



International Journal of Applied Smart Interdisciplinary Technologies (IJASIT)

Home Page: <https://ijasit.org>

ISSN: 3139-759X

<https://doi.org/10.68104/ijasit.v1.i3.64>

Vision-Driven framework For Interpreting Human Gestures into Natural Language

Karthikeyan V¹, Hemlathadhevi A², Jackulin C³, Dhanush Abhishek J⁴, Balamurugan P⁵, Jabasheela L⁶

^{1 2 3 4 5 6}

Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, Tamil Nadu, India

* Corresponding author.

E-mail: author@gmail.com

ABSTRACT

Article History:

Received Date: 06 August 2026

Revised Date: 20 August 2026

Accepted Date: 28 August 2026

Published Date: 30 August 2026

Keywords:

Sign Language Recognition; Computer Vision; Gesture Recognition; Deep Learning; Human-Computer Interaction; Assistive Technology; Accessibility

Communication limitations between sign language and the hearing community have remained an impediment for inclusion and access of deaf and hard-of-hearing persons into vital and basic life requirements. Towards overcoming this major challenge, this paper advances a vision-based framework for interpreting and processing natural language from human gestures using computer vision and machine learning approaches. The proposed framework, chiefly, employs camera devices in capturing real-time gestures and understanding hand, body, and facial movements. Efficient techniques have been incorporated, such as detecting gestures and landmarks using OpenCV and MediaPipe, and further analyzing and processing gestures with landmark detection techniques and machine learning-based classification using TensorFlow. Experimental results indicate the accuracy and reliability of a proposed system for vision-based natural language translation. Thus, the proposed vision-based framework can be a potential and inclusive technology in addressing the challenges and limitations in natural language understanding and accessibility, towards promoting inclusive and equitable societies by ensuring accessibility for all, thus promoting and sustaining inclusive growth and social stability.

1. Introduction

The role of human communication is critical to social interactions, learning processes, medical attention, and access to critical services. For instance, communication among deaf and hard-of-hearing persons is mostly dependent on sign language. However, this language is rarely comprehended by hearing persons. The communication gap has remained an impediment to accessibility and inclusiveness for hearing persons. Despite this, communication can be facilitated by Identify applicable funding agency here. If none, delete this. Interpreters [1]. However, interpreters are few and come with a high cost. The communication gap has created an urgent demand for sign language recognition systems that can facilitate communication between sign language users and non-signers. With the help of computer vision and deep learning technologies, considerable advances have been achieved in the field of gesture recognition technology [2]. Existing technologies, including sensor-based and glove-based technologies, have shown promising accuracy in the recognition of hand gestures. However, these technologies have several drawbacks, including the requirement of special hardware and the imposition of physical constraints. In contrast, vision-based technologies have proven to be a promising approach to the recognition of natural human gestures using conventional camera devices. Sign language

communication is a complex and multimodal communication system that includes not only hand gestures but also other components such as body posture and facial expressions. Most of the existing technologies have concentrated only on the recognition of hand gestures, which limits their ability to recognize complex and continuous sign language communication. Therefore, a comprehensive framework is required to recognize different components of the gestures [3].

In this paper, a vision-driven system for interpreting human gestures in natural language will be proposed, utilizing computer vision and deep learning approaches. The proposed system will be capable of capturing video feed in real-time using conventional RGB cameras and will be able to detect hand, body, and facial landmarks using landmark-based detection. The detected landmarks will then be classified using a deep learning model and a natural language processing component to produce meaningful texts and speech [4].

2. Literature Review

In recent studies on sign language translation and recognition, scientists have used various techniques to improve the communication gap between the deaf or hard-of-hearing population and the hearing population. Various researchers have proposed sensor-based approaches or glove-based approaches, which showed remarkable accuracy in hand movement detection but were not completely effective due to limitations such as high costs, hardware dependency, etc. For example, smart glove-based approaches have been used in sign language translation, utilizing various machine learning techniques with sensor-based approaches, but the smart glove-based approach restricts the usage of such approaches in real-time environments due to hardware limitations [5][6].

However, recent survey-based research has also shown the issues with generalization, the lack of support for various languages, and insufficient bodily expression recognition assistance. IoT-based translators for specific languages, such as Indian Sign Language, have also been proposed, though these work with specific hardware and have shown poor flexibility and adaptability. The proposed work has focused on overcoming these issues by using a vision-driven approach, including hand, body, and facial expression detection followed by deep learning-based gestures and natural language translation [7]. This work has eliminated the need for specific hardware and has provided high flexibility, thus promoting inclusive human-computer interaction. In the research paper titled "JUMLA-QSL-22: A Novel Qatari Sign Language Continuous Dataset (2023)," a large-scale continuous sign language dataset was proposed, which includes hand, facial, and body gestures. However, the research paper that created the dataset focused on creating the dataset, which can be used for advanced gesture modeling. A real-time translation approach from gestures to language is not addressed in this paper." An Efficient Two-Stream Network for Isolated Sign Language Recognition (2022)" was a paper that described a deep learning-based model that learned spatial and temporal features for isolated signs. However, even though the model showed good recognition results, it is still an isolated sign recognition technique, unable to deal with other signs, sequences of signs, or language translation [8].

The authors, in the paper "Vision-Based Continuous Sign Language Spotting Using Hidden Markov Models (2022)," have specifically focused on the identification of sign language using probabilistic modeling techniques [9]. This method has proved promising for sign language segmentation but has only limited potential for gesture spotting and does not include gesture recognition or language conversion." Vision-Based Continuous Sign Language Spotting Using Hidden Markov Models" (2022) has concentrated on probabilistic modeling techniques that find boundaries of signs within continuous video streams. Though very effective in sign segmentation, the approach remains limited to gesture spotting and does not really perform full gesture recognition or semantic language conversion. The work "Smart IoT System for Indian Sign Language Translator" in the year 2024 presented an IoT-based framework for translating Indian Sign Language gestures. Considering its relevance to regional accessibility, this system requires the use of specific hardware devices. Thus, this reduces scalability and increases costs in public and large-scale deployments of this system. The survey paper "Sign Language Recognition and Translation Systems for Enhanced Communication" reviews methodologies, datasets, and challenges from the research paper in 2023. While comprehensive, the study does not propose or validate the practical implementation desirably; it thus leaves room for real-time, end-to-end systems [10].

This problem was further addressed in "Research on Sensor-Integrated Smart Gloves for Sign Language Recognition (2025)," in which researchers emphasized the union of machine learning algorithms and smart

gloves in the recognition of gestures. Although the technique was successful in the recognition of finger movement, the problem that arises is that the technique is dependent on the hardware and that the movements are recognized when they are made individually. Considering the limitations identified in the above studies, the present research proposes the following vision-based integrated solutions: hand, body, and face region detection, along with gesture classification by deep learning and natural language translation in real time. With the elimination of wearable devices, the present research acknowledges the role of inclusivity in overcoming the limitations of research in the stated area.

3. System Overview and Architecture

The proposed system is explained to be a vision-based system for gestural interpretation that is capable of converting the signs and language from the human gestural system into text and natural language sounds in real time. Specifically, the system is explained to run based on the acquisition of various video devices, which is achieved through the provision of camera devices that function in normal manners, providing a hardware-free system. In the input step, live video feeds are constantly captured, and frames are constructed from these videos for processing. In the frames, computer vision approaches are used to recognize hand, body, and facial landmarks. The landmarks correspond to significant parts of a gesture, including finger joint positions, hand orientation, body posture, and facial expressions, which cumulatively define a visual representation.

The identified landmarks are then passed to the feature extraction phase, in which spatial and temporal features are computed. The spatial features define the relationship between landmarks, while temporal features define the motion dynamics between successive frames. This allows for accurate modeling of both isolated and continuous gestures. The use of machine or deep learning for training on sign language data is achieved in the classification module of gestures, whereby extracted features are recognized. The proposed system also includes support for continuous gestures, whereby there is consistency in the transition of gestures. Once gestures are recognized, the language processing module maps classified gestures to corresponding natural language tokens and constructs meaningful sentences using predefined linguistic rules. The final output is generated in two formats: text, which is displayed on the user interface, and audio, produced through a text-to-speech engine to facilitate real-time communication. Overall, the proposed system provides an efficient, scalable, and accessible solution for sign language interpretation, addressing limitations of existing hardware-dependent and isolated-gesture-based approaches while promoting inclusive human-computer interaction.

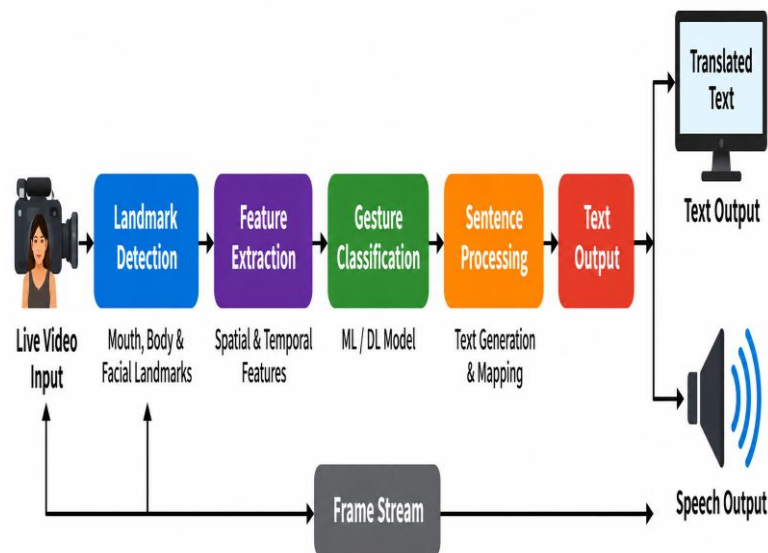


Fig. 1. System Overview

3.1 Gesture Acquisition Layer

This layer utilizes a conventional RGB camera to capture real-time video frames comprising human gestures, posture, and facial movements. In this regard, the OpenCV library is employed to effectively interact

with the aforementioned devices to capture video frames at an appropriate frame rate. It is worth mentioning that using conventional devices such as cameras enables the proposed system to achieve uniform deployment within a variety of environments, such as classrooms and service centers.

3.2 Preprocessing Layer

The pre-processing layer involves cleaning up the video frames captured from various environmental conditions to a decent quality. Preprocessing will involve a number of operations like resizing of frames, noise removal, and normalization that can address swings in lighting conditions, background complexity, and camera resolution. This layer ensures consistent input quality, which is essential for reliable landmark detection and gesture recognition.

3.3 Landmark Detection Layer

The landmark detection layer is the principal aspect of the visual analysis functionality. In this layer, MediaPipe is used to identify and track various important landmarks of hands, bodies, and faces, respectively, in a video frame. These include finger joints, palm orientation, upper body pose, and facial expression points, among others. However, the extracted landmarks are stored as three-dimensional spatial coordinates, thus enabling the precise modeling of both standing and moving gesture aspects accurately. The addition of hand, body, and facial landmarks means that all elements of sign language gestures are obtained accurately.

3.4 Feature Extraction and Normalization Layer

From the detected landmarks in this layer, meaningful features are derived. Joint distances along with angular relationships in space and motion between the frames over some times are computed. Normalization of features is performed to make them invariant under scaling and rotation, thus making the system robust across users and camera viewpoints. The structured feature representation eliminates redundancy and allows for higher classification efficiency.

3.5 Gesture Classification Layer

Gesture classification layer uses a supervised deep learning model built using TensorFlow. Normalized feature vectors are fed into the trained model, which makes the mapping of gesture patterns to previously defined sign language classes. Architecture supports multilingual sign language recognition, and easy extension of languages like American Sign Language (ASL), Indian Sign Language (ISL), British Sign Language (BSL) is possible. This modular classification design enhances adaptability and future scalability.

3.6 Natural Language Processing and Output Layer

As soon as the gesture is recognized, the corresponding semantic meaning is generated in natural language form. The gesture label recognized from the gestural input has been mapped to text output, and this output will be displayed on the user interface in real time. In addition to this, a text-to-speech subsystem is also used to produce speech output from the natural language sentences that have been generated for effective communication between the SLS user and non-signers.

3.7 Deployment and Execution Layers

Finally, the last level is responsible for system execution and deployment. This framework is based on Python and is able to run in both local and cloud environments. This allows the system to run as a standalone system, as well as part of many other applications like mobile apps, web applications, and assistive communication systems. This creates room for its maintenance and further upgrades.

The proposed system utilizes a vision-driven system for the real-time recognition and generation of sign language. The system framework includes gesture acquisition using an ordinary RGB camera, which captures the video stream from the live feed or recorded media. The acquired frames are preprocessed to improve the system's performance in diverse environmental conditions. Landmark detection methods are employed to extract the key points from the hand, body, and face, which are arranged in a structured manner to represent the gesture. The extracted features are fed into a classification module using a deep learning algorithm to classify the gesture. The classified gesture is converted into text and speech using the natural language processing and speech synthesis.

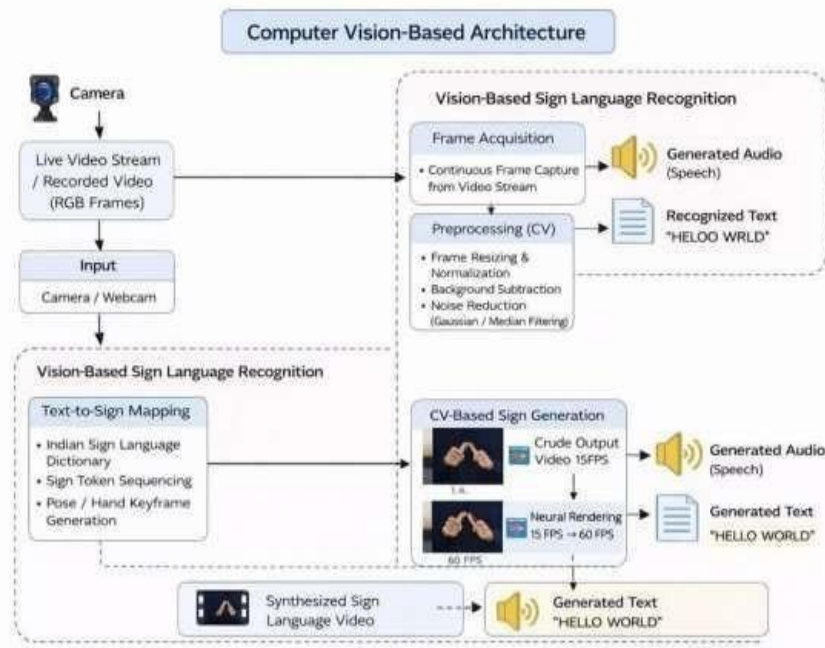


Fig. 2. System Architecture

4. Methodology

The process begins with the acquisition of videos in real-time, facilitated by a camera interface. Frames acquired are constantly processed frame by frame, facilitated by the OpenCV library. Frames undergo necessary preprocessing, which includes resizing, noise reduction, and normalization, to ensure robustness in different conditions, particularly in environments with changing lighting conditions or different backgrounds. Following this step, landmark detection is performed by utilizing the MediaPipe system. MediaPipe is used to detect precise spatial coordinates of specific hand landmarks (finger joints and palm orientation), body pose landmarks (upper body posture), and facial landmarks (expressions and mouth movements). The integration of hand, body, and facial landmarks is crucial to provide a wide range of gestures, as sign language incorporates various movements rather than hand gestures alone.

The detected landmarks then undergo a feature extraction process to transform raw feature data into discriminative and compact feature representations. Information that can be extracted as features includes the relative distance between two or more joints, the angle formed by two or more fingers or limbs, and the pattern of motion from one frame to another. To obtain a measure of invariance as regards scales, rotations, and translations, the features are normalized to ensure that the system can function appropriately despite the presence of noise and other differences among users and camera positions. The normalized feature vectors thus obtained are input to the gesture classification model, and the model was built employing TensorFlow. It was trained following the supervised learning approach, specifically to map the various gesture patterns to specific sign language labels.

The architecture of the system provides scope for multi-lingual sign language recognition, including ASL, ISL, and BSL, due to the gesture-label mapping mechanism, which addresses the major limitation of the state of the art. Typically, the systems available only recognize a single sign and, moreover, lack adaptability. Subsequent to gesture recognition, the recognized sign language word or phrase is converted into its corresponding semantic meaning in natural language form. The natural language text output derives from the constructed module and is shown on the user interface in real time. Furthermore, the system also includes a text-to-speech module, allowing the generated text to be converted into sound, and facilitating communication among the deaf and hard of hearing and hearing people.

The entire framework is implemented in Python using OpenCV, MediaPipe, TensorFlow, and NumPy. The system supports both local execution and cloud-based deployment, allowing future scalability and integration into larger platforms. Performance evaluation is conducted using metrics such as recognition accuracy,

response latency, and user feedback on ease of use, validating the effectiveness and practicality of the proposed approach.



Fig. 3. Landmark Detection

5. Experimental Evaluation

The experimental execution of the presented vision-driven gesture recognition framework was carried out to assess the efficacy of this process of translating human gestures to natural language text and speech. The experiments were carried out within a controlled indoor environment using a regular RGB camera to process video input in real time.

The entire system was developed using Python programming language along with OpenCV, MediaPipe, and TensorFlow libraries. During the experiment, continuous video frames were captured and preprocessed to remove noise, as well as to unify the dimension of the input data. MediaPipe was used to detect the landmarks of hands, bodies, and faces within a video frame and acquire three-dimensional coordinates from the captured video frames. After that, spatial-temporal features that incorporate distance between joints and angle between them were extracted and normalized to unify various users and gestures.

Table I

Performance Evaluation Results

Training Epoch	Accuracy	Gesture	Response Latency
Epoch 1	72%	Hello	120 ms
Epoch 3	78%	Thank You	135 ms
Epoch 4	84%	Yes	110 ms
Epoch 5	87%	No	105 ms
Epoch 7	91%	Help	140 ms
Epoch 10	94%	–	–

These normalized feature vectors were then used as a form of input to the supervised deep learning model with predefined sign gesture patterns. These patterns were used as classifications in real-time, generating labels that corresponded with each of the gestures. These labels were then used to create natural language output, which was then displayed in the user interface. Furthermore, there was also the incorporation of text-to-speech, which enabled real-time communication through audio. To validate these experiments, different experiments

regarding system performance have been conducted. These experiments involved several samples of gestures and different users. The experiments resulted in validating the proposed system, which can be used to recognize various hand, body, and facial gestures, converting these gestures into meaningful text or speech.

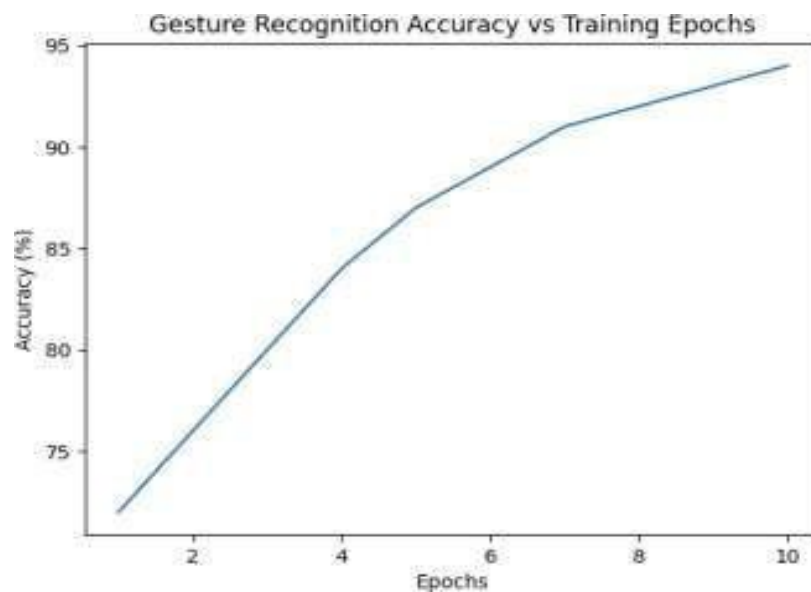


Fig. 4. Accuracy vs Training Epochs

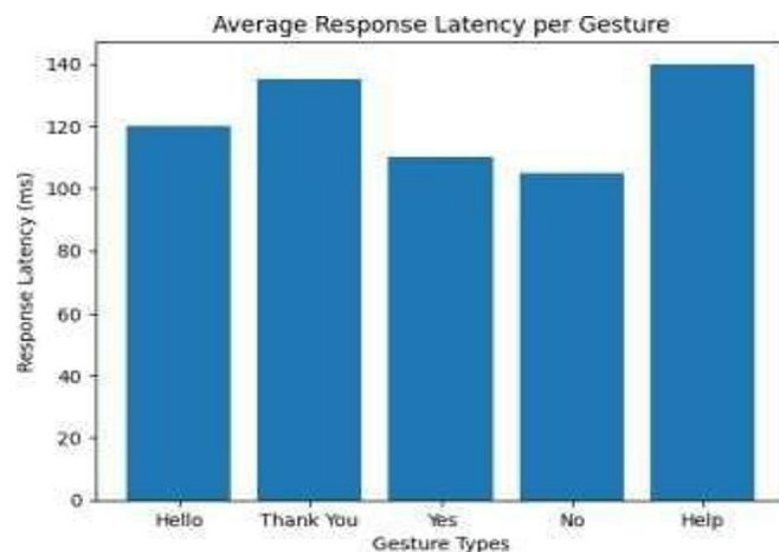


Fig. 5. Response Latency per Gesture

$G = f(X)$, where G is the Recognized gesture class, f is the trained deep learning model, and x is the normalized landmark feature vector.

6. Conclusion And Future Works

This paper introduced the vision-driven framework for the interpretation of human gestures into natural language by employing computer vision and deep learning techniques. In this context, the proposed framework is facilitated by the extraction of human hand, body, and facial key points through the use of landmark-based MediaPipe, eliminating the necessity for the use of wearables and specialized hardware to extract such information. The framework is able to achieve such an objective by relying on the extraction of the relevant

features while employing the sign language recognition capabilities of the suggested framework, thereby facilitating the translation of the identified signs into relevant text and speech forms. Experimental results validate the efficacy of the suggested framework to achieve high accuracy in the context of low latency, thereby facilitating the translation of sign language into machine-readable formats while promoting inclusivity in the context of hearing-impaired individuals with the general populace to promote inclusivity in society. Future enhancements of the proposed system would be to enhance its ability to withstand challenges related to its robustness, scalability, and language variety. The system would be improved to recognize sentences in sign language on a continuous basis using temporal deep learning architectures such as CNN-LSTM and CNN-Transformer. The variety of gestures would also be increased to cover other signs, users, and environmental conditions. The system would also be enhanced to include multilingual natural language processing modules to enable it to translate to more than one language. Future plans would also include increasing the system's ability to be implemented on a cloud and mobile platforms to enhance scalability. Future plans would also include incorporating emotion recognition and contextual understanding as well as developing a prototype to enable accessibility and usability.

References

- [1]. A. Al-Khazraji, M. Al-Ali, and S. Al-Maadeed, "JUMLA-QSL-22: A Novel Qatari Sign Language Continuous Dataset," in Proceedings of the 2023 International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2023, pp. 1–6.
- [2]. Vision-Base Y. Zhou, H. Shi, and L. Wang, "An Efficient Two-Stream Network for Isolated Sign Language Recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022, pp. 1234–1242. d Continuous Sign Language Spotting Using HMM (2022)
- [3]. A. K. Talukdar and M. K. Bhuyan, "Vision-Based Continuous Sign Language Spotting Using Gaussian Hidden Markov Model," IEEE Sensors Letters, vol. 6, no.7, pp. 1–4, July 2022.
- [4]. R. Singh, P. Kumar, and S. Verma, "Smart IoT System for Indian Sign Language Translator," in Proceedings of the International Conference on Internet of Things and Smart Systems (IoTSS), 2024, pp. 1–6
- [5]. M. Ahmed, S. Khan, and R. Ali, "Sign Language Recognition and Translation Systems for Enhanced Communication," in Proc. Int. Conf. Artificial Intelligence and Communication Technologies (AICT), 2023, pp. 1–6.
- [6]. C. Liang, C. Wu, and S. Shimamoto, "Research on Sensor-Integrated Smart Gloves for Sign Language Recognition," in Proc. Asia-Pacific Conf. Communications (APCC), 2025, pp. 1–5.
- [7]. M. S. Amin, S. T. H. Rizvi, and M. M. Hossain, "A Comparative Review on Applications of Different Sensors for Sign Language Recognition," J. Imaging, vol. 8, no. 4, art. 98, 2022.
- [8]. N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Neural Sign Language Translation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 7784–7793.
- [9]. K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2014, pp. 568–576.
- [10]. A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2006, pp. 369–376.