International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

AI-Based Sign Language Proficiency Assessment and Learning System using Hand Gesture Recognition

Muthamma A A[1], Meghaswini K M [2], Abhilash Shetty[3]
Student, Department of BCA, St. Francis Colleeg, Bengaluru[1]
Student, Department of BCA, St. Francis Colleeg, Bengaluru[2]
Assistant Professor, Department of BCA, St. Francis Colleeg, Bengaluru[3]
Email id: abhilashshetty633@gmail.com

Abstract - The use of sign language serves as an essential means of communication for deaf and mute individuals. Nevertheless, learning sign language may be difficult for beginners due to the requirement of consistent practice, appropriate guidance, and feedback. Presently, most sign language recognition systems are oriented to recognizing or translating the signs while providing little attention to assessment and enhancement of the sign language proficiency. This research aims at designing an AI-enabled interactive system that will assess and improve the proficiency of the learner through hand gesture recognition and intelligent feedback. The hand gestures are captured via webcam and hand joint positions are extracted using MediaPipe. Hand gesture recognition and evaluation are performed using deep learning techniques like Dense Neural Network (DNN). Learner's gestures are evaluated against the predefined reference signs and any inconsistencies in terms of hand shape, position, and movement are identified. The AI-enabled feedback module guides the user by giving immediate instructions that help them in correcting the mistakes while practicing. This system tries to ensure independence in learning, increase gesture accuracy, and develop sign language fluency through continued practice. Future improvements could come in the form of increased databases, advanced deep learning algorithms, and personal learning methods to improve recognition accuracy and learner-specific feedback.

Keywords: Sign Language, Artificial Intelligence, Hand Gesture Recognition, MediaPipe, Dense Neural Network (DNN), Deep Learning, Computer Vision.

1.INTRODUCTION:

There is also an idea of a model based on Transformer, which was proposed as a way of recognizing and translating sign language gestures. The recognition part is responsible for processing continuous signing and is trained by connectionist temporal classification. This approach reached the BLEU-4 metric of 21.80 when tested on the PHOENIX14T dataset.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

However, there is an alternative approach that allows for real-time continuous recognition of sign language gestures through the use of a regular RGB camera. After the initial estimation of the approximate position of the palm, the subsequent estimation of 21 keypoints is performed. The absence of any additional equipment in the process allows us to implement an efficient real-time gesture recognition algorithm.

Furthermore, another research conducted a corpus of 21,083 video clips that contained 2000 signs in American Sign Language, made by 119 unique persons. It was proved that as the number of layers in the model increased, the ability to recognize a large number of words became even more important. It was also noted in the paper that as the vocabulary of the signer grew, the difficulty of SLR also increased.

Part Affinity Fields (PAFs) is a continuous pose estimation model that detects the key points of the human body, hands, and facial landmarks to construct pose estimation. It constructs an efficient deep learning model for pose estimation, which serves as the feature extractor. Although PAFs were not developed for the purpose of recognizing sign languages, the body pose can be represented in terms of a graph, whereby the graph nodes represent the body joints and the edges symbolize their relation to adjacent joints. Therefore, it allows for the continuous sign language pose estimation and exhibits great promise for developing graph-based multimodal systems for processing dynamic gestures.

Moreover, a huge benchmark has more than 25,000 videos of American Sign Language annotations having more than 1000 signs performed by more than 200 speakers, which allows testing of the models on an even multi-speaker benchmark. Additionally, the benchmark contains over 80 hours of sign language, more than 35,000 utterance annotations with RGB videos, depth videos, audio, and English transcripts providing diverse data for conducting experiments. Since it is necessary to have a benchmark to conduct multimodal research, it provides a variety of data like videos, audio, and texts.

This study covered more than 300 papers on sign language recognition and listed 400 of the studies which have been done on them. The main trend was towards automatic vision based sign language recognition, along with increased attention towards general vocabulary recognition as well as non-rigid continuous sign language recognition techniques. An open source graph based visual SLR framework which can process continuous video streams with GPU acceleration has been introduced through the paper.

The use of deep learning techniques, especially convolutional networks, has made a tremendous contribution to improving the output of automatic sign language recognition systems and has revealed the importance of the hands, face, and body in this kind of recognition. The recognition of continuous signs is harder than the recognition of isolated signs since in the case of continuous signs it is required to split the video into individual signs during the recognition process. An end-to-end system that recognizes continuous sign language must be designed for the recognition of continuous sign language because of its effect on recognition accuracy.

However, the SLR systems should also consider other factors like various facial expressions and the motions of the head and body, because signing language includes much more than just hand gestures. The capability to recognise continuous sign language can be considered a very important part of automatic sign language recognition systems. Different solutions can be employed for improving such systems' performance, for example, multiview techniques, triangulation, recurrent 3D-CNN, etc. But the problem of occlusions and complex dynamic gesture patterns persists when it comes to continuous sign language recognition.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

The use of wearable sensors for automatic identification of sign languages is one of the most interesting fields of investigation since when applied to the skin, IMU and sEMG sensors can detect many different features. However, despite that these approaches are characterized by a high precision in identifying particular signs, they have a restricted vocabulary and small amount of users.

The problem of continuous recognition can be solved using MediaPipe Holistic Landmark Detection because it involves multiple streams of data, making it a more reliable approach that involves information from various sources. This approach allows the inclusion of more data streams and more features in solving particular problems related to the continuous sign language recognition based on the initial dataset.

2. RELATED WORK

2.1 Sign Language Recognition

In recent times, with the advent of advanced techniques, there have been developments in methods and techniques for Sign Language Recognition. This includes the use of methods like the use of Convolutional Neural Networks which help in learning patterns in the data and Long Short-Term Memory networks that work well in dealing with time series. The 3D-CNN works on three-dimensional signs with respect to time, while the CNN-LSTM gives the results for dynamic signs.

Moreover, there are methods for detection of position of different parts of the body using MediaPipe landmarks, suitable both for isolated and continuous signing. However, there are some limitations of these approaches which influence the accuracy of the results. First of all, a great number of words is the main issue for Sign Language Recognition systems. Secondly, signs may be executed in different ways depending on users, which means the accuracy of recognition will be decreased. Furthermore, the lighting plays an important role because the accuracy of the system will be lowered with poor lighting conditions. The next obstacle for development of such systems is the issue of occlusion which leads the system to give wrong results. The final component that needs to be considered is the background which should be excluded from the process to avoid any distractions. Temporal complexity is also the factor which influences the accuracy of the process.

These problems must be considered and addressed in order to improve sign language recognition systems. The improvement of these kinds of systems is very important in order to facilitate communication.

Table 1 Importance of datasets in the present study

Input	Content used	Purpose
WLASL[3]	21,083 videos of 2,000 ASL words from 119 signers, in subsets of 100, 300, 1,000 and 2,000 words	Word-level recognition; shows how accuracy falls as vocabulary grows
MS-ASL [6]	Over 25,000 annotated videos, 1,000 signs, more than 200 signers	Signer-independent testing
How2Sign [7]	80+ hours of continuous ASL with English transcripts, depth and body, face and hand keypoints	Continuous recognition and keypoint-based models

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

RWTH-PHOENIX-Weather 2014T [15]	Over 25,000 annotated videos, 1,000 signs, more than 200 signers	Sign language translation; gloss-based and end-to-end baselines
Chinese Sign Language set [18]	25,000 videos, 50 signers, 178-word vocabulary	Continuous recognition without temporal segmentation
ChaLearn 2014 gestures [16]	20 Italian gestures, 27 users; Kinect colour, depth and skeleton	Gesture spotting and CNN-based recognition
NVIDIA gesture set [14]	25 gesture classes, 1,532 gestures, 20 subjects; colour, depth and infrared	Online detection and early classification

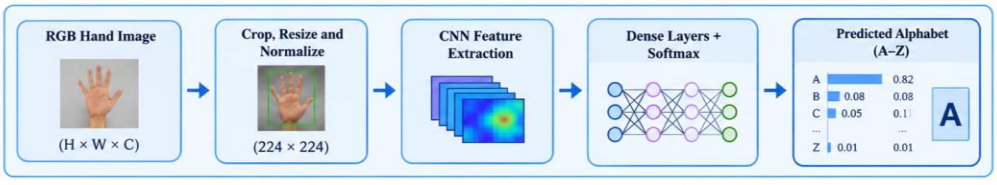
2.2 Hand Landmarks / Pose Recognition

The MediaPipe technology can be applied for the real-time detection of the hand landmarks using the camera-based lightweight model. The 21 landmarks can be further applied for the computation of measures including joint position, joint angle, and distance between specific points. OpenPose hand model provides the opportunity for detecting all body, hand, and face key-points. At the same time, the MediaPipe holistic model takes this concept one step further by providing all three key-point detections in order to create an integral motion tracking system. Consequently, the detected pose landmarks can be applied for the generation of an extended number of measures related to the motion tracking. In comparison with video frames detection, such a technique requires less computational power.

Algorithms 1: Preprocess and classify sign language images

Input: RGB hand image frame (H * W* C)
Output: Predicted alphabet (A-Z)
1.Read the input from a webcam or dataset.
2.Detect and crop the hand region.
3.Resize the images to 224*224 pixels.
4.Normalize pixel values to the range [0,1].
5.Pass the image through convolution and max-pooling layers.
6.Flatten features maps and pass them through dense layers.
7.Apply softmax to calculate probabilities for 26 alphabet classes.
8.Return the alphabet with the highest probability.

Figure 1. Basic workflow of the AI-based sign language alphabet recognition system.



2.3 Deep Learning-Based Recognition

Convolutional Neural Networks are mostly applied for feature extraction from images. They are capable of identifying patterns in the images. However, Long Short-Term Memory and Recurrent Neural Networks are used in analyzing time-dependent data. They have the ability to analyze temporal data better. 3D

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Convolutional Neural Networks are mostly an extension of Convolutional Neural Networks. They can be applied to videos because besides being able to extract spatial features, they can extract temporal features too. Transformers are mostly applied to analyze sequential data. The most beneficial aspect about Transformers is the fact that they can analyze the dependency between different elements of a sequence. By combining Hidden Markov Models and Convolutional Neural Networks, spatio-temporal feature extraction can be done and better predictions can be achieved. Spatio-Temporal Graph Convolutional Networks are used for analyzing spatio-temporal data.

2.4 Datasets

The total number of videos in the project is 21,083. Within a bigger database, the project consists of 25,000 videos. The project provides for involvement of 119 signers and the project includes 50 signers who have contributed their knowledge. The project is made up of linguistic elements, with the project consisting of 1,000 signs and 178 words with the project having a database which contains 67,000 unique signs. The project includes types of data: the project combines videos, gloss and spoken language in order for the project to enable understanding. The project takes 80 hours of content. The total number of sentences used in the project is 35,000. The project incorporates 120,000 images within 24 alphabet sets with the project having all images taken in 0.95 million frames. The project even incorporates landmark animation sequences which facilitate better learning and the project shows evidence of meeting the objective of providing an open learning environment.

2.5 Sensor-Based Recognition

Introducing Inertial Measurement Unit (IMU) and surface Electromyography (sEMG) to the wearable system is an intriguing idea. The technology allows for tracking movements and muscle movements with flexible finger sensors. The obtained data is wirelessly transmitted to smartphones via Bluetooth, enabling instant recognition on cell phones. Nevertheless, this technology requires advanced hardware components, which may make the use of such devices uncomfortable and challenging for the user. Moreover, in comparison with camera-based solutions, such systems may be less available and convenient to use.

Table 3 Existing approaches and their limitations

Input	Content	Purpose
MediaPipe Hands [2]	Palm detector and 21-landmark hand model from one RGB camera	Real-time hand tracking without special hardware
MediaPipe framework [9]	Pipeline built as a graph of reusable components	Running landmark extraction in a real-time pipeline
OpenPose [4]	Body, hand and face keypoints using Part Affinity Fields	Pose input for skeleton-based models
Multiview bootstrapping [13]	21-keypoint hand detector trained with labels triangulated from multiple cameras	Dealing with occluded hand joints
ST-GCN [5]	Graph convolution over joints within and across frames	Skeleton-based temporal modelling
IMU and sEMG sensors [19]	Wrist inertial unit and four forearm EMG channels; 80 ASL signs	Camera-free recognition of similar-motion signs

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Strain-sensor glove [20]	Stretchable finger sensors with PCA and SVM; 11 gestures	Low-cost wearable sign-to-speech output

3. EXISTING LIMITATIONS / RESEARCH GAP

Recognition and translation are two fundamentally different tasks in the context of sign language recognition (SLR), with the latter being highly challenging due to the complexity of linguistic structure. While the tracking of the performer €™s hands is undoubtedly an important step towards developing sign language models, it does not solve the problems of comprehensive understanding or learning. For words with larger vocabularies, recognition becomes more complicated by the fact that considerable datasets have to be created which would reflect the wide range of idiosyncrasies of different users. Pose estimation provides certain key points essential for recognition, but it does not necessarily engage with the multifaceted nature of the problem. Similarly, skeletal approaches require finding spatiotemporal regularities in a skeleton which general action recognition does not have to address since it ultimately does not learn to recognize sign language. Evaluation has to be independent of the user since it is reasonable to suggest that they tend to perform signs differently. Continuous and multimodal recognition is particularly challenging since it requires substantial datasets to be built, while non-manual elements and continuous recognition pose additional difficulties. What the paper offers is a perceptual infrastructure to recognize signs rather than learn sign languages: comprehensive identification requires multiple underlying modalities to be taken into account, as well as the different grammatical structure compared to traditional language. Moreover, the use of external hardware such as Kinect introduces a question of whether such technology can be portable and widespread. Continuous recognition in sign language requires accurate alignment and segmentation of video frames, which can be particularly taxing on the processor. Recognition is often hindered by such factors as hands being too small and occluded, inconsistencies in terms of illumination, and an absence of facial expressions. With the constraints on the size of the vocabulary and the number of users, recognition tends to lag significantly when it comes to different subjects as compared to the same user. Recognition of continuous sentences is particularly challenging, especially with smaller vocabularies and a limited set of users. Finally, there are additional hindrances such as the orientation of hand landmarks, processing of subsequent signs which involve heavy computations, and gestures which require multiple limbs and faces to be taken into account at once. Even though CNN-based computations are a viable way of improving accuracy, they add to the computational burden. As a result, considerations have to be made regarding processing time which is particularly important for real-time recognition. Feature fusion is a potential way of improving information aggregation.

Table 4 Technical components and options used in the reviewed studies

Approach	Content	Limitation
HMMs with hand-crafted features [12], [8]	Lexical signs recognised from hand-crafted features; non-manual signals rarely modelled	Misses grammar and facial expression; hard to scale
Depth cameras [2]	Earlier hand tracking relied mostly on depth sensors	Needs special hardware
Wearable sensors and gloves [19], [20]	IMU and sEMG on 80 signs; strain-sensor glove on 11 gestures	Per-user training; small gesture sets

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Large video models [3], [22]	I3D on WLASL; 3D CNN reaching 92.4% in a comparison with LSTM	Needs large labelled data and a GPU; accuracy falls with vocabulary
Landmark and LSTM systems [21], [24], [25]	MediaPipe landmarks with compact recurrent or CNN models	Few signers; weak on two-handed signs; faces ignored
Continuous recognition and translation [15], [1]	Gloss supervision raises BLEU-4 from 9.58 to 18.13; transformer reaches 21.80	Needs gloss annotation; one large continuous benchmark [8]

4. PROBLEM STATEMENT AND OBJECTIVES

Real-time recognition of signs is still a challenging task, and one of the reasons for this is dynamic gestures requiring temporal modelling. Information from only one part of the body can be insufficient, and factors like lighting and occlusion can interfere with recognition. Furthermore, different people often perform signs differently, and this is especially evident when it comes to two-hand signs. One of the critical points I find is the need for a diverse dataset, as it addresses both the signer-independent recognition problem and the issue of large vocabulary recognition, both of which are serious challenges. Additionally, continuous signing adds complexity to the recognition process. While using wearables and sensors could enable high recognition performance, such solutions can be limited in terms of convenience and comfort due to their nature. On the other hand, using camera- and landmark-based solutions can help in enabling real-time processing with the help of tools such as MediaPipe being suitable for such a solution by enabling relatively light processing.

5. SYSTEM OVERVIEW

The use of webcams or RGB cameras is mentioned in papers 2, 4, 21, 24, and 25. The term MediaPipe is widely used in the research papers 2, 9, 21, 23, 24, and 25. In paper 4, it mentions the OpenPose technique. Meanwhile, the determination of the hand landmark is described in papers 2, 13, 21, 24, and 25. Furthermore, the calculation of the joint angle and distance is discussed in paper 24. However, spatial features are elaborated in detail in papers 16, 21, and 25. In addition, temporal features are identified in papers 14, 17, 21, 22, and 24. Some of these papers use CNN as the architecture, and it is mentioned in 16, 21, 25, and also 23. LSTMs are discussed in papers 21, 22, 23, 24. 3D-CNN is one of the architectures used by the researchers in this study, particularly in papers 14, 18, 22. In paper 1, the Transformer architecture is utilized, while there is a study, paper 5, that explores the ST-GCN architecture. The recognition of the signs and calculating the confidence level and predicted probabilities are also discussed in these research papers. Some of these mention the translation of the sign language, and several of them talk about the sign language recognition field.

6. SYSTEM ARCHITECTURE AND COMPONENTS

Input Layer

A camera capable of recording RGB videos with hand tracking capabilities enables the acquisition of images with a high amount of detail on the RGB spectrum in combination with precise pose estimation. By utilizing a common webcam as an underlying hardware foundation, applications using the combination of RGB-based 3D computer vision and hand tracking can become widely accessible and viable.

Landmark Extraction Layer

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

MediaPipe perception pipeline provides holistic tracking of hands and bodies using 543 landmarks in total. By having a total of 21 key points for hands, MediaPipe manages to identify hand movements quite successfully. The process involves MediaPipe landmark detection using advanced computer vision. The 21 points for each hand are used to detect and track hand movement, while the other 522 key points are used for bodies. All of this makes MediaPipe one of the most advanced holistic action detection systems out there.

Feature Processing Layer

The analysis of human motion can be improved by using graph-based skeleton relations, allowing to represent the spatial configuration of the human body with key points. Moreover, applying 3D hand keypoints, it becomes possible to describe the location of each hand part fully in three dimensions. Another advantage of graph-based representations is the opportunity to consider normalized landmark vectors to evaluate movements. In addition, it is possible to use joint, bone, and motion representations, considering coordinates, angles, and distances. All these features help to build a detailed motion description model, suitable for representing each detail of movement.

Recognition Layer

In the context of advanced machine learning and deep learning techniques, various models were created to operate and process data. Convolutional Neural Networks (CNN) were the basic concept that later improved and developed into other models such as Long Short-Term Memory (LSTM). Additionally, 3D-CNN is an evolution of the original CNN, made to process diverse types of data, and it is currently used in video processing and medical imaging. With the introduction of the Transformer model, processing natural language became efficient and could be done while considering context. Moreover, CNN can be combined with other concepts such as Hidden Markov Models resulting in CNN-HMM, which showcases versatility. This model is applied in sequence prediction problems. Spatio-Temporal Graph Convolutional Networks (ST-GCN) is another model that demonstrates how convolutional networks can be used in distinct fields and processes. In short, these are the different models of machine learning and deep learning that currently exist.

Output Layer

The process of recognition and translation includes several elements, one of which is the sign-to-text translation. The latter enables a user to transform the recognized sign into text. Moreover, the system employs predicted signs and measures of confidence, which allows it to realize what signed word is intended by the user. Finally, by utilizing real-time recognition and translation, the system provides the ability of instant communication.

Algorithm

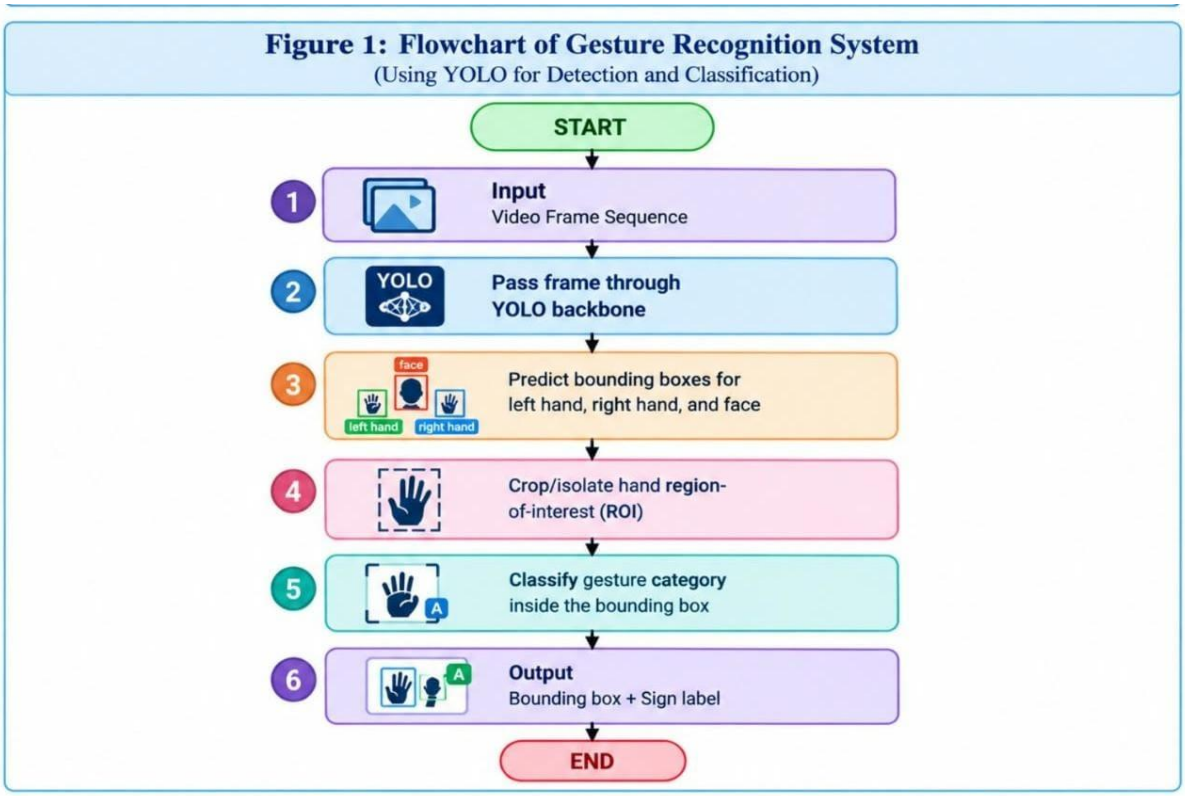
INPUT: video Frame Sequence

1. Pass frame through YOLO backbone
2. Predict bounding boxes for left hand, right hand, and face
3. Crop/isolate hand region-of-interest(ROI)
4. Classify gesture category inside the bounding box

OUTPUT: Bonding box + sign label

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026



7. TECHNICAL IMPLEMENTATION

MediaPipe

The application of the newest computer vision technologies to recognize image landmarks in real-time with a camera is an extremely promising direction. Such a beneficial outcome can be achieved due to the use of MediaPipe models, which allow extracting hand, body, and face landmarks. As a result, there will be more opportunities for developers to expand the capabilities of their applications. In addition, this approach is helpful when the use of specific sensors is limited, which makes the technology applicable in various fields. Therefore, the described solution is appropriate for implementing real-time recognition tasks.

CNN

Spatial feature extraction is one of the most significant approaches used in image or video-based classification, such as the case with static gesture recognition. This method is particularly effective in recognizing and classifying gestures. At the same time, a notable downside is that depending on the chosen architecture, this approach may be resource-heavy, which makes it relatively slow.

LSTM

Temporal sequence modelling is an effective method, which is appropriate for dynamic signs.

Table 5 Reported performance of the reviewed systems

Component	Options reported	Importance
Capture	Webcam at 30 FPS [21]; Kinect RGB-D [16]; wearable sensors [19]	Webcam needs no extra hardware

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Landmark extraction	MediaPipe Hands and Holistic [2], [9]; OpenPose [4]	Holistic gives 543 landmarks per frame [21]
Normalisation	Wrist as local origin [25]; scaling by distance between extreme landmarks [21]	Result does not depend on where the signer stands
Features	Coordinates plus 41 angles and 26 distances [24]; bone and motion vectors [23]	Compact description that helps separate similar signs
Temporal model	LSTM [21], [24]; 3D CNN [22]; 3D CNN with CTC [14]; ST-GCN [5]	Captures how a sign changes over time
Training	Adam optimiser, dropout, data augmentation [16], [21]	Reduces overfitting; validation error fell from 18.9% to 8.3% in [16]
Output	On-screen text; speech through a phone app [20]	Lets a hearing person follow the sign

REFERENCES

[1] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, arXiv:2003.13830.
[2] F. Zhang et al., "MediaPipe Hands: On-device real-time hand tracking," arXiv:2006.10214, 2020.
[3] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2020, arXiv:1910.11006.
[4] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," IEEE Trans. Pattern Anal. Mach. Intell., arXiv:1812.08008, 2019.
[5] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in Proc. AAAI Conf. Artif. Intell., 2018.
[6] H. R. Vaezi Joze and O. Koller, "MS-ASL: A large-scale data set and benchmark for understanding American Sign Language," in Proc. British Mach. Vis. Conf. (BMVC), 2019.
[7] A. Duarte et al., "How2Sign: A large-scale multimodal dataset for continuous American Sign Language," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, arXiv:2008.08143.
[8] O. Koller, "Quantitative survey of the state of the art in sign language recognition," arXiv:2008.09918, 2020.
[9] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," arXiv:1906.08172, 2019.
[10] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," Expert Syst. Appl., Art. no. 113794, 2020, doi: 10.1016/j.eswa.2020.113794.
[11] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," Int. J. Mach. Learn. Cybern., 2017, doi: 10.1007/s13042-017-0705-5.
[12] S. C. W. Ong and S. Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 6, pp. 873–891, Jun. 2005.
[13] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017.
[14] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 4207–4215.
[15] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7784–7793.
[16] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in Proc. ECCV Workshops, 2014.
[17] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep Sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs," Int. J. Comput. Vis., 2018, doi: 10.1007/s11263-018-1121-3.
[18] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in Proc. AAAI Conf. Artif. Intell., 2018.
[19] J. Wu, L. Sun, and R. Jafari, "A wearable system for recognizing American Sign Language in real-time using IMU and surface EMG sensors," IEEE J. Biomed. Health Inform., vol. 20, no. 5, pp. 1281–1290, Sep. 2016.
[20] Z. Zhou et al., "Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays," Nat. Electron., 2020, doi: 10.1038/s41928-020-0428-6.
[21] S. Kamble, "SLRNet: A real-time LSTM-based sign language recognition system," arXiv:2506.11154, 2025.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

[22] A. Anturkar et al., “Real-time sign language to text translation using deep learning: A comparative study of LSTM and 3D CNN,” Int. J. Comput. Appl., vol. 187, no. 55, pp. 31–35, Nov. 2025.
[23] N. P. Thanh et al., “Exploring MediaPipe optimization strategies for real-time sign language recognition,” CTU J. Innov. Sustain. Dev., vol. 15, special issue ISDS, pp. 142–152, 2023, doi: 10.22144/ctujoisd.2023.045.
[24] Y. Farhan, Z. Haimer, O. Lamaamar, and A. Ait Madi, “Dynamic sign language recognition: A hybrid approach combining MediaPipe and LSTM,” CommIT J., vol. 20, no. 2, pp. 363–377, 2026.
[25] P. More, S. Koli, M. Khandekar, P. Arsh, and D. Ahmad, “A MediaPipe and CNN-based framework for real-time sign language recognition and translation,” ITM Web Conf., vol. 82, Art. no. 03024, 2026, doi: 10.1051/itmconf/20268203024.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)