10 participants doing 9 different activities that are aimed at 4 different objects.	The standard error of 4.60% and a mean recognition rate of 89.30%	Because of its recurrent connection, the suggested model has the potential to perform well on object identification tasks.
Trajectory-based Fisher kernel representation for action recognition in videos	I.Atmosukarto , B. Ghanem, and N. Ahuja	Action recognition in videos using a trajectory-based Fisher kernel representation approach	Fisher Kernel combined with Gaussian Mixture Model(GMM) and Support Vector Machines(SVM)	Two Benchmark Datasets for video recognition	Improved accuracy compared to BOF approach for the same.	Analysis using a comparison between the Fisher Kernel method and BOF approach. Fisher kernel shows a superior performance for action recognition.
Action Localization and Recognition through Unsupervised I3D and TSN	M. Umran, K. Muchtar, T. F. Abidin, and F. Arnia	Automating human action recognition in CCTV cameras and systems.	Inflated 3D(I3D) and temporal segment networks(TSN)	Collected CCTV video data.	Prediction Similarity rate of above 84.6% for 54 out of 101 action classes.	Development of methods to automate action recognition by using local GPU capabilities to identify regions of interest(ROI).

A Multi-Scale Hierarchical Codebook Method for Human Action Recognition in Videos Using a Single Example	M. J. Roshtkhari and M. D. Levine	Using a technique in local spatiotemporal video volume (STVs), human activity recognition	Hierarchical codebook approach.	Three benchmark datasets- KTH, Weizmann, and MSR II	Improvement of over 50% in accuracy compared to BOF methods.	Understanding spatial and temporal changes and deformation in videos. Using a codebook to identify actions in real-time.
Video action recognition method based on attention residual network and LST	Y. Zhang and P. Dong	Video action recognition using attention residual networks in combination with LSTM.	Attention residual network and LSTM	Dataset obtained from YouTube videos.	A recognition rate of 95.45% based on the dataset.	An improvement over traditional data preprocessing and sampling techniques. Integration of Convolutional block attention module to identify features.
Large-Scale Weakly-Supervised Pre-Training for Video Action Recognition	D. Ghadiyaram, D. Tran, and D. Mahajan	To enhance pre-training techniques for video action detection and overcome the shortcomings of the supervised video datasets currently available.	Techniques for pre-training a high number of online videos.	A large volume of over 65 million videos was obtained from the web.	Significant improvements in recognition rates compared to current methods.	Importance of optimizing label spaces for videos. Developing benchmarks to identify videos in detail. Focus on temporal localization of videos.
Human Action Recognition Using Deep Learning Methods	Zeqi Yu, Wei Qi Yan	Exploring the effectiveness of different deep learning architectures such as Two-Stream CNN, CNN+LSTM, and 3D CNN for accurately identifying human actions in videos.	Two-Stream CNN, CNN+LSTM, and 3D CNN	HMDB-51 dataset	CNN+LSTM model achieved an accuracy of 89.74%. The two-Stream CNN model achieved an accuracy of 82.37%. The 3D CNN model achieved an accuracy of 86.54%.	The effectiveness of Two-Stream CNN and CNN+LSTM models, challenges with 3D CNN in handling complex temporal relationships, and successful recognition across different viewing angles

						and lighting conditions.
Human Activity Recognition Using Convolutional Neural Networks	Gulustan Dogan, Sinem Sena Ertas, and İremnaz Cay	Using convolutional neural networks (CNNs) to recognize human activities based on smartphone sensor data, particularly from the linear accelerometer, gyroscope, and magnetometer sensors.	Convolutional Neural Networks (CNNs)	Sussex-Huawei Locomotion (SHL) dataset	accuracies ranging from approximately 58% to 79% for various experiments conducted.	It indicates better recognition for slower movements but struggles with similar activities like train and subway travel. Integration of additional sensor data is suggested for potential accuracy improvements.
Video-based Human Action Recognition Using Deep Learning: A Review	Hieu H. Pham, Louahdi Khoudour, Alain Crouzil, Pablo Zegers, and Sergio A. Velastin	Its primary aim is to accurately describe human actions and their interactions from previously unseen data sequences acquired by sensors.	Deep Belief Networks (DBNs), Convolutional Neural Networks (CNNs), Stacked Denoising Autoencoders (SDAs), and RNN-LSTMs.	KTH IXMAS Hollywood-1 Hollywood-2 YouTube MuHAVi UT-Interaction MSR Action3D HMDB-51 UCF-101 Sports-1M ActivityNet NTU RGB+D	CNNs achieved accuracies of 74.22% (spatial), 82.34% (temporal), and 85.94% (spatio-temporal) on UCF-101 and HMDB51 datasets.	Deep learning techniques have recently shown significant improvement in human action recognition. Also found that combining different deep learning architectures can lead to even better performance.
Transfer Learning For Videos: From Action Recognition To Sign Language Recognition	Noha Sarhan; Simone Frintrop	They aimed to address challenges in sign language recognition by leveraging transfer learning from action recognition and proposed a novel approach based on two-stream inflated 3D ConvNets.	Inflated 3D (I3D) Convolutional Neural Network	ChaLearn249 Isolated Gesture Recognition dataset	Accuracy of 64.44% when utilizing combined RGB and flow data, 57.73% for the RGB stream, and 57.68% for the optical flow stream.	The proposed method demonstrates promising results, even in the absence of depth data.

Human Action Recognition in Video Using DB-LSTM and ResNet	Akram Mihanpour; Mohammad Javad Rashti; Seyed Enayatallah Alavi	creating a technique that combines Deep Bidirectional Long Short-Term Memory (DB-LSTM) networks with Convolutional Neural Networks (CNN) to recognize human actions in videos.	Deep bidirectional long short-term memory (DB-LSTM) with convolutional neural networks (CNN)	UCF 101 dataset	Observe an average of about 95% accuracy	Parallel processing using multiple threads reduces training time, making the method suitable for real-time action recognition tasks.

Table 2: Literary survey

VI. CONCLUSION

While the LSTM type network has shown good results in detecting and interpreting sign language gestures, challenges and limitations still need to be addressed to enhance their performance and accuracy. The study emphasizes the importance of overcoming limitations such as the availability of comprehensive datasets, variability in sign language gestures, lighting conditions, camera angles, and real-time processing constraints. Specifically, LSTMs are adept at understanding and recognizing continuous sequences of signs or words due to their ability to capture long-term dependencies and temporal relationships in sequential data. Despite the strengths, LSTMs can face issues like vanishing gradient problems, affecting their accuracy in recognizing sign language gestures. The research underscores the need for further investigation to address these challenges and develop more robust and efficient sign language recognition systems that leverage the strengths of LSTMs effectively.

VII. CONTRIBUTION

Suraj worked on the preprocessing and the extraction of landmarks from the videos part, Arjun worked on creating the dataset and converting the videos into arrays using Suraj's work and Heet worked on making the LSTM model and also saving the numpy values back to the array from the dataset so that it can be used for training and making the plots. Together we worked on implementing the real time recognition code.

VIII. REFERENCES

[1] P. Dasari, L. Zhang, Y. Yu, H. Huang and R. Gao, "Human Action Recognition Using Hybrid Deep Evolving Neural Networks," 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 2022, pp. 1-8

[2] Y. Shao and G. Chen, "Sign language recognition based on spatiotemporal convolutional neural network and attention mechanism," 2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Xi'an, China, 2022, pp. 519-522,

[3]J. Johnson, J. Joseph, M. Reji, M. E. George and T. D. Sajanraj, "Sign Language Identification using Skeletal Point-based Spatio-Temporal Recurrent Neural Network," 2023 9th International Conference on Smart Computing and Communications (ICSCC), Kochi, Kerala, India, 2023, pp. 465-470

[4] V. Sharma, M. Jaiswal, A. Sharma, S. Saini and R. Tomar, "Dynamic Two Hand Gesture Recognition using CNN-LSTM based networks," 2021 IEEE International Symposium on Smart Electronic Systems (iSES), Jaipur, India, 2021, pp. 224-229

[5] H. Lee, M. Jung and J. Tani, "Recognition of Visually Perceived Compositional Human Actions by Multiple Spatio-Temporal Scales Recurrent Neural Networks," in IEEE Transactions on Cognitive and Developmental Systems, vol. 10, no. 4, pp. 1058-1069, Dec. 2018

[6] I. Atmosukarto, B. Ghanem and N. Ahuja, "Trajectory-based Fisher kernel representation for action recognition in videos," Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 2012, pp. 3333-3336.

[7] M. Umran, K. Muchtar, T. F. Abidin and F. Arnia, "Action Localization and Recognition through Unsupervised I3D and TSN," 2023 3rd International Conference on Computing and Information Technology (ICCIT), Tabuk, Saudi Arabia, 2023, pp. 269-273, doi: 10.1109/ICCIT58132.2023.10273877.

[8] M. J. Roshtkhari and M. D. Levine, "A Multi-Scale Hierarchical Codebook Method for Human Action Recognition in Videos Using a Single Example," 2012 Ninth Conference on Computer and Robot Vision, Toronto, ON, Canada, 2012, pp. 182-189, doi: 10.1109/CRV.2012.32.

[9] Y. Zhang and P. Dong, "Video action recognition method based on attention residual network and LSTM," 2021 33rd Chinese Control and Decision Conference (CCDC), Kunming, China, 2021, pp. 3611-3616, doi: 10.1109/CCDC52312.2021.9601577.

[10] D. Ghadiyaram, D. Tran and D. Mahajan, "Large-Scale Weakly-Supervised Pre-Training for Video Action Recognition," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 12038-12047, doi: 10.1109/CVPR.2019.01232.

[11] Z. Yu and W. Q. Yan, "Human Action Recognition Using Deep Learning Methods," 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ), Wellington, New Zealand, 2020, pp. 1-6, doi: 10.1109/IVCNZ51579.2020.9290594.

[12] G. Dogan, S. S. Ertas and İ. Cay, "Human Activity Recognition Using Convolutional Neural Networks," 2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Melbourne, Australia, 2021, pp. 1-5, doi: 10.1109/CIBCB49929.2021.9562906.

[13] Pham, Hieu & Khoudour, Louahdi & Crouzil, Alain & Zegers, Pablo & Velastin, Sergio. (2022). Video-based Human Action Recognition using Deep Learning: A Review. 10.48550/arXiv.2208.03775.

[14] N. Sarhan and S. Frintrop, "Transfer Learning For Videos: From Action Recognition To Sign Language Recognition," 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 2020, pp. 1811-1815, doi: 10.1109/ICIP40778.2020.9191289.

[15] A. Mihanpour, M. J. Rashti, and S. E. Alavi, "Human Action Recognition in Video Using DB-LSTM and ResNet," 2020 6th International Conference on Web Research (ICWR), Tehran, Iran, 2020, pp. 133-138, doi: 10.1109/ICWR49608.2020.9122304.