This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier

Deep Learning Approaches for Continuous Sign Language Recognition: A Comprehensive Review

ASMA KHAN[1], SEYONG JIN[2], GEON-HEE LEE[2], GUL E ARZU[1], TAN N. NGUYEN[3], L. MINH DANG[4], WOONG CHOI[5], and HYEONJOON MOON[1]

[1]Department of Computer Science and Engineering, Sejong University, Seoul 05006, Republic of Korea (e-mail: Asmakhan28@sju.ac.kr, hmoon@sejong.ac.kr, arzurabbani12@gmail.com)
[2]Department of Artificial Intelligence, Sejong University, Seoul 05006, Republic of Korea (e-mail: jinseyong@naver.com,lsch6214@gmail.com)
[3]Department of Architectural Engineering, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea (e-mail:tnnguyen@sejong.ac.kr)
[4]Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, Republic of Korea (e-mail: minhdl@sejong.ac.kr)
[5]College of ICT Construction & Welfare Convergence, Kangnam University, Yongin-si 16979, Republic of Korea (e-mail: wchoi@kangnam.ac.kr)

Corresponding authors: Woong Choi (e-mail: wchoi@kangnam.ac.kr) and Hyeonjoon Moon (e-mail: hmoon@sejong.ac.kr).

ABSTRACT Sign language utilizing hand gestures for visual mode of communication, body actions, and facial expressions. Due to the increasing incidence of hearing deficiencies, the field of Continuous Sign Language Recognition (CSLR) has seen a considerable increase in research, which involves identifying consecutive signs in video streams without previous information of their sequential limitations. This survey reviews CSLR research, presenting deep knowledge into the development of CSLR systems. It critically analyzes numerous studies, organizing them into a comprehensive taxonomy covering aspects such as sign language, data collection, input method, gesture signals, identification methods, applied data collections, and comprehensive efficiency. The article further categorizes deep-learning CSLR models according to spatial, temporal, and alignment approaches, highlighting their benefits and drawbacks. Furthermore, it explores various research aspects, such as the challenges of CSLR, the significance of nonverbal elements in CSLR systems, and the gaps in the body of current research. This classification serves as a helpful tool for researchers developing and organizing cutting-edge CSLR methods. The study highlights the effectiveness of deep learning systems capture different sign language signals. On the other hand, several challenges remain, such as the need for diverse, naturalistic datasets, improved signer diversity, and real-time CSLR systems. Addressing these gaps will be essential for advancing CSLR's real-world applications and developing more robust, efficient models for the future. The conclusions give a wider Comprehension of sign language recognition and set the groundwork for future studies focused on addressing the current challenges and issues in this developing area.

INDEX TERMS Continuous Sign Language Recognition, deep learning, Hand Gesture Recognition (HGR), computer vision

I. INTRODUCTION

SIGN language is an important mode of communication that utilizes visual cue such as facial expressions and hand motions [1]. Interpreting these motions from video sequences and translating them into comprehensible glosses is known as sign language recognition (SLR). This method records the signer's hand and body movements, frequently capturing their facial expressions [2]. The principal aim of SLR is to enable effective communication between those who are deaf Deaf and the wider community by interpreting sign language in a comprehensible manner for others who are not familiar with it.

Continuous Sign Language Recognition (CSLR) includes two main types: isolated SLR, which focuses on recognizing individual signs from video clips, and CSLR itself, which interprets sequences of signs to produce corresponding glosses [3]. CSLR is particularly useful in real-world situations as it accurately identifies the flow of signs in natural conversations.

A primary challenge in CSLR involves interpreting sign

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

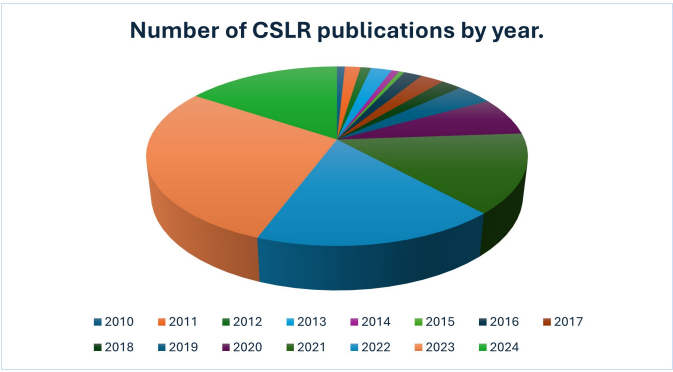


FIGURE 1. Distribution of CSLR publications by year from 2010 to 2024.

gestures within their contextual framework. Continuous sign language sentences encompass finger-spelled, static, and dynamic signs. Signs written with the fingers are often static and stationary, while dynamic signs involve hand and body movements accompanied by various non-manual signals like facial expressions. Additionally, the variability among signers in how signs are performed further complicates CSLR systems.

Weakly supervised learning faces difficulty in CSLR since it is hard to align video frames exactly with their annotations. When using sign language, Sentences in sign language are performed constantly and without boundaries. making it challenging to discern where each sign gesture begins and ends. Consequently, CSLR systems must learn to identify these boundaries from the continuous flow of sign language, which is a significant challenge due to the absence of clear delineations between signs within sentences [4].

CSLR systems follow four main stages. Initially, the input video stream undergoes preprocessing, which involves resizing and normalizing video frames. At this point, certain CSLR systems make use of pose or skeletal data, while others skip this step if they rely on sensor-based sign-capturing methods [5]. Following preprocessing, spatial and temporal features are extracted from the sequence of frames. Techniques including 2D, 3D, Graph Convolutional, and Vision Transformer neural networks (ViTs) are used in spatial feature extraction to extract feature representations from sign frames. Subsequently, temporal learning techniques like Recurrent Neural Networks (RNNs) or Temporal Convolutional Networks (TConv) are used to understand the temporal dynamics of sign gestures. The final stage involves learning the the position among the video frames and notation glosses, typically addressed using techniques as Connectionist Temporal Classification (CTC), Dynamic Time Warping (DTW), and Hidden Markov Models (HMMs). CTC is one of these techniques that has demonstrated superior results and is frequently utilized in CSLR studies for sequence alignment training. [4].

Three different protocols are used to evaluate CSLR systems: unseen sentences (Unseen-Sent), signer-independent (Signer-Indep), and signer-dependent (Signer-Dep). The same signers who were employed in training produce the sentences on which the models are tested in Signer-Dep evaluation. This approach can yield high accuracy, but there's a risk of overfitting to specific signer characteristics, which limits generalizability to new signers not included in the training set.In Signer-Indep assessment, CSLR models are trained on one set of signers and then evaluated on an additional set of signers that were not seen during training. This protocol ensures systems can generalize across diverse signing styles and appearances, which is essential for creating inclusive and adaptable CSLR systems. The Unseen-Sent evaluation assesses the system's capability to recognize sign language sentences that were not part of the training dataset. This challenging evaluation simulates real-world scenarios where CSLR systems must accurately interpret novel signs or sentences, providing crucial insights into their robustness and real-world applicability.

The majority of state-of-the-art (SOTA) models rely significantly on a limited number of benchmark datasets, even with the widespread use of CSLR techniques. These include the CSL dataset [6], which is used for Chinese sign language (CSL), and the RWTH-PHOENIX-Weather-2014 (Phoenix2014) dataset [7], which is used for German sign language (GSL). Other sign language datasets exist but are not used much in contemporary research, such as Arabic (ArSL) [8], Greek (GrSL) [9], and Russian (RSL) [10]. Continuous Sign Language Recognition (CSLR) has gained significant attention from researchers because deep learning (DL) has advanced so quickly over the past ten years. A large portion of studies in this area, about 70%, have been issued within the past seven years, indicating the increasing interest and progress in CSLR. However, CSLR systems still have significant scope for enhancement in relation to speech recognition systems. Existing CSLR systems frequently lack sufficient vocabulary and are unsuitable for use in real-time applications or for business use.

The objective of this study is to provide an extensive evaluation of the literature that highlights the advancements in CSLR and points out areas that require more investigation. The following are this review's main goals:

- Provide an in-depth understanding of CSLR, highlighting its challenges and problem description.
- Present a comprehensive summary of the datasets and techniques developed throughout the previous 20 years.
- Analyze different characteristics of CSLR, including techniques, features, input channels, and data collection approaches.
- Determine the present shortcoming in the literature, and make recommendations for prospective future studies to strengthen the validity and usefulness of CSLR frameworks.

II. CONTINUOUS SIGN LANGUAGE RECOGNITION

Continuous Sign Language Recognition (CSLR) involves identifying and understanding sign language signals performed in an unbroken stream, without pauses between signs.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

This field has gained significant importance due to the growing number of individuals with hearing deficiencies whose main source of communication is sign language. CSLR is a sequence-to-sequence task, aiming to show a series of frames $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ to a sequence of glosses $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}$, where $\mathbf{x}$ and $\mathbf{y}$ are not necessarily equal. The generated glosses must align in order with the signs shown in the video. So, the challenge involves forecasting the accurate order of glosses based on a video depicting a series of gestures. This task encompasses two main steps: establishing time boundaries from loosely labeled video clips and recognizing the signs presented

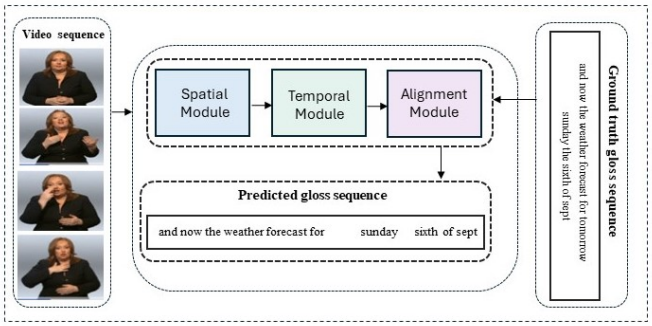


FIGURE 2. A General framework for Continuous Sign Language Recognition (CSLR).

AThe CSLR framework for applying deep learning has three main modules: spatial, temporal, and alignment, as shown in figure 5. The spatial module extracts visual features and physical time from the video input, captures the sequential patterns of the sign gestures, and the alignment module matches the sequences with gloss annotations, producing a structured output of glosses [11]. The performance evaluation of Continuous Sign Language Recognition (CSLR) is primarily conducted through the assessment of the Word Error Rate (WER). This metric quantifies the number of deletions, insertions, and substitutions required to align the predicted label sequence with the actual label sequence. A lower WER indicates a higher degree of accuracy. Traditionally, accuracy is quantified by calculating the ratio of correctly identified signs to the total number of signs present within a given sentence. Furthermore, the Bilingual Evaluation Understudy (BLEU) metric can serve as a valuable supplementary tool for assessing CSLR systems by enabling the comparison of n-grams from generated outputs against those derived from reference sentences, although BLEU is primarily intended for translation evaluations.

The tasks of Sign Language Translation (SLT) and Cross-Modal Sign Language Recognition (CSLR) are interconnected in the fields of computer vision and Natural Language Processing (NLP). CSLR systems output a sequence of sign labels that represent the order of signs demonstrated in the video, which may not align with the grammatical structure of natural language sentences [4]. Conversely, SLT builds upon the output of the Conversational Sign Language Recognition (CSLR) system by converting the identified signs into spoken or written natural languages, such as English. Effective SLT necessitates a comprehension of the context, grammar, and semantics of sign language, which renders it a more complex challenge compared to CSLR. SLT is combined with CSLR through two primary methods: sign2gloss2text [12], [13] and sign2text [14]. In the sign2gloss2text method, a CSLR model produces intermediate glosses, The SLT system employs specific methodologies to generate coherent and grammatically accurate text in the target language. Conversely, the sign-to-text approach eliminates the glossing phase, opting instead to translate expressions from sign language directly into natural language text.

TABLE 1. Input Modalities for Continuous Sign Language Recognition (CSLR).

Modality	Description	References
RGB	High-res gesture representation; feature extraction; often combined with depth and skeleton.	[14, 15, 16, 17, 18, 6, 19, 20]
Skeleton	Encodes joints; no background removal needed; effective in cluttered scenes.	[21, 24, 25, 26, 27, 28, 29]
Depth	Shows distance from objects; enhances RGB data representation.	[30, 31, 32, 33, 34, 35]
Pose Key Points	Models joints for GCNs; can be heatmaps; identifies key regions.	[13, 28, 36, 37, 38, 39]
Optical Flow	capture motion patterns; improves accuracy with RGB data.	[40, 41, 32]

A. INPUT MODALITIES FOR CONTINUOUS SIGN LANGUAGE RECOGNITION (CSLR)

Researchers have employed three primary input modalities for Continuous Sign Language Recognition (CSLR): RGB, depth, and skeleton information. RGB is the most commonly used modalityBecause of its superior clarity and intricate visual representation of sign movements [15]–[21], progress in computer vision and deep learning (DL) methodologies has facilitated the extraction of valuable features from RGB images. While RGB data serves as the principal modality, the integration of depth and skeletal information has the potential to enhance the performance of CSLR.

Skeletal or human posture information is the second most commonly employed modality in CSLR research. This type of data captures joint trajectories to create an abstract skeletal depiction of the signer, removing the necessity for background elimination and hand tracking, which are typical pre-processing steps for RGB images. Posture features have the capability to effectively address occlusions and clutter present in RGB images. Skeleton data can be acquired through specialized sensors or extracted from RGB imagery. Kinect and Leap Motion Controller (LMC) are common devices for capturing skeleton data. Recent high-accuracy pose estimation models like MediaPipe [22], OpenPose [23], and MMpose [24]provide more detailed pose features than sensor-based systems and are lightweight enough for mobile applications. OpenPose is particularly popular in CSLR literature [25]–[30].

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

Depth information, which indicates the distance between objects in an image and the capturing device,is often used alongside RGB data to provide a more comprehensive representation of sign gestures. Although its use has declined in recent years, depth data remains a valuable complementary modality in some CSLR studies [31]–[34]. Recent research continues to explore its potential, as seen in recent studies from 2023 and 2024 [35], [36].

CSLR researchers have employed various methods to utilize estimated human pose key points. For instance, [29] modeled the key points as a graph and processed through a Graph Convolutional Network (GCN), while others [37] used raw 3D pose coordinates. Another approach involved representing key points as heatmap images to reduce noise [14]. Some researchersutilized pose data to detect facial and hand areas, generating cropped images for vision-based recognition systems [38]–[40]. In addition to these primary modalities, optical flow has been used to describe motion patterns in consecutive frames. Studies have shown that combining optical flow with RGB data can improve recognition accuracy by capturing dynamic gesture movements more effectively [41]–[43].

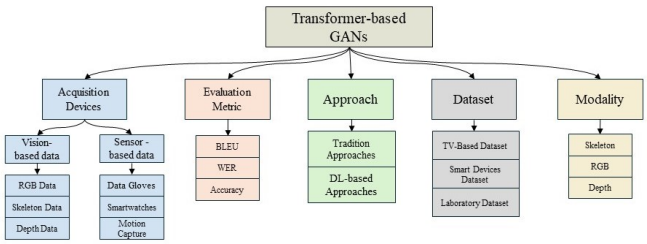


FIGURE 3. A General framework for Continuous Sign Language Recognition (CSLR).

These studies underscore the importance of combining multiple modalitiesto improve the performance and resilience of CSLR systems. The combination of RGB, depth, and skeletal data helps in capturing the complex dynamics of sign language more effectively, leading to more accurate recognition outcomes.

III. ACQUISITION DEVICES

Sign language recognition (SLR) approaches can be categorized based on the vision- orVision-based approaches use cameras to gather data, depicting signs as videos or images, while data acquisition can also be achieved through sensor-based devices. Studies have demonstrated the effectiveness of vision-based methods in recognizing sign language using deep learning techniques such as convolutional neural networks (CNNs) [9], [12], [20], [43]–[51]. The use of advanced camera systems like the Microsoft Kinect, which provides RGB, depth, and skeleton data, has enabled the capture of detailed hand and body movements [52]–[56]. Vision-based approaches have also been enhanced through the integration of transfer learning and multi-modal systems [57]–[60]. Methods utilizing sensors incorporate tools like data gloves and armbands equipped with sensors to track and gather sign language information [61]–[64]. In contrast, vision-based methods utilize cameras to capture signs as images or videos. Additionally, sensor-based methods utilize data gloves and armbands for tracking and recording sign data. The advantage of sensor-based Sign Language Recognition (SLR) techniques is that they eliminate the requirement for a picture pre-processing process, leading to lower computational demands.Before 2007, most studies preferred collection of data through sensors due to the challenges of vision-based recognition. which often required Comprehensive feature extraction and faced challenges such as occlusions, varying illumination, complex backgrounds, and different viewpoints. However, sensor-based approaches have high costs and practical limitations, Since the signer is required to wear the sensors, their use in practical situations is limited. In contrast, vision-based methods are more affordable and user-friendly, with cameras being widely available in households. According to a survey, 82% of studies on Continuous Sign Language Recognition (CSLR) used vision-based approaches, while 18% employed sensor-based methods

A. SENSOR-BASED METHODS

There have been various studies recognizing sign language through the use of sensors such as sensor gloves, wristbands, and smartwatches. Examples of sensor devices used to recognize sign language performed by users are illustrated in Figure N. According to Table N, which classifies sensor-based CSLR studies [65]–[72] it can be observed that many of these studies relied on gloves equipped with sensors to recognize the user's sign language. Recently, researchers studying sensor-based CSLR have been exploring methods to extend from wired systems to wireless devices using WiFi or Bluetooth signals [73].This section comprehensively discusses and analyzes sensor-based devices used in CSLR.

a: *Data Gloves*

Sensors affixed to the gloves are utilized to monitor movements. The information provided by these sensors encompasses the movements, orientation, and positional data of the user's hands and fingers.Flex sensors on each finger of the glove calculate the joint flexion, and sensors located in the glove's centre calculate the hand's position and orientation. The reliability of the data provided by the data glove can vary significantly depending on how well the glove size matches the user's hand size. Initial research [74] collected CSLR data using sensors attached to a single glove; however, due to the experimental limitation that most sign languages are two-handed, subsequent studies integrated data collection for both hands. The Cyberlove, like the Data Glove, is another type of device that tracks the user's hand using sensors. Several studies have used the Cyber Glove for CSRL [65], [67], [68], [75] . The Cyber Glove offers improved hand-tracking accuracy and convenient attachment compared to the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

Data Glove, but it has the disadvantage of being high-cost for commercial use. The DG5 VHand is a low-cost alternative that includes multiple flex sensors and a 3D accelerometer to detect the orientation and movement of the hand. This device has demonstrated its utility in several CSLR systems [69], [72], [76]. In addition to the aforementioned devices, due to the high cost associated with actual application fields, custom sensor gloves have been developed for experimental and research purposes at a lower cost.These custom sensor gloves are designed to replicate the functionality of high-end devices while being accessible for academic and experimental use. They integrate advanced sensor technologies, such as strain sensors and 3D-knitted electronics, to provide accurate hand gesture recognition without the financial burden of commercial products. Researchers have utilized these gloves in various fields, including sign language translation, post-operative rehabilitation, and virtual reality, to enhance the interaction between humans and machines. The adaptability and lower cost of these custom gloves make them ideal for testing and prototyping new interaction methods, enabling a broader range of applications in both healthcare and technology sectors [77]–[81].

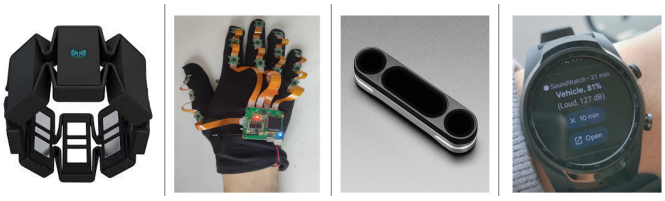


FIGURE 4. Types of sensor devices applied in the collection of sign language data (a) Myo armband [82] (b) Data gloves [81] (c) Ultraleap Motion [83] (d) SoundWatch [84].

b: *Smartwatches and Armbands*

Smartwatches, as embedded-system wrist-worn wearable devices, are designed to capture users' biometric signals, infer health conditions, and provide timely notifications. These smartwatches are equipped with sensors for detecting physical activities, such as 3D accelerometers and 3D gyroscopes, enabling the development of various smart applications [85]. However, similar to how spoken languages exhibit differences in articulation, pronunciation, and intonation among speakers, sign language users also demonstrate significant variations in their movements, such as elevation and rotation. These variations can lead to substantial differences in the data collected. Consequently, sensors like accelerometers and gyroscopes included in smartwatches have limitations when used for continuous sign language recognition (CSLR) [86].

Recent research has explored various approaches to address these challenges. For example, [87]proposed a framework for translating American Sign Language (ASL) video data into corresponding IMU data to enhance CSLR performance . Similarly, Caballero-Hernandez et al. (2023) developed an IoT-based system for translating Mexican Sign Language (MSL) into Spanish using machine learning, achieving high accuracy in gesture recognition [88]. Gu introduced a method for collecting an ASL dataset using a wearable inertial motion capture system, achieving high accuracy in sign language recognition and translation [89].

Additionally, Carlsson and Samuelsson developed an application for the Apple Watch that translates Swedish sign language into text and sound. This application leverages the watch's accelerometer, gyroscope, and rotation rate sensors, employing machine learning models to achieve significant accuracy in both letter and word detection [90][6].Researchers also the use of smartwatches for augmentative and alternative communication (AAC), focusing on discreet and effective communication for individuals with language impairments [91].In addition to sensor-equipped gloves, band-type devices with sensors that can be attached to the arm are also another option for capturing user movements [82], [92], [93]. Electromyography (EMG) sensor [82], which detect muscle movements, can be attached to these arm bands to capture specific movements by detecting muscle activity during the performance of these movements. However, EMG sensors can be influenced by physiological factors such as body temperature, experimental environment, volume, and body fluid levels, making it difficult to obtain accurate data due to noise when detecting adjacent muscle activity [92].

c: *Motion Capture Devices*

Sensors worn directly on the user's body must be worn while performing sign language, which presents limitations in terms of convenience and usability. Motion capture technology, which can capture user movements without the need for wearing cumbersome sensors, has been proposed to address these issues. The Leap Motion Controller (LMC) is a motion capture device that uses two infrared (IR) cameras to capture the positional information of a person's hands and fingers. LMC has been used for CSLR in studies by Lang et al. [94], Fang et al. [25], Mittal et al. [28], and Akkar et al. [95]. However, LMC can fail to accurately capture hands when performing specific actions at certain angles. For instance, if a finger that needs to be captured at a specific angle is obscured by another finger, LMC cannot detect the obscured finger, making it difficult to determine whether the finger is bent or extended, resulting in a confused state. Moreover, while LMC is not directly connected to the body of users, it captures the user's body through infrared cameras, requiring the user to be within the camera's effective range. Since infrared cameras use a wide field of view to capture the target, there can be noise caused by surrounding objects or other people. To mitigate these issues, LMC should be used in an isolated environment to avoid capturing other objects and people, and additional preprocessing is necessary for data that includes noise.

d: *Signal-based methods.*

Gestures in sign language can be captured by analyzing the impact on surrounding Wi-Fi signals. Zhang et al. [73] utilized the Channel State Information (CSI) of Wi-Fi signals

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

to collect American Sign Language sentences using a laptop as a transmitter and three laptops as receivers. The experimenter performed sign language between the transmitter and receiver laptops, achieving an average recognition accuracy of approximately 70%. However, the experimenter had to be positioned near the receivers and within a maximum range of 3 meters, which presents environmental limitations. Meng et al. [96] argued that recent studies on Sign Language Recognition primarily focus on three categories: computer vision-based, wearable sensor-based, and Wi-Fi signals-based methods. These methods are not user-friendly for everyday use, have significant noise, and are susceptible to signal interference. To overcome these limitations, Meng proposed using radio frequency identification (RFID) readers and tags with directional antennas, which do not interfere with each other's signals, to utilize radio frequency signals for sign language recognition. The proposed system demonstrated a high recognition accuracy ranging from 96% to 98.11%.The generalizability of the proposed framework has not been assessed in a signer-neutral context. An alternate method involves using Doppler radar to recognize sign language by evaluating the frequency changes of signals reflected from the target to calculate velocity data. This radar can achieve higher accuracy by overcoming basic noise in the detected movements. Ye et al. [97] used micro-Doppler signatures to recognize sign language, with the experimenter performing movements within 10 cm of the radar. Unlike previous studies, the data collected with this device was limited to the signal hand, restricting the system to signal hand sign language.

B. VISION-BASED DATA

Vision-based continuous sign language recognition (CSLR) uses video footage as the main source for capturing signing gestures [98]. High-quality cameras with a high frame rate are essential to accurately record the fast and intricate movements of sign language. In many households,Webcams and smartphone cameras offer a practical solution for capturing sign language. The popularity of high-resolution, 4K cameras in the latest smartphones has increased among researchers for capturing detailed sign language gestures, as demonstrated by Mukushev et al. [10]. Additionally, the Microsoft Kinect is widely used due to its ability to provide RGB, depth, and skeleton data through its dual cameras and infrared sensor. Researchers like Huang et al. [6] and Jebali et al. [32] have extensively used this combination of data types to develop comprehensive multi-modal CSLR datasets. Kagirov et al. [99], Ganesan et al. [100],Aloysiuset al. [101],Srivastava et al. [102] and Padmavathi et al. [103] have also contributed to the literature, by utilizing these technologies.

C. SENSOR-BASED VS. VISION-BASED METHODS

Sensor-based methods utilize specialized hardware such as data gloves, EMG sensors, or inertial measurement units to capture precise hand movements and muscle activity. These systems excel in controlled environments where occlusion or background interference is minimal. However, their high costs and the requirement for wearable devices often limit their scalability for broader adoption. For example, bending sensor gloves capture the degree of finger curvature effectively but struggle to differentiate gestures with similar curvature patterns. In contrast, vision-based systems use cameras to acquire RGB or depth data, enabling non-intrusive, scalable solutions. These systems benefit from widespread hardware availability and advancements in deep learning for gesture recognition. However, they often require large datasets and face challenges such as occlusions, variable lighting, and background clutter, which can reduce their performance in uncontrolled environments [104], [105], [106].

D. HYBRID APPROACHES

Hybrid methods combine sensor-based and vision-based modalities, leveraging the strengths of both to address their individual limitations. For instance, a system that fuses bending sensor data with RGB video has been shown to improve recognition accuracy, increasing performance from 68.34% with vision-only data to 84.13% when both modalities are integrated [104]. Another example involves the combination of MediaPipe holistic landmarks with LSTM models and YOLOv6, achieving 96% accuracy for static gestures and 92% for continuous gestures, demonstrating the effectiveness of multimodal fusion in recognizing complex sign sequences [107]. Additionally, error-elimination techniques have been implemented in hybrid systems, such as merging data from gloves and cameras to compensate for inaccuracies inherent in each modality, further enhancing robustness in dynamic and uncontrolled settings [105] [106]. These approaches have also found applications in AR and VR systems, particularly for immersive sign language learning environments.

IV. METHOD AND MATERIAL

A. CONTINUOUS SIGN LANGUAGE RECOGNITION (CSLR) DATASETS

The development of an intelligent Sign Language Recognition (SLR) system depends heavily on the presence and use of well-annotated datasets. SLR datasets are commonly classified as datasets for isolated and continuous sign language. Isolated sign datasets typically have the graphical information like videos or images capturing separate signs. These datasets give importance to discrete signs without context or continuous conversation and are used for training and evaluating systems that recognize isolated signs. On the other hand, CSLR data feature signs performed in a continuous flow, forming sentences without breaks and are essential for the development of Continuous Sign Language Recognition (CSLR) systems.

Isolated sign datasets are increasingly gaining prominence in various fields because recording isolated signs is simpler and is not essential for ex- tensive sign language experience. Continuous signing, however, is more complex and requires skilled signers. Due to the scarcity of CSLR datasets, scholars usually used their datasets of small size

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

prepared by themselves for training and evaluation, Several studies emphasize that larger datasets improve model generalization. The RWTH-PHOENIX-Weather-2014-T dataset, for instance, has been widely used to enhance recognition accuracy. However, empirical studies show varying impacts of dataset size on generalization. For example, Koller et al. [108] observed diminishing returns in performance gains beyond a certain dataset size, indicating that simply increasing data may not always be beneficial. Additionally, data enhancement methods, such as augmentation techniques (e.g., frame interpolation, random cropping, and motion blurring), impact model performance differently. Wang et al. [109] demonstrated that applying temporal augmentation significantly improved model robustness, while spatial transformations had minimal impact. These findings suggest that not all data enhancement methods contribute equally to generalization, warranting a more detailed analysis of augmentation strategies in CSLR. posing challenges for inter-model performance evaluation. Several crucial factors should be consid- ered for SLR datasets:

- **Vocabulary Size:** The dataset should have a broad vocabulary to enable application across various domains.
- **Number of Signers:** Including a diverse range of signers aids in the creation of signer-independent CSLR systems. Variability in appearance, signing speed, and handedness contributes to the development of robust models that generalize well to new signers.
- **Number of Samples per Sign or Sentence:** It's essential to have an adequate number of samples, When signers are few, it's important to improve variety to avoid overfitting.
- **Dynamic Signs:** In contrast to static signs, which are frequently used for alphabet letters or signs not present in the language. Additionally, the dataset must incorporate dynamic signs that require movement.

With the previously mentioned standards, CSLR datasets must contain a significant number of naturally performed sentences covering a diverse vocabulary. These sentences should flow naturally without undue pauses or explicit segmentations between signs, as the model is expected to automatically handle alignment and segmentation. The dataset must encompass a variety of sign formats, including static, dynamic, and finger-spelled signs, as well as sentences of varying lengths. Furthermore, gloss annotations are required to train the algorithm to detect sentences in sign language. Including natural language translations for each sentence would also be beneficial, as it would enable further investigation into the linguistic features of sign communication and the development of automated translation frameworks designed to convert sign language expressions into verbal language.

Dataset size plays a critical role in determining the model's ability to generalize effectively to unseen data. Larger datasets, such as the RWTH-PHOENIX-Weather-2014-T dataset [7], typically contain diverse examples and are better suited for training deep learning models. However, creating large datasets is resource-intensive. To address limitations posed by small datasets, researchers often employ data augmentation techniques such as flipping, cropping, scaling, and introducing noise. Additionally, synthetic data generation using Generative Adversarial Networks (GANs) and virtual signer models has shown promise in mimicking real-world signing conditions, enhancing generalization while minimizing annotation costs [42, 60]. However, over-reliance on synthetic data risks introducing biases or artifacts that may not align with natural signing, underlining the need for a balanced approach

A thorough comparison of publicly available Continuous Sign Language Recognition (CSLR) datasets is provided in Table 3, which evaluates the datasets based on a number of criteria,including sign language, vocabulary size, sentence count, signer count, modality, and domain. The dataset that we used includes coverage for the following languages: Arabic, Greek, Russian, Chinese, German, and American. Transitioning from controlled to real-world environments introduces significant challenges for CSLR systems. Controlled datasets like the CSL dataset [6] provide uniform lighting, consistent backgrounds, and signer postures, ensuring data quality but often lack variability. Real-world datasets, such as FluentSigners-50 [10], capture signing in diverse environments with varying lighting, dynamic backgrounds, and emotional expressions, enabling better generalization. Domain adaptation techniques and multi-modal datasets that combine RGB, depth, and infrared data are crucial to bridge this gap. For example, Scene-PHOENIX [104] enriches realism by incorporating artificial backgrounds, and adversarial training methods have been proposed to simulate environmental variability, improving robustness against real-world complexities [105].

Figure 5 illustrates the distribution of CSLR research across various sign languages, which illustrates the prevalence of these languages in the field. The Phoenix2014 dataset, which focuses on German Sign Language (GSL), boasts the largest vocabulary, comprising approximately 2048 signs [7]. In contrast, both the CSL dataset [6] and the FluentSigners-50 dataset [10]feature the highest number of signers, with 50 signers each. Over time, there has been an increase in the number of video samples, with the FluentSigners-50 dataset [10] currently holding the largest collection, totaling about 43,250 videos. The SIGNUM dataset [110], created in 2007, contains around 19,000 GSL sentences recorded in a controlled environment across general domains.In 2012, Forster et al. introduced the RWTH-PHOENIX-Weather-2012 [104], which was the largest publicly available dataset for CSLR at that time, specifically for German Sign Language (GSL). This dataset features a vocabulary of around 1,389 signs and includes 6,841 sentences. This dataset was expanded in 2014 (Phoenix2014) [7], and the vocabulary size is doubled. Due to its real-life weather forecast recordings, the Phoenix2014 dataset presents challenges. It has an out-of-vocabulary (OOV) rate of 0.54%, with 30% of the vocabulary being represented

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

TABLE 2. Summary of Acquisition Devices for Continuous Sign Language Recognition (CSLR).

Category	Description	References
Vision-based Methods	Cameras capture signs as videos or images; effective with CNNs and advanced systems like Kinect; integrates transfer learning and multi-modal systems.	[9, 42, 43, 44, 11, 45, 46, 19, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59]
Sensor-based Methods	Devices like data gloves, wristbands, and smartwatches track sign language; bypass image pre-processing, reducing computational demands; high costs and practical limitations; examples include Data Glove, Cyber Glove, DG5 VHand, and custom gloves.	[60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 76, 77, 78, 79, 80]
Motion Capture Devices	Capture movements without sensors; examples include Leap Motion Controller (LMC) with infrared cameras; issues with accuracy at certain angles and potential noise.	[91, 24, 27, 92]
Signal-based Methods	Analyze impacts on Wi-Fi signals or use radio frequency; examples include Wi-Fi CSI, RFID systems, and Doppler radar; challenges with signal interference and environmental limitations.	[72, 93, 94]
Vision-based Data	Uses video footage with high-quality cameras; popular technologies include webcams, smartphones, and Kinect; recent research utilizes high-resolution cameras and multi-modal datasets.	[95, 10, 7, 31, 96, 97, 98, 99, 100]

only once in the training dataset. make it difficult to use. Additionally, the dataset includes Phoenix2014SI, which is a signer-independent split, while signer-5 is not incorporated in the testing phase. The RWTH-PHOENIX-Weather-2014-T (Phoenix2014T) dataset features enhanced annotations, including gloss and German translations, which apply to both Continuous Sign Language Recognition (CSLR) and Sign Language Translation (SLT). Improvements have also been made to the dataset's annotations and sentence segmentation.

The CSL dataset [6], released in 2018, has been widely used in recent research. Created in a controlled laboratory setting, It comprises approximately 100 sentences captured from 50 signers. The dataset is provided in two versions: The CSL dataset comprises two distinct splits for evaluation purposes. CSL Split I is designed for signer-independent assessment and features videos from 40 unique signers in the training set, as well as 10 signers in the testing set. In contrast, CSL Split II is focused on testing previously unseen sentences, encompassing 6% of sentences that were not utilized during training. Despite the involvement of numerous signers, the CSL dataset is limited to 100 unique sentences and 178 distinct signs.To overcome these limitations, a new dataset called CSL-Daily [40]was introduced.Ten signers recorded 2000 signs and 6598 words in a lab setting for this dataset.FluentSigners-50 [10] was not prepared in a controlled environment, in contrast to other datasets. Rather, it used public to record movies in diverse real-world environments with a variety of gadgets, including webcams and cellphones. This variance in recording settings can help CSLR systems become more generalizable and perform better at identifying everyday sign movements.

In a comparable way, the Scene-PHOENIX dataset [111] added several artificial backgrounds to the Phoenix2014 dataset in order to improve the realism of CSLR. This method can be used to generate more realistic data from other lab-recorded CSLR datasets. Additionally, some datasets (Duarte et al. [112]; Huang et al. [6]; Luqman [8]) provide depth and skeleton information alongside RGB data. By providing a more diverse range of features for models to train from, multi-modal data sources allow for the extraction ofenhanced distinguishing characteristics that can better represent the nuances of sign language and enhance recognition performance. Fig. 3 shows examples of several dataset types, such as lab-created (CSL), real-world (Phoenix2014), and crowd-sourced (FluentSigners-50).The creation of multiple significant datasets in recent years, each intended to address distinct sign language processing issues, has greatly helped the area of Continuous Sign Language Recognition (CSLR). The King Saud University Saudi-SSL (KSU-SSL) dataset is notable for being the biggest Saudi Sign Language (SSL) dataset, with 293 signs, 33 signers, and 145,035 samples spread across 10 domains [113]. Intending to promote communication between the deaf and hearing communities, this dataset serves as a foundation for CSLR research in Arabic sign languages.

FIGURE 5. Sample images from the CSLR dataset illustrating diverse hand gestures and signs performed by different individuals.

A significant advancement in the field is the Continuous Word-Level Sign Language Recognition technology..Dataset created for Indian sign language [114], containing 80 static signs. This dataset, coupled with advanced machine learning models like YOLOv4 and Support Vector Machines (SVM), enables highly accurate real-time gesture recognition.Zhou et al. [115] introduced the CSL-Daily dataset, a large-scale Chinese Sign Language corpus focused on daily life scenarios, covering topics like family life, medical care, and shopping. The dataset includes videos recorded by 10 native signers,

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

with annotations for over 2,000 sign glosses and corresponding spoken language translations. The dataset is captured in 1920×1080 resolution at 30 FPS, and a sign dictionary (SignDict) is also provided for tasks such as sign spotting and isolated sign language recognition (SLR).Starner et al. [116]introduced the PopSign ASL v1.0 dataset, comprising over 210,000 examples of 250 isolated American Sign Language (ASL) signs, collected via smartphone cameras from 47 Deaf signers, with the aim of advancing real-time sign language recognition in educational gaming

Further expanding the diversity of sign language resources, the LSA64 Dataset [117] offers 3,200 videos of 64 distinct signs from Argentinian Sign Language (LSA), recorded by 10 subjects.The use of colored gloves enhances hand-tracking and segmentation, making this dataset valuable for machine learning tasks focused on recognition. Additionally, YouTube-ASL [124] presents a large-scale, open-domain dataset of American Sign Language (ASL), comprising over 984 hours of video and featuring more than 2,500 unique signers. This dataset provides a significant boost to ASL-to-English translation systems, offering a large corpus for pretraining and evaluation, including zero-shot translation capabilities. Collectively, these datasets are invaluable for advancing CSLR research, facilitating more accurate and inclusive communication technologies for deaf communities worldwide.

B. CSLR APPROACHES

The literature on continuous sign language recognition (CSLR) extensively explores different techniques, which can be broadly categorized into traditional methods and deep learning (DL)-based methods. Traditional techniques involve manual feature extraction from sign gesture videos or images, while DL-based methods autonomously acquire the required features directly from the data, resulting in a more flexible and adaptive feature representation.

1) Deep learning approaches

Since 2015, the majority of research in Conversational Speech Language Recognition (CSLR) has shifted from hand-crafted features to a focus on deep learning methods for feature extraction. The advent of Convolutional Neural Networks (CNNs) led researchers to employ deep CNN architectures for improved frame-level feature extraction, resulting in significant improvements in CSLR model performance. These extracted features were then utilized in various temporal learning models, such as Hidden Markov Models (HMM) and Recurrent Neural Networks (RNN). Recently, there has been a growing interest in exploring alternative techniques for CSLR, including Graph Convolutional Networks (GCN) and Transformers.We organize and assess CSLR-based research publications on popular architectures, including CNN-HMM, in the subsections that follow. [47], [126], [127], CNN-RNN [25], [42], [92], [128], Transformer networks [20], [129], [130] 3D CNN [43], [131], [132], and GCN [29]. While traditional deep learning models have shown strong performance in CSLR, their computational complexity can hinder real-time applications. Recent studies have explored lightweight architectures to address this limitation.Recent advancements have explored the feasibility of lightweight models for real-time Continuous Sign Language Recognition (CSLR), demonstrating that efficient architectures can achieve low-latency inference while maintaining competitive accuracy. For instance, Liu et al. [133] proposed RealTimeSignNet, a lightweight 3D deep learning network designed specifically for real-time sign language recognition, optimizing computation cost without compromising recognition accuracy (Liu et al., 2025). Similarly,Mnif et al. [134] introduced a lightweight CNN model for hand gesture classification, achieving rapid inference (10.2s training time) and efficient edge computing performance, proving the viability of lightweight architectures in CSLR applications.These studies provide empirical evidence supporting the claim that lightweight models and computationally efficient methods can enhance real-time CSLR performance. However, despite these promising findings, further comparative analyses across different architectures are necessary to fully understand the accuracy-speed trade-off in real-world settings. Future work should benchmark various lightweight models under standardized real-time constraints to solidify their role in CSLR applications.

Table 8 provides a overview of the examined deep learning-based CSLR methods.

2) CNN and RNN

The challenges associated with Hidden Markov Models (HMMs) in effectively capturing the broader context of sign language sentences highlight an opportunity for improvement. This recognition has led to the implementation of Recurrent Neural Networks (RNNs) in Continuous Sign Language Recognition (CSLR), enhancing the model's ability to interpret complex sign language expressions more accurately. RNNs, adept at handling long-range dependencies in sequences, have introduced new capabilities in this field. The introduction of Long Short-Term Memory (LSTM) networks has been instrumental in addressing challenges such as the problem of vanishing gradient in conventional RNNsThe use of bidirectional RNNs, in particular bidirectional LSTMs (BLSTMs), is noteworthy has enabled bidirectional information flow in a sequence, allowing for the modeling of more intricate relationships in CSLR. The pioneering work of [135] introduced the application of BLSTMs to CSLR, proposing a CNN-BLSTM model

combined with Connectionist Temporal Classification (CTC).Recent research has focused on various strategies to enhance the performance of CNN-BLSTM models. One notable approach is the introduction of SubUNets by Camgoz et al. [137]. This methodology demonstrates the effectiveness of training two SubUNets utilizing both segmented hand images and complete-frame visuals, significantly improving the performance of Continuous Sign Language Recognition (CSLR) [151] . . [138] presented an iterative training technique to

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

TABLE 3. Overview of Sign Language Datasets.

Dataset	Year	Language	Vocab.	Participants	Instances	Modality	Domain
Purdue RVL-SLLL [118]	-	ASL	104	14	2,576	RGB	Emergency
RWTH-BOSTON-104 [119]	2007	ASL	104	3	843	RGB	General Use
SIGNUM [110]	2008	German	450	25	19,500	RGB	General Use
Assaleh [120]	2010	Arabic	80	1	760	RGB	General Use
RWTH-PHOENIX-Weather [121]	2012	German	1,081	7	2,640	RGB	Weather
RWTH-PHOENIX-Weather-2014 [7]	2014	German	2,048	9	6,841	RGB	Meteorology
RWTH-PHOENIX-Weather-2014-T [45]	2018	German	1,066	9	8,257	RGB	Meteorology
CSL [6]	2019	Chinese	178	50	25,000	RGB, Depth, Skeleton	General Use
TheRuSLan [99]	2020	Russian	164	13	N/A	RGB, Depth	Supermarket
ArSL for Deaf Drivers [122]	2021	Arabic	N/A	3	N/A	RGB	Driving
LMSLR [29]	2021	Chinese	298	10	10,000	Skeleton	General Use
Continuous GrSL [9]	2021	Greek	310	7	10,295	RGB, Depth	Public Services
CSL-Daily [115]	2021	Chinese	2,000	10	21,000	RGB	Daily Life
How2Sign [112]	2021	ASL	N/A	11	2,456	RGB	Various
ArabSign [8]	2022	Arabic	95	6	9,335	RGB, Depth, Skeleton	General Use
FluentSigners-50 [10]	2022	Kazakh-Russian	278	50	43,250	RGB	General Use
ASL-Homework [123]	2022	ASL	N/A	45	935	RGB, Depth	General Use
Scene-PHOENIX [111]	2022	German	2,048	9	6,841	RGB	Weather
KSU-SSL [113]	2023	Saudi Sign Language	293	33	145,035	RGB	Multiple Fields
CSL-Daily [115]	2023	Chinese	2,000	10	21,000	RGB	Daily Life
LSA64 [117]	2023	Argentinian Sign Language	64	10	3,200	RGB	General Use
PopSign ASL v1.0 [116]	2024	ASL	250	47	210,000	RGB	Educational
YouTube-ASL [124]	2024	ASL	N/A	2,500	984 hours	RGB	Translation
ArSL Multimodal Dataset [125]	2024	Arabic	N/A	2	262	RGB, Audio	Religious

TABLE 4. Summary of Deep Learning-based Methods for CSLR.

Year	Study	Approach	Dataset
2012	Gweth et al. [136]	MLP-HMM	SIGNUM
2017	Koller et al. [135]	CNN-BLSTM-HMM	Phoenix2014 SI
2017	Koller et al. [135]	CNN-BLSTM-HMM	SIGNUM
2017	Camgoz et al. [137]	CNN-BLSTM	Phoenix2014
2017	Cui et al. [138]	CNN-BLSTM	Phoenix2014
2018	Huang et al. [6]	3D-CNN with Attention	CSL Split I
2019	Cui et al. [42]	CNN-BLSTM	SIGNUM
2019	Zhang et al. [129]	3D-CNN-Transformer	Phoenix2014
2019	Pei et al. [139]	3D-CNN-BGRU	Phoenix2014
2019	Zhou et al. [140]	3D-CNN-GRU	Phoenix2014
2019	Wei et al. [141]	W3D-CNN-BLSTM with N-Grams	CSL Split II
2020	Pu et al. [128]	CNN-BLSTM	Phoenix2014 SI
2020	Cheng et al. [142]	Fully Connected Network (FCN)	CSL Split I
2020	Papastratis et al. [49]	CNN-1D-CNN-BLSTM	CSL Split I
2020	Tateno et al. [82]	LSTM	Private Dataset
2021	Papastratis et al. [143]	Generative Adversarial Networks (GAN)	GrSL-SD
2021	Adaloglou et al. [9]	EnStimCTC (CNN-1D-CNN)	GrSL-SD
2021	Adaloglou et al. [9]	EnStimCTC (I3D-BLSTM)	GrSL-SI
2021	Min et al. [46]	CNN-1D-CNN-BLSTM	CSL Split I
2021	Hao et al. [144]	CNN-1D-CNN-BLSTM	Phoenix2014T
2022	Zhu et al. [145]	CNN-1D-CNN-Transformer	Phoenix2014
2022	Aditya et al. [37]	CNN-Attention-BLSTM	Phoenix2014
2023	Hu et al. [146]	CNN-1D-CNN-BLSTM	CSL Split I
2023	Jiao et al. [26]	Co-Sign (GCN-1D-CNN-BLSTM)	CSL-Daily
2023	Zheng et al. [147]	CNN-Attention-BLSTM	Phoenix2014
2023	Xie et al. [50]	Fully Connected Network (FCN)	CSL Split I
2023	Cui et al. [148]	Spatial Transformer (ViT-Transformer)	Phoenix2014
2023	Chen et al. [14]	Two-Stream SLR (3D-CNN)	Phoenix2014
2024	Hu et al. [18]	PA-CMA (CNN-1D-CNN-BLSTM)	CSL Split II
2024	Hu et al. [149]	AdaSize (CNN-1D-CNN-BLSTM)	CSL Split I
2024	Zuo et al. [150]	SRM (CNN-Transformer)	CSL Split I

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

improve feature extraction in CSLR. The network-like approach was trained using expectation maximization in this manner, and the feature extractor was then optimized using the pseudo labels that the network produced. When evaluated against their previous framework, Deep-Sign [126], on the Phoenix2014 dataset, the resulting system, ReSign [135], integrated CNN-BLSTM with HMM and obtained a 12% reduction in Word Error Rate (WER). Camgoz et al. [137] made significant strides in the field with their introduction of SubUNets, which effectively combine Bidirectional Long Short-Term Memory networks (BLSTMs) with 2D Convolutional Neural Networks (CNNs). Their research highlights a promising approach, showing that training two separate SubUNets—one focused on cropped hand images and the other on full-frame images—can lead to optimal performance. Furthermore, in 2020, Pu et al. [128] contributed to the ongoing development in this domain by introducing Cross Modal Augmentation (CMA). This innovative technique allows for the generation of pseudo text-video pairs through strategic operations such as deletion, substitution, and addition, thereby enhancing both the text labels and their corresponding video frames. These advancements pave the way for further exploration and improvement in multimodal approaches. However, this method involved manual efforts to create pseudo labels. Hu, Pu, et al. [18] improved Prior Aware CMA (PA-CMA), which made use of a language model to automatically provide pseudo labels, in order to expedite this procedure. Furthermore, Cui et al. [42] proposed employing gloss-level alignment proposals for training the feature extractor, which resulted in a 4% WER decrease on the Phoenix2014 dataset in comparison with the ReSign model. Chinese Sign Language (CSL) and German Sign Language (GSL) have been the main subjects of several CSLR models. Recent studies have highlighted the effectiveness of various sign languages, including those with limited proprietary datasets such as Japanese Sign Language (JSL), in the realm of research that employs Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [152]. The integration of CNN and RNN architectures has demonstrated significant promise in modeling sensor data acquired from devices such as the Leap Motion Controller (LMC) [25] and sensor gloves [92], in addition to vision-based techniques. Sharma et al. [92] investigated the efficacy of pre-training Continuous Sign Language Recognition (CSLR) models on datasets focused on isolated sign language recognition. Meanwhile, Fang et al. [25] introduced a hierarchical bidirectional RNN (HB-BRNN) specifically designed to model skeletal data effectively. It's crucial to remember that this method was only tested on a tiny, private sample of 40 Indian Sign Language (ISL) expressions.

C. BODY POSE AS A NON-MANUAL CUE

Body pose has emerged as a critical non-manual cue in various fields such as human-computer interaction, emotion recognition, and gesture analysis. It provides essential information about an individual's posture, movement, and overall intention, contributing significantly to the understanding of non-verbal communication. Researchers have explored body pose as an independent feature, focusing on the estimation of upper body pose, which includes the arms, shoulders, and torso. Studies such as Brock et al. [152], Jiao et al. [26], Ko et al. [27], and Wang and Zhang [29] have investigated methods for accurately detecting and analyzing the upper body pose. These works show that upper body posture can be used to infer activities, emotions, and interactions, enhancing systems that depend on understanding human behavior.

In addition to the upper body, body pose is often integrated with other cues, such as hand gestures, to improve the precision and context of recognition systems. Research by Cui et al. [42], Forster et al. [153], and Gweth et al. [136] examines how hand movements, within a full-body frame, add important context to pose estimation, especially for gesture recognition applications. This combination of body pose with hand gestures is crucial for understanding complex actions, where the position and movement of the hands are intricately linked with the overall body posture.

Moreover, full-body pose estimation has garnered attention, with studies like Aditya et al. [37], Chen et al. [14], Wei and Chen [154], Li and Meng [155],and Zuo and Mak [150] focusing on the holistic analysis of body pose. These works aim to capture the full-body movement dynamics, not only the upper body, to interpret complex gestures and human behaviors. Full-body pose systems can detect subtle motions that might be missed in isolated hand or head pose recognition, thus providing a more comprehensive understanding of the individual's actions.Body pose is frequently combined with facial expressions, mouth movements, or gaze direction to enhance gesture and emotion recognition systems. Notable works, such as those by Forster et al. [153], Koller, Forster, and Ney [7], and Zhang et al. [34], explore the integration of body pose with facial features to improve the accuracy of systems designed for non-verbal communication recognition. The synergy between body and facial cues can help systems discern emotions or gestures more accurately, offering a richer analysis of human behavior. Furthermore, integrating body pose with gaze and head movements, as studied by Jebali et al. [32] and von Agris et al. [156], provides valuable contextual information that is especially important for applications like sign language recognition, human-robot interaction, and advanced emotion detection systems.

D. TRANSFORMER-BASED NETWORKS

Transformers [157] have fundamentally transformed the field of machine learning through their introduction of the self-attention mechanism. Various CSLR frameworks have utilized Transformers for sequence learning [20], [38], [145], [150], [158]integrated 3DRes-Net into an encoder-decoder architecture for Transformers, demonstrating the ability of Transformers to learn in series. Stochastic Fine-Grained Labeling (SFL) was introduced by Niu and Mak et al. [130] o optimize the alignment of sequences within a ResNet-Transformer framework. Zuo and Mak et al. [159]proposed

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

a refined self-attention technique called the Local Context-Aware Transformer Encoder (LCTE), which substanially improved the per formance over traditional Transformers.

Multi-modal Transformer-based models have also been explored. Slimane and Bouguessaet al. [20] used a 2D CNN-Transformer to encode both cropped hand and full-frame images. Alternately, in the C2SLR VGG11-Transformer model, posture heatmaps were employed to impose spatial attention [39]. On the CSL signer-independent dataset, this framework was further improved using the Signer Removal Method (SRM), yielding new state-of-the-art results [150]. The multi-modal SignBERT [38] encoded frames alongside cropped hand images using a ResNet-BERT model.

Recent studies have investigated leveraging textual informationfor CSLR model training. Guo et al. [158] implemented a BERT language model trained with gloss sequences to enhance a contextual module built with BLSTM. Similar to this, Zheng et al. [147] used gloss sequences to pre-train a language model that merges self-attention with BLSTM to strengthen contextual encoding and alignment, although this approach resulted in a modest performance gain of about 0.6 WER.The sign2gloss2text technique employs transformers to facilitate spoken language translations utilizing gloss predictions derived from the Continuous Sign Language Recognition (CSLR) module, which are subsequently integrated into a Sign Language Translation (SLT) module. This approach has been implemented in combined CSLR-SLT systems, as noted in the work by Papastratis et al. [143]. In SLRGAN [143], for example, a GAN is used to produce glosses, which a Transformer model eventually translates into text. Vision Transformers (ViT) treat images as sequences of patches, adding positional encodings, and encoding these sequences with standard Transformer encoders [155]. Research on CSLR has revealed a limited focus on the utilization of Vision Transformers (ViTs) for visual feature extraction. A noteworthy contribution is presented by [148], which developed a two-stream model that utilizes a pre-trained ViT to extract visual features from RGB frames. This model integrates an Attention-Enhanced Multi-Scale 3DGCN (AM3DGCN) to encode OpenPose features, resulting in Word Error Rates (WERs) of 1.9% on the CSL and Phoenix2014T datasets. They presented the Spatial Temporal Transformer (ST-Transformer), a fully Transformer-based framework that achieved a new state-of-the-art result on the CSL dataset with a 1.2% WER. Zhang et al. [13] attained notable word error rates (WERs) of 17.7% and 18.9% on the Phoenix2014 and Phoenix2014T datasets, respectively. They have further developed previous methodologies by introducing the Cross-modal Contextualized Sequence Transduction (C2ST) framework. This advanced architecture proficiently integrates textual information obtained from gloss sequences utilizing the BERT language model, thereby improving both contextual understanding and sequence transduction capabilities across diverse modalities.

Recent advancements in continuous sign language recognition (CSLR) aim to improve the extraction of complex temporal information in sign language videos, which involve fluent transitions and varying temporal scales. Huang et al. [160] propose a dual-stage temporal perception module (DTPM) that combines temporal convolutions and transformers to address these challenges by capturing richer and more comprehensive temporal features through a hierarchical structure. Alyami et al. [161] focused on isolated Arabic sign language recognition using a Transformer-based model and landmark keypoints, achieving high accuracy of 99.74% in signer-dependent and 68.2% in signer-independent modes on the KArSL-100 dataset. Liu et al. [162] proposed the Adaptive Video Representation Enhanced Transformer (AVRET) to address issues with temporal correspondence and weakly supervised sequence labeling in end-to-end sign language translation, achieving competitive performance on the CSL-FocusOn, PHOENIX14T, and CSL-Daily datasets.

Transformer-based models have demonstrated state-of-the-art performance in CSLR due to their ability to capture long-range dependencies and process sequential information efficiently. However, when compared with traditional CNN-RNN architectures, Transformers come with higher computational costs, which can hinder real-time applications. Studies such as Du et al. [163] have shown that while Transformers achieve higher recognition accuracy, CNN-RNN hybrids remain more computationally efficient, making them suitable for low-power devices. Additionally, Ranjbar and Taheri et al. [164] introduced a hybrid CNN-Transformer model that combines the strengths of both architectures, outperforming standalone models in CSLR tasks .Furthermore, lightweight architectures have gained attention as alternatives to computationally expensive Transformer models. Liu et al. [133] proposed RealTimeSignNet, a 3D deep learning network optimized for CSLR, demonstrating faster inference while maintaining competitive accuracy (Liu et al., 2025). Similarly, Mnif et al. [165] developed a lightweight CNN model that achieved high efficiency in CSLR tasks (Mnif et al., 2024). Lastly. Camgöz et al. [12] explored Neural Sign Language Recognition, comparing CNN-RNN models with Transformer-based approaches, concluding that while Transformers provide greater flexibility in learning representations, RNN-based models still hold advantages in efficiency and training stability.

These comparative studies highlight the trade-offs between different CSLR architectures, demonstrating that while Transformers improve accuracy, CNN-RNN hybrids and lightweight models remain viable solutions for real-time performance. Future research should focus on benchmarking these models on standardized datasets, evaluating both recognition rates and computational efficiency to determine the most suitable architecture for CSLR applications.

E. TRADITIONAL APPROACHES

Before 2015, most research on CSLR [68], [74], [166]–[170], focused on traditional methods, as outlined in Table 4. These studies commonly used handcrafted features like Histogram of Oriented Gradients (HOG) [171], and Fourier descriptors

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

[172]. The sequence of signs was primarily modeled using Hidden Markov Models (HMM) [166] and Dynamic Time Warping (DTW) [67].

1) Dynamic Time Warping (DTW)

A dynamic programming approach called dynamic time warping (DTW) is used to compare sequences with varying durations and speeds. DTW has been applied in various CSLR studies [85], [173], [174]. For example, [70] employed a CyberGlove for data collection and modeled the data using DTW. Additionally, DTW has been used to vision-based CSLR systems. Dynamic programming was used by Yang et al. [168]to apply a level building (LB) strategy for segmenting signs in American Sign Language (ASL).Yang and Lee (2011) improved this strategy even further by combining DTW with layered dynamic programming. [169] achieved better performance by combining the Level Building method with HMM rather of DTW in a different research. In Zhang et al.'s [175] framework,HMM is used for classification, and DTW is utilized for segmentation. in order to fully use the advantages of both DTW and HMM. In citehassan2019multiple,zadghorban2018algorithm,hassan2016user, K-Nearest Neighbor (KNN)

was utilized as an alternative for classifying sentences segmented by DTW. However, despite its effectiveness, DTW is sensitive to variations in signing speed and has difficulty handling large vocabularies and unseen sentences [4].

Traditional techniques in CSLR literature also include graph modeling [177] and Conditional Random Fields (CRF) [185]. Graph modeling has been used for recognizing sentences in ArSL and ISL , When the phrases are shown as connected graphs and the detection is performed via a graph matching technique. To effectively capture the temporal dependencies inherent in sign language movements, Conditional Random Fields (CRFs) have been employed in Continuous Sign Language Recognition (CSLR) for probabilistic modeling [75]. The capability of CRFs to proficiently model both local and global contexts within sign language underscores their significance in advancing the field of sign language recognition,they provide a strong substitute for HMMs. Large vocabulary sets, however, may provide challenges for CRF-based CSLR systems, which also need a substantial amount of annotated data.

2) Hidden Markov Model (HMM)

The effectiveness of Hidden Markov Models (HMM) in audio recognition motivated scholars to explore their application in Continuous Sign Language Recognition (CSLR) [4]. HMM, a probabilistic model, is used in CSLR to determine the most likely sequence of signs corresponding to a sentence. The first application of HMM for CSLR can be traced back to Liang and Ouhyoung [74], who used it with a DataGlove to identify 196 phrases in Taiwanese Sign Language (TSL). A similar methodology was used in later research [65], [68], which employed HMM to analyzesign data collected using gloves. with sensors.By using colored gloves, Bauer et al. [166] were able to enhance hand tracking and detection.

Several hidden Markov model (HMM)-based continuous sign language recognition (CSLR) systems have been developed for various sign languages, leveraging the proven efficacy of HMMs in vision-oriented approaches. These include Italian Sign Language (ItSL) by Infantino et al. [65], [68], Spanish Sign Language (SSL) by Cortés et al. [41], Arabic Sign Language (ArSL) by Assaleh et al. [120], and Taiwanese Sign Language (TSL) by Yu et al. [178]. However, many of these models were trained on limited datasets with restricted vocabularies. Dreuw et al. [182] introduced RWTH-BOSTON-104, the first public CSLR dataset, featuring 104 American Sign Language (ASL) signsThrough the combination of HMM models trained on hand movement paths and speeds, a word error rate (WER) of 17.9% was achieved. In a later study, including depth data, a WER of 19.6% was reported Dreuw, Steingrube, et al. [31]. Subsequently, two public datasets for German Sign Language (GSL) were released: SIGNUM and Phoenix2014.Koller, Forster, and Ney [7] introduced the Phoenix2014 dataset for GSL, obtaining a 53% WER by utilizing HMM with maximum likelihood linear regression. In a multi-stream HMM framework, HMMs were also extended to represent non-manual and manual factors, like body stance and facial expressions [153], therefore increasing recognition accuracy. Jebali et al. [32] achieved a notable accuracy of 95.1% in the recognition of phrases in French Sign Language (FSL) by employing the Kinect sensor along with LMC for feature extraction; however, it is important to note that their dataset was limited to just 33 distinct signs. Researchers have proposed an intriguing approach to improve recognition rates—segmenting signs into smaller, manageable subunits—especially as the vocabulary of sign languages continues to grow [176], [180], [186].

Taking cues from the progress made in the field of speech recognition, several Hidden Markov Models (HMM) have been developed to identify these subunits, making use of techniques like K-means clustering to refine the process [180]. However, the inherent independence assumption of HMMs presents a significant challenge. This limitation makes it difficult to effectively capture the rich and complex features, as well as the broader contextual elements, of sign languages, consequently hindering the overall accuracy of Continuous Sign Language Recognition (CSLR).

3) Other Traditional Techniques

In the field of Continuous Sign Language Recognition (CSLR), several conventional methodologies have been developed, including the application of Conditional Random Fields (CRF) [185]. Graph-based modeling has been employed to recognize sentences in Indian Sign Language (ISL) [187], wherein sentences are represented as interconnected graphs. This representation facilitates the identification of signed sentences through graph matching techniques. To better capture the temporal dynamics inherent in sign lan-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

TABLE 5. Summary of Traditional CSLR Approaches.

Year	Reference	Method	Dataset
2021	Jebali et al. [32]	Hidden Markov Model (HMM)	Set of 33 signs
2019	Hassan et al. [72]	k-Nearest Neighbors (KNN)	Collection of 80 signs and 40 sentence examples
2019	Elakkiya and Selvamani et al. [176]	HMM	Phoenix2014 dataset
2017	Hassan et al. [76]	KNN-based approach	80 distinct signs, 40 sentences
2017	Ekiz et al. [85]	DTW, Logistic Regression	13 sentence gestures
2016	Yang et al. [30]	HMM-based technique	20 sentence-based signs
2016	Li et al. [71]	HMM	Dataset of 510 signs and 1,024 sentences
2015	Tripathi and Nandi et al. [173]	Dynamic Time Warping (DTW)	11 sentence gestures
2015	Tubaiz et al. [69]	KNN	Dataset of 80 signs and 40 sentences
2015	Koller, Forster, and Ney et al. [7]	HMM	Phoenix2014 dataset
2015	Koller, Forster, and Ney et al. [7]	HMM	SIGNUM dataset
2014	Kong and Ranganath et al. [75]	CRF-SVM hybrid model	107 signs and 74 sentences
2014	Zhang et al. [174]	DTW combined with HMM	180 sentence gestures
2013	Tolba et al. [177]	Graph Matching	Dataset with 100 signs, 30 sentences
2013	Forster et al. [153]	Multi-stream HMM	SIGNUM dataset
2011	Yu et al. [178]	HMM	40 signs, 3 sentence examples
2011	Sarkar et al. [179]	HMM-based segmentation	25 sentences
2010	Yang et al. [169]	DTW	Perdue dataset with 10 sentences
2010	Roussos et al. [180]	K-means clustering	400 signs and 843 sentence dataset
2009	Dreuw, Steingrube, et al. [31]	HMM, Principal Component Analysis (PCA)	RWTH-BOSTON-104 dataset
2009	Kelly et al. [181]	Multi-channel HMM	160 sentence gestures
2007	Dreuw et al. [182]	HMM-based method	RWTH-BOSTON-104 dataset
2007	Yang et al. [168]	Dynamic Programming (DP)	25 sentence gestures
2006	Vassilia and Konstantinos et al. [183]	HMM	71 distinct signs
2006	Guilin et al. [68]	HMM-based approach	543 sentence gestures
2006	Cortés et al. [41]	HMM	Dataset of 33 signs
2004	Gao et al. [67]	DTW	1,500 sentence examples
2002	Fang et al. [66]	SRN-HMM hybrid	Dataset of 100 sentences
2002	Bauer and Kraiss et al. [170]	HMM, K-means clustering	Dataset of 12 signs
2001	Wang et al. [65]	HMM-based method	100 sentence gestures
2001	Vogler and Metaxas et al. [184]	Parallel HMMs	22 signs
2000	Bauer et al. [166]	HMM	Dataset containing 97 signs

guage gestures, CRFs are utilized as a probabilistic modeling approach within CSLR [185]. Due to their capability to model both local and global contexts in sign language, CRFs present a viable alternative to Hidden Markov Models (HMM). Nonetheless, CRF-based CSLR systems necessitate considerable amounts of annotated data and may encounter challenges when processing extensive vocabularies of signs.

F. CSLR CHALLENGES

Sign Language Recognition (SLR) presents significant intricacies and challenges, particularly with regards to capturing features from multiple sources concurrently.Sign language encompasses gestures from various body parts, including hand movements and facial expressions, necessitating a robust SLR system capable of effectively capturing and integrating both manual and non-manual features.

Another substantial challenge pertains to ensuring signer independence. Variations in physical attributes, such as skin tone, body morphology, and stature, in addition to differences in sign execution due to factors like hand dominance, signing speed, and skill level, contribute to substantial individual divergences. These variations pose difficulties for SLR systems to generalize effectively to new signers, particularly in real-time applications.

In comparison to finger-spelled or isolated SLR systems, Continuous Sign Language Recognition (CSLR) presents even more formidable challenges, incorporating additional computational and linguistic complexities. Beyond the challenges mentioned earlier, we highlight some of the key complexities specific to CSLR in the following discussion:

- The appearance of a sign can change depending on the signs that come before and after it, which complicates recognition. Signs may look different based on their context, leading to variability in how they are perceived and recognized.
- Recognizing a sequence of signs requires understanding the relationships between signs over time. Some systems treat each sign independently, which can result in errors. More advanced models aim to capture how signs interact with each other to improve accuracy.
- Finding the start and end points of each sign in continuous signing is challenging because there are no clear breaks between signs. Some methods attempt to separate the signs before recognition, but this depends on how well the signs are divided. Techniques like Hidden Markov Models (HMM) and Connectionist Temporal Classification (CTC) can align signs automatically without needing explicit separation. However, accurately segmenting continuous signs remains a persistent challenge due to transitions and overlaps between signs that blur boundaries. Active learning approaches can iden-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

tify ambiguous or uncertain segments for annotation, thereby reducing manual effort while improving segmentation accuracy. Additionally, crowdsourcing offers a scalable solution to generate diverse and comprehensive annotations for these complex datasets.

- In real conversations, finger-spelled signs are sometimes used alongside regular signs. These finger-spelled signs are short and less dynamic, which makes recognizing them along with other signs more challenging. Transitions or finger movements can further complicate the recognition process.
- When transitioning between signs, additional movements that are not part of the actual signs can occur. These transitions can make it difficult to determine where one sign ends and another begins. Some systems are specifically designed to detect these transitions, while others incorporate them into the overall recognition process.

Addressing these challenges is essential for advancing CSLR systems beyond controlled research environments into real-world applications. While deep learning advancements have improved recognition accuracy, several fundamental gaps remain, particularly in dataset diversity, signer independence, and real-time performance. The next section explores these research gaps and outlines potential future directions to enhance CSLR models.

V. RESEARCH GAPS AND FUTURE DIRECTIONS

Despite recent advancements in CSLR, several challenges remain unaddressed, limiting the practical deployment of these systems. Key limitations include data scarcity, model generalization, and computational efficiency. This section categorizes these research gaps into three key areas: data, model, and computational constraints. The recent progress in deep learning-based continuous speech recognition (CSLR) systems has led to notable enhancements in accuracy. Despite these advancements, there are several persistent challenges. This section delineates the current constraints, divided into data-related and model-related challenges.

A. DATA-RELATED CHALLENGES

Dataset-related constraints further compound the complexity of CSLR. While advancements in deep learning have improved recognition accuracy, current datasets still suffer from limitations that hinder the generalizability of CSLR models.

- The importance of diverse and comprehensive datasets in the field of sign language recognition cannot be overstated. Most datasets are recorded in controlled settings with fixed environments, which limits the system's performance in real-world scenarios. This restriction highlights the critical need for more data from naturalistic, unconstrained environments with variable backgrounds, lighting, and angles. Such diverse data is essential for enabling sign language recognition models to perform effectively in real-world situations.
- Another area of concern is that existing datasets cover only a small range of sign languages, leaving many languages unrepresented. This limitation underscores the necessity for a broader range of datasets, particularly for underrepresented languages, in order to enhance the generalizability of sign language recognition models. Additionally, the alignment of gloss annotations across languages is a significant challenge. Standardized gloss annotations and the use of multilingual datasets can facilitate cross-lingual generalization and support the development of models capable of recognizing signs across diverse linguistic contexts. By incorporating more languages into the datasets, these models can become more inclusive and accessible to a wider range of sign language users.
- Furthermore, considering that sign language recognition models are expected to operate on mobile devices, datasets recorded from selfie-view perspectives on smartphones and tablets are essential. This approach ensures that the models developed are robust and device-friendly, allowing for practical and seamless integration into everyday mobile usage.
- In addition, the current datasets, such as Phoenix2014, have a narrow vocabulary and are domain-specific (e.g., weather forecasts). To make sign language recognition more applicable to diverse real-world contexts, datasets need to include signs from various topics and domains. This expansion will enable the development of models that can effectively recognize signs across a wide range of subjects, furthering the practical utility of sign language recognition in everyday life.

B. MODEL-RELATED CHALLENGES

- The existing models encounter difficulties when applied to new signers, primarily due to the limited diversity of signers in the available datasets. It is imperative to enhance signer diversity through data augmentation and explore signer-independent methods, such as pose-based techniques.
- The integration of information from multiple modalities (e.g., RGB, depth, skeletal data) has the potential to enhance accuracy but comes with increased computational demand. Therefore, future research should prioritize the development of more efficient approaches to fuse these data types without excessively complicating the system.
- Training models to simultaneously perform related tasks (e.g., sign language recognition and translation) has the potential to improve model performance through shared representations, thereby enhancing both recognition and translation accuracy.
- Non-manual cues, including facial expressions and eye movements, play a crucial role in sign language recognition. Despite their significance, these features are often underutilized in current systems. Further research is essential to effectively incorporate these cues.
- Interest in the development of real-time isolated sign

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

language recognition (SLR) has grown significantly, with several studies [172], [188] focusing on this area. However, research on real-time Continuous Sign Language Recognition (CSLR) has been relatively limited. This is mainly due to the challenges associated with developing real-time and online CSLR systems, which require low latency and high performance. In response to these challenges, there is a need for the creation of lightweight and efficient models [17], [149] capable of meeting the demands of real-time applications.

- Sign language recognition (SLR) encompasses not only a visual task but also involves language understanding. It is imperative to conduct more studies to integrate advanced language modeling techniques to enhance overall recognition and translation quality.
- The development of systems that can adapt based on user feedback during sign recognition has the potential to significantly enhance accuracy and robustness over time, especially in real-world applications.

C. COMPUTATIONAL-RELATED CHALLENGES

- Computational efficiency is a vital aspect, particularly in real-time and resource-constrained scenarios such as mobile and embedded systems. Many existing CSLR models demand substantial computational resources, making them less practical for devices with limited processing power or memory. Addressing this challenge is crucial for enabling wider adoption in everyday applications.
- Lightweight models that achieve a balance between high performance and reduced complexity are essential for real-time applications. Techniques such as model quantization, pruning, and knowledge distillation have shown promise in this area, as they significantly reduce model size and enhance inference speed while maintaining accuracy.
- Architectures optimized for low-latency scenarios, such as transformer-based models and attention mechanisms, offer further potential to improve computational efficiency. These designs need to be adapted to minimize the trade-offs between speed and accuracy, ensuring seamless integration into real-time systems.
- Training pipelines also require optimization to lower computational demands. Approaches like distributed training, adaptive learning rates, and efficient data augmentation can enable researchers to experiment with complex models more feasibly, even on constrained resources.
- Finally, integrating energy-efficient techniques into model design and execution can support sustainable, long-term operation on battery-powered devices. This approach ensures that CSLR systems are not only computationally efficient but also environmentally conscious in their deployment.

VI. CONCLUSION

The field of Computer-assisted Sign Language Recognition (CSLR) has been experiencing a surge in interest due to the continuous advancements in machine learning and related technologies. This surge is evident in the increasing number of research studies and the development of new models and frameworks aimed at improving sign language recognition capabilities. As part of this ongoing progress, a thorough review has been conducted to outline the achievements in the field and to identify critical areas that require further attention and development.One of the persistent challenges in CSLR is the need to address co-articulation effects and detect sign boundaries effectively. These challenges have been known to hinder the development and accuracy of sign language recognition systems. Additionally, the review of publicly available datasets has highlighted limitations in the diversity of languages represented, which in turn constrains the creation of models capable of generalizing across various sign languages and contexts.

Furthermore, a wide range of studies organized into different categories has covered various aspects of CSLR, including data acquisition, recognition methods, and input modalities. Despite the recent introduction of deep learning (DL)-based frameworks, many models still operate within restricted environments, limiting their practical applicability. While there has been a shift towards vision-based approaches in recent times, and the potential of multi-modal CSLR systems, there has been a noted lack of effective incorporation of linguistic knowledge in many of the systems, which could significantly enhance recognition performance.

Most recent models in CSLR utilize standard spatial and temporal feature extraction techniques, but they encounter limitations such as overfitting, especially when relying on alignment methods. Addressing these challenges through improved training strategies remains a critical focus for researchers in the field. It is important to ensure that the models are robust and capable of generalizing across different sign languages and contexts. The field of CSLR is evolving, and there are several promising directions for future exploration. These directions include the exploration of multi-modal fusion, integrating vision-language techniques, and adapting systems for complex environments, such as multi-person scenarios. Furthermore, research may expand into adjacent areas like finger-spelling and sign language translation, as well as exploring new developments in sign language generation.In summary, the field of CSLR is dynamic, with ongoing advancements and promising avenues for further exploration. It is clear that researchers and developers are committed to overcoming the existing challenges and making significant strides to make the better the accuracy and applicability in the recognition of sign language systems.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) under Grant 2022R1A2C1092178 and by Institute of Information communications Technology

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

Planning Evaluation (IITP) under the metaverse support program to nurture the best talents (IITP-2024-RS-2023-00254529) grant funded by the Korea government(MSIT) and by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00106, Development of explainable AI-based diagnosis and analysis frame work using energy demand big data in multiple domains).

REFERENCES

[1] W. Suliman, M. Deriche, H. Luqman, and M. Mohandes, "Arabic sign language recognition using deep machine learning," in 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT). IEEE, 2021, pp. 1–4.

[2] A. Wadhawan and P. Kumar, "Sign language recognition systems: A decade systematic literature review," Archives of Computational Methods in Engineering, vol. 28, pp. 785–813, 2021.

[3] S. Alyami, H. Luqman, and M. Hammoudeh, "Reviewing 25 years of continuous sign language recognition research: Advances, challenges, and prospects," Information Processing & Management, vol. 61, no. 5, p. 103774, 2024.

[4] N. Aloysius and M. Geetha, "Understanding vision-based continuous sign language recognition," Multimedia Tools and Applications, vol. 79, no. 31, pp. 22 177–22 209, 2020.

[5] E. M. El-Alfy and H. Luqman, "A comprehensive survey and taxonomy of sign language research," Engineering Applications of Artificial Intelligence, vol. 114, p. 105198, 2022.

[6] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.

[7] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," Computer Vision and Image Understanding, vol. 141, pp. 108–125, 2015.

[8] H. Luqman, "Arabsign: a multi-modality dataset and benchmark for continuous arabic sign language recognition," in 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). IEEE, 2023, pp. 1–8.

[9] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras, "A comprehensive study on deep learning-based methods for sign language recognition," IEEE Transactions on Multimedia, vol. 24, pp. 1750–1762, 2021.

[10] M. Mukushev, A. Ubingazhibov, A. Kydyrbekova, A. Imashev, V. Kimmelman, and A. Sandygulova, "Fluentsigners-50: A signer independent benchmark dataset for sign language processing," Plos one, vol. 17, no. 9, p. e0273649, 2022.

[11] L. Tightiz, L. M. Dang, S. Padmanaban, and K. Hur, "Metaverse-driven smart grid architecture," Energy Reports, vol. 12, pp. 2014–2025, 2024.

[12] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10 023–10 033.

[13] B. Zhang, M. Müller, and R. Sennrich, "Sltunet: A simple unified model for sign language translation," arXiv preprint arXiv:2305.01778, 2023.

[14] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, and B. Mak, "Two-stream network for sign language recognition and translation," Advances in Neural Information Processing Systems, vol. 35, pp. 17 043–17 056, 2022.

[15] L. Hu, L. Gao, Z. Liu, and W. Feng, "Temporal lift pooling for continuous sign language recognition," in European conference on computer vision. Springer, 2022, pp. 511–527.

[16] ——, "Continuous sign language recognition with correlation network," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2529–2539.

[17] L. Hu, L. Gao, Z. Liu, C.-M. Pun, and W. Feng, "Adabrowse: Adaptive video browser for efficient continuous sign language recognition," in Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 709–718.

[18] H. Hu, J. Pu, W. Zhou, H. Fang, and H. Li, "Prior-aware cross modality augmentation learning for continuous sign language recognition," IEEE Transactions on Multimedia, vol. 26, pp. 593–606, 2023.

[19] Y. Jang, Y. Oh, J. W. Cho, M. Kim, D.-J. Kim, I. S. Kweon, and J. S. Chung, "Self-sufficient framework for continuous sign language recognition," in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.

[20] F. B. Slimane and M. Bouguessa, "Context matters: Self-attention for sign language recognition," in 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 7884–7891.

[21] K. Koishybay, M. Mukushev, and A. Sandygulova, "Continuous sign language recognition with iterative spatiotemporal fine-tuning," in 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 10 211–10 218.

[22] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee et al., "Mediapipe: A framework for building perception pipelines," arXiv preprint arXiv:1906.08172, 2019.

[23] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7291–7299.

[24] M. Contributors, "Mmpose, openmmlab pose estimation toolbox and benchmark," 2020.

[25] B. Fang, J. Co, and M. Zhang, "Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation," in Proceedings of the 15th ACM conference on embedded network sensor systems, 2017, pp. 1–13.

[26] P. Jiao, Y. Min, Y. Li, X. Wang, L. Lei, and X. Chen, "Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 20 676–20 686.

[27] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho, "Neural sign language translation based on human keypoint estimation," Applied sciences, vol. 9, no. 13, p. 2683, 2019.

[28] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian, and B. B. Chaudhuri, "A modified lstm model for continuous sign language recognition using leap motion," IEEE Sensors Journal, vol. 19, no. 16, pp. 7056–7063, 2019.

[29] Z. Wang and J. Zhang, "Continuous sign language recognition based on multi-part skeleton data," in 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021, pp. 1–8.

[30] W. Yang, J. Tao, and Z. Ye, "Continuous sign language recognition using level building based on fast hidden markov model," Pattern Recognition Letters, vol. 78, pp. 28–35, 2016.

[31] P. Dreuw, P. Steingrube, T. Deselaers, and H. Ney, "Smoothed disparity maps for continuous american sign language recognition," in Pattern Recognition and Image Analysis: 4th Iberian Conference, IbPRIA 2009 Póvoa de Varzim, Portugal, June 10-12, 2009 Proceedings 4. Springer, 2009, pp. 24–31.

[32] M. Jebali, A. Dakhli, and M. Jemni, "Vision-based continuous sign language recognition using multimodal sensor fusion," Evolving Systems, vol. 12, no. 4, pp. 1031–1044, 2021.

[33] Y. Ye, Y. Tian, M. Huenerfauth, and J. Liu, "Recognizing american sign language gestures from within continuous videos," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2064–2073.

[34] C. Zhang, Y. Tian, and M. Huenerfauth, "Multi-modality american sign language recognition," in 2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016, pp. 2881–2885.

[35] J. Shin, M. A. M. Hasan, A. S. M. Miah, K. Suzuki, and K. Hirooka, "Japanese sign language recognition by combining joint skeleton-based handcrafted and pixel-based deep learning features with machine learning classification," Comput. Model. Eng. Sci, vol. 139, no. 3, pp. 2605–2625, 2024.

[36] Z. Deng, Y. Leng, J. Chen, X. Yu, Y. Zhang, and Q. Gao, "Tms-net: A multi-feature multi-stream multi-level information sharing network for skeleton-based sign language recognition," Neurocomputing, vol. 572, p. 127194, 2024.

[37] W. Aditya, T. K. Shih, T. Thaipisutikul, A. S. Fitriajie, M. Gochoo, F. Utaminingrum, and C.-Y. Lin, "Novel spatio-temporal continuous sign language recognition using an attentive multi-feature network," Sensors, vol. 22, no. 17, p. 6452, 2022.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

[38] Z. Zhou, V. W. Tam, and E. Y. Lam, "Signbert: a bert-based deep learning framework for continuous sign language recognition," IEEE Access, vol. 9, pp. 161 669–161 682, 2021.

[39] ——, "A cross-attention bert-based framework for continuous sign language recognition," IEEE Signal Processing Letters, vol. 29, pp. 1818–1822, 2022.

[40] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for sign language recognition and translation," IEEE Transactions on Multimedia, vol. 24, pp. 768–779, 2021.

[41] G. Cortés, L. García, M. C. Benítez, and J. C. Segura, "Hmm-based continuous sign language recognition using a fast optical flow parameterization of visual information." in INTERSPEECH, 2006.

[42] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," IEEE Transactions on Multimedia, vol. 21, no. 7, pp. 1880–1891, 2019.

[43] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman, "Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. Springer, 2020, pp. 35–53.

[44] T. Ananthanarayana, P. Srivastava, A. Chintha, A. Santha, B. Landy, J. Panaro, A. Webster, N. Kotecha, S. Sah, T. Sarchet et al., "Deep learning methods for sign language translation," ACM Transactions on Accessible Computing (TACCESS), vol. 14, no. 4, pp. 1–30, 2021.

[45] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7784–7793.

[46] Y. Min, A. Hao, X. Chai, and X. Chen, "Visual alignment constraint for continuous sign language recognition," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 11 542–11 551.

[47] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, "Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos," IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 9, pp. 2306–2320, 2019.

[48] C. Wei, J. Zhao, W. Zhou, and H. Li, "Semantic boundary detection with reinforcement learning for continuous sign language recognition," IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 3, pp. 1138–1149, 2020.

[49] I. Papastratis, K. Dimitropoulos, D. Konstantinidis, and P. Daras, "Continuous sign language recognition through cross-modal alignment of video and text embeddings in a joint-latent space," IEEE Access, vol. 8, pp. 91 170–91 180, 2020.

[50] P. Xie, Z. Cui, Y. Du, M. Zhao, J. Cui, B. Wang, and X. Hu, "Multi-scale local-temporal similarity fusion for continuous sign language recognition," Pattern Recognition, vol. 136, p. 109233, 2023.

[51] P. Xie, M. Zhao, and X. Hu, "Pisltrc: Position-informed sign language transformer with content-aware convolution," IEEE Transactions on Multimedia, vol. 24, pp. 3908–3919, 2021.

[52] Z. Halim and G. Abbas, "A kinect-based sign language hand gesture recognition system for hearing-and speech-impaired: a pilot study of pakistani sign language," Assistive Technology, vol. 27, no. 1, pp. 34–43, 2015.

[53] T. Raghuveera, R. Deepthi, R. Mangalashri, and R. Akshaya, "A depth-based indian sign language recognition using microsoft kinect," Sādhanā, vol. 45, pp. 1–13, 2020.

[54] P. Kaushik, E. Jain, K. S. Gill, R. Chauhan, and H. S. Pokhariya, "Deep learning for sign language recognition utilizing vgg16 and resnet50 models," in 2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS). IEEE, 2024, pp. 1355–1359.

[55] Q. Han, Z. Huangfu, W. Min, T. Ding, and Y. Liao, "Sign language recognition based on skeleton and sk3d-residual network," Multimedia Tools and Applications, vol. 83, no. 6, pp. 18 059–18 072, 2024.

[56] R. Stamp, D. Cohn, H. Hel-Or, and W. Sandler, "Kinect-ing the dots: Using motion-capture technology to distinguish sign language linguistic from gestural expressions," Language and Speech, vol. 67, no. 1, pp. 255–276, 2024.

[57] M. A. Rahaman, K. U. Oyshe, P. K. Chowdhury, T. Debnath, A. Rahman, and M. S. I. Khan, "Computer vision-based six layered convneural network to recognize sign language for both numeral and alphabet signs," Biomimetic Intelligence and Robotics, vol. 4, no. 1, p. 100141, 2024.

[58] M. A. Rahaman, M. H. Ali, and M. Hasanuzzaman, "Real-time computer vision-based gestures recognition system for bangla sign language using multiple linguistic features analysis," Multimedia Tools and Applications, vol. 83, no. 8, pp. 22 261–22 294, 2024.

[59] G. S. Özcan, Y. C. Bilge, and E. Sümer, "Hand and pose based feature selection for zero-shot sign language recognition," IEEE Access, 2024.

[60] M. Islam, M. Aloraini, S. Aladhadh, S. Habib, A. Khan, A. Alabdulatif, and T. M. Alanazi, "Toward a vision-based intelligent system: A stacked encoded deep learning framework for sign language recognition," Sensors, vol. 23, no. 22, p. 9068, 2023.

[61] İ. Umut and Ü. C. Kumdereli, "Novel wearable system to recognize sign language in real time," 2024.

[62] K. Waldow, A. Fuhrmann, and D. Roth, "Deep neural labeling: Hybrid hand pose estimation using unlabeled motion capture data with color gloves in context of german sign language," in 2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR). IEEE, 2024, pp. 1–10.

[63] A. Salmankhah, A. Rajabi, N. Kheirmand, A. Fadaeimanesh, A. Tarabkhah, A. Kazemzadeh, and H. Farbeh, "Penslr: Persian end-to-end sign language recognition using ensembling," arXiv preprint arXiv:2406.16388, 2024.

[64] A. Tashakori, Z. Jiang, A. Servati, S. Soltanian, H. Narayana, K. Le, C. Nakayama, C.-l. Yang, Z. J. Wang, J. J. Eng et al., "Capturing complex hand movements and object interactions using machine learning-powered stretchable smart textile gloves," Nature Machine Intelligence, vol. 6, no. 1, pp. 106–118, 2024.

[65] C. Wang, W. Gao, and Z. Xuan, "A real-time large vocabulary continuous recognition system for chinese sign language," in Advances in Multimedia Information Processing—PCM 2001: Second IEEE Pacific Rim Conference on Multimedia Beijing, China, October 24–26, 2001 Proceedings 2. Springer, 2001, pp. 150–157.

[66] G. Fang, W. Gao, X. Chen, C. Wang, and J. Ma, "Signer-independent continuous sign language recognition based on srn/hmm," in Gesture and Sign Language in Human-Computer Interaction: International Gesture Workshop, GW 2001 London, UK, April 18–20, 2001 Revised Papers. Springer, 2002, pp. 76–85.

[67] W. Gao, G. Fang, D. Zhao, and Y. Chen, "Transition movement models for large vocabulary continuous sign language recognition," in Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings. IEEE, 2004, pp. 553–558.

[68] G. Yao, H. Yao, X. Liu, and F. Jiang, "Real time large vocabulary continuous sign language recognition based on op/viterbi algorithm," in 18th International Conference on Pattern Recognition (ICPR'06), vol. 3. IEEE, 2006, pp. 312–315.

[69] N. Tubaiz, T. Shanableh, and K. Assaleh, "Glove-based continuous arabic sign language recognition in user-dependent mode," IEEE Transactions on Human-Machine Systems, vol. 45, no. 4, pp. 526–533, 2015.

[70] M. Tuffaha, T. Shanableh, and K. Assaleh, "Novel feature extraction and classification technique for sensor-based continuous arabic sign language recognition," in Neural Information Processing: 22nd International Conference, ICONIP 2015, November 9-12, 2015, Proceedings, Part IV 22. Springer, 2015, pp. 290–299.

[71] K. Li, Z. Zhou, and C.-H. Lee, "Sign transition modeling and a scalable solution to continuous sign language recognition for real-world applications," ACM Transactions on Accessible Computing (TACCESS), vol. 8, no. 2, pp. 1–23, 2016.

[72] M. Hassan, K. Assaleh, and T. Shanableh, "Multiple proposals for continuous arabic sign language recognition," Sensing and Imaging, vol. 20, no. 1, p. 4, 2019.

[73] L. Zhang, Y. Zhang, and X. Zheng, "Wisign: Ubiquitous american sign language recognition using commercial wi-fi devices," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 11, no. 3, pp. 1–24, 2020.

[74] R.-H. Liang and M. Ouhyoung, "A real-time continuous gesture recognition system for sign language," in Proceedings third IEEE international conference on automatic face and gesture recognition. IEEE, 1998, pp. 558–567.

[75] W. Kong and S. Ranganath, "Towards subject independent continuous sign language recognition: A segment and merge approach," Pattern Recognition, vol. 47, no. 3, pp. 1294–1308, 2014.

[76] M. Hassan, K. Assaleh, and T. Shanableh, "User-dependent sign language recognition using motion detection," in 2016 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE, 2016, pp. 852–856.

[77] A. I. M. Thaim, N. Sazali, and K. Kadirgama, "Smart glove for sign language translation," Journal of Applied Mechanics and Materials, 2024. [Online]. Available: https://semarakilmu.com.my/journals/index.php/appl$_m ech/article/view/4581$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

[78] R. Gorka, A. K. Subramaniyan, and R. Velu, "Integrating advanced technologies in post-operative rehabilitation: 3d-knitting, 3d-printed electronics, and sensor-embedded textiles," in Digital Design and Manufacturing of Medical Devices and Systems. Springer, 2024, pp. 93–110.

[79] M. Hu, P. He, W. Zhao, X. Zeng, and J. He, "Machine learning-enabled intelligent gesture recognition and communication system using printed strain sensors," ACS Applied Materials & Interfaces, 2023. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/acsami.3c10846

[80] D. Omary and G. Mehta, "Multi-modal interactions of mixed reality framework," in 2024 IEEE 17th Dallas Circuits and Systems Conference (DCAS). IEEE, 2024, pp. 1–6.

[81] A. Ji, Y. Wang, X. Miao, T. Fan, B. Ru, L. Liu, R. Nie, and S. Qiu, "Dataglove for sign language recognition of people with hearing and speech impairment via wearable inertial sensors," Sensors, vol. 23, no. 15, p. 6693, 2023.

[82] S. Tateno, H. Liu, and J. Ou, "Development of sign language motion recognition system for hearing-impaired people using electromyography signal," Sensors, vol. 20, no. 20, p. 5807, 2020.

[83] D. Enikeev and S. Mustafina, "Recognition of sign language using leap motion controller data," in 2020 2nd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA). IEEE, 2020, pp. 393–397.

[84] D. Jain, H. Ngo, P. Patel, S. Goodman, K. Nguyen, R. Grossman-Kahn, L. Findlater, and J. Froehlich, "Soundwatch: Deep learning for sound accessibility on smartwatches," Communications of the ACM, vol. 65, no. 6, pp. 100–108, 2022.

[85] D. Ekiz, G. E. Kaya, S. Buğur, S. Güler, B. Buz, B. Kosucu, and B. Arnrich, "Sign sentence recognition with smart watches," in 2017 25th Signal Processing and Communications Applications Conference (SIU). IEEE, 2017, pp. 1–4.

[86] P. R. I. Gomes, M. S. d. Castro, and T. H. Nascimento, "Gesture recognition methods using sensors integrated into smartwatches: Results of a systematic literature review," in Proceedings of the XXII Brazilian Symposium on Human Factors in Computing Systems, 2023, pp. 1–11.

[87] P. S. Santhalingam, P. Pathak, H. Rangwala, and J. Kosecka, "Synthetic smartwatch imu data generation from in-the-wild asl videos," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 7, no. 2, pp. 1–34, 2023.

[88] H. Caballero-Hernandez, V. Muñoz-Jiménez, and M. A. Ramos-Corchado, "Translation of mexican sign language into spanish using machine learning and internet of things devices," Bulletin of Electrical Engineering and Informatics, vol. 13, no. 5, pp. 3369–3379, 2024.

[89] Y. Gu, H. Oku, and M. Todoh, "American sign language recognition and translation using perception neuron wearable inertial motion capture system," Sensors, vol. 24, no. 2, p. 453, 2024.

[90] E. Carlsson and A. Samuelsson, "Sign language detection using an apple watch and machine learning: Integrating apple watch sensors and machine learning for development of hand sign recognition application," 2024.

[91] H. Curtis and T. Neate, "Watch your language: Using smartwatches to support communication," in Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility, 2023, pp. 1–21.

[92] S. Sharma, R. Gupta, and A. Kumar, "Continuous sign language recognition using isolated signs data and deep transfer learning," Journal of Ambient Intelligence and Humanized Computing, pp. 1–12, 2023.

[93] K. Suri and R. Gupta, "Continuous sign language recognition from wearable imus using deep capsule networks and game theory," Computers & Electrical Engineering, vol. 78, pp. 493–503, 2019.

[94] S. Lang, M. Block, and R. Rojas, "Sign language recognition using kinect," in International Conference on Artificial Intelligence and Soft Computing. Springer, 2012, pp. 394–402.

[95] A. Akkar, S. Cregan, Y. Zeleke, C. Fahy, P. Sarkar, and T. K. Mohd, "Evaluation of accuracy of leap motion controller device," in International Conference on Intelligent Human Computer Interaction. Springer, 2021, pp. 391–402.

[96] X. Meng, L. Feng, X. Yin, H. Zhou, C. Sheng, C. Wang, A. Du, and L. Xu, "Sentence-level sign language recognition using rf signals," in 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC). IEEE, 2019, pp. 1–6.

[97] L. Ye, S. Lan, K. Zhang, and G. Zhang, "Em-sign: A non-contact recognition method based on 24 ghz doppler radar for continuous signs and dialogues," Electronics, vol. 9, no. 10, p. 1577, 2020.

[98] P. K. Chowdhury, K. U. Oyshe, M. A. Rahaman, T. Debnath, A. Rahman, and N. Kumar, "Computer vision-based hybrid efficient convolution for isolated dynamic sign language recognition," Neural Computing and Applications, pp 1–16, 2024.

[99] I. Kagirov, D. Ivanko, D. Ryumin, A. Axyonov, and A. Karpov, "Theruslan: Database of russian sign language," in Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 6079–6085.

[100] P. Ganesan, S. K. Jagatheesaperumal, S. Gaftandzhieva, R. Doneva et al., "Novel cognitive assisted adaptive frame selection for continuous sign language recognition in videos using convlstm." International Journal of Advanced Computer Science & Applications, vol. 15, no. 7, 2024.

[101] N. Aloysius, P. Nedungadi et al., "Continuous sign language recognition with adapted conformer via unsupervised pretraining," arXiv preprint arXiv:2405.12018, 2024.

[102] S. Srivastava, S. Singh, Pooja, and S. Prakash, "Continuous sign language recognition system using deep learning with mediapipe holistic," Wireless Personal Communications, pp. 1–14, 2024.

[103] S. Padmavathi et al., "Continuous sign language recognition using convolutional neural network," in 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE). IEEE, 2024, pp. 1–6.

[104] C. Lu, M. Kozakai, and L. Jing, "Sign language recognition with multimodal sensors and deep learning methods," Electronics, vol. 12, no. 23, p. 4827, 2023.

[105] M. S. Amin, S. T. H. Rizvi, and M. M. Hossain, "A comparative review on applications of different sensors for sign language recognition," Journal of Imaging, vol. 8, no. 4, p. 98, 2022.

[106] V. Ellappan, R. Sathiskumar, P. Saikrishna, C. Sivakumar, and R. Selvam, "Multimodal deep neural networks for robust sign language translation in real-world environments," in 2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE). IEEE, 2024, pp. 1–6.

[107] A. M. Buttar, U. Ahmad, A. H. Gumaei, A. Assiri, M. A. Akbar, and B. F. Alkhamees, "Deep learning in sign language recognition: a hybrid approach for the recognition of static and dynamic signs," Mathematics, vol. 11, no. 17, p. 3729, 2023.

[108] H. R. V. Joze and O. Koller, "Ms-asl: A large-scale data set and benchmark for understanding american sign language," arXiv preprint arXiv:1812.01053, 2018.

[109] J. Ye, W. Jiao, X. Wang, Z. Tu, and H. Xiong, "Cross-modality data augmentation for end-to-end sign language translation," arXiv preprint arXiv:2305.11096, 2023.

[110] U. Von Agris and K.-F. Kraiss, "Towards a video corpus for signer-independent continuous sign language recognition," Gesture in Human-Computer Interaction and Simulation, Lisbon, Portugal, vol. 11, 2007.

[111] Y. Jang, Y. Oh, J. W. Cho, D.-J. Kim, J. S. Chung, and I. S. Kweon, "Signing outside the studio: Benchmarking background robustness for continuous sign language recognition," arXiv preprint arXiv:2211.00448, 2022.

[112] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i Nieto, "How2sign: a large-scale multimodal dataset for continuous american sign language," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2735–2744.

[113] M. Alsulaiman, M. Faisal, M. Mekhtiche, M. Bencherif, T. Alrayes, G. Muhammad, H. Mathkour, W. Abdul, Y. Alohali, M. Alqahtani et al., "Facilitating the communication with deaf people: Building a largest saudi sign language dataset," Journal of King Saud University-Computer and Information Sciences, vol. 35, no. 8, p. 101642, 2023.

[114] R. Sreemathy, M. Turuk, S. Chaudhary, K. Lavate, A. Ushire, and S. Khurana, "Continuous word level sign language recognition using an expert system based on machine learning," International Journal of Cognitive Computing in Engineering, vol. 4, pp. 170–178, 2023.

[115] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li, "Improving sign language translation with monolingual data by sign back-translation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1316–1325.

[116] T. Starner, S. Forbes, M. So, D. Martin, R. Sridhar, G. Deshpande, S. Sepah, S. Shahryar, K. Bhardwaj, T. Kwok et al., "Popsign asl v1. 0: an isolated american sign language dataset collected via smartphones," Advances in Neural Information Processing Systems, vol. 36, 2024.

[117] F. Ronchetti, F. M. Quiroga, C. Estrebou, L. Lanzarini, and A. Rosete, "Lsa64: an argentinian sign language dataset," arXiv preprint arXiv:2310.17429, 2023.

[118] A. M. Martínez, R. B. Wilbur, R. Shay, and A. C. Kak, "Purdue rvl-slll asl database for automatic recognition of american sign language," in Proceedings. Fourth IEEE International Conference on Multimodal Interfaces. IEEE, 2002, pp. 167–172.

[119] P. Dreuw, D. Stein, and H. Ney, "Enhancing a sign language translation system with vision-based features," in Gesture-Based Human-Computer Interaction and Simulation: 7th International Gesture Workshop, GW 2007, Lisbon, Portugal, May 23-25, 2007, Revised Selected Papers 7. Springer, 2009, pp. 108–113.

[120] K. Assaleh, T. Shanableh, M. Fanaswala, F. Amin, and H. Bajaj, "Continuous arabic sign language recognition in user dependent mode," 2010.

[121] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. H. Piater, and H. Ney, "Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus." in LREC, vol. 9, 2012, pp. 3785–3789.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

[122] S. Abbas, H. Al-Barhamtoshy, and F. Alotaibi, "Towards an arabic sign language (arsl) corpus for deaf drivers," PeerJ Computer Science, vol. 7, p. e741, 2021.

[123] S. Hassan, M. Seita, L. Berke, Y. Tian, E. Gale, S. Lee, and M. Huenerfauth, "Asl-homework-rgbd dataset: An annotated dataset of 45 fluent and non-fluent signers performing american sign language homeworks," arXiv preprint arXiv:2207.04021, 2022.

[124] D. Uthus, G. Tanzer, and M. Georg, "Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus," Advances in Neural Information Processing Systems, vol. 36, 2024.

[125] S. Abbas, D. Alahmadi, and H. Al-Barhamtoshy, "Establishing a multimodal dataset for arabic sign language (arsl) production," Journal of King Saud University-Computer and Information Sciences, p. 102165, 2024.

[126] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep sign: Hybrid cnn-hmm for continuous sign language recognition." in BMVC, 2016, pp. 136–1.

[127] O. Koller, H. Ney, and R. Bowden, "Deep learning of mouth shapes for sign language," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 85–91.

[128] J. Pu, W. Zhou, H. Hu, and H. Li, "Boosting continuous sign language recognition via cross modality augmentation," in Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 1497–1505.

[129] Z. Zhang, J. Pu, L. Zhuang, W. Zhou, and H. Li, "Continuous sign language recognition via reinforcement learning," in 2019 IEEE international conference on image processing (ICIP). IEEE, 2019, pp. 285–289.

[130] Z. Niu and B. Mak, "Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. Springer, 2020, pp. 172–186.

[131] Z. Yang, Z. Shi, X. Shen, and Y.-W. Tai, "Sf-net: Structured feature network for continuous sign language recognition," arXiv preprint arXiv:1908.01341, 2019.

[132] J. Pu, W. Zhou, and H. Li, "Dilated convolutional network with iterative optimization for continuous sign language recognition." in IJCAI, vol. 3, 2018, p. 7.

[133] N. Liu, X. Li, B. Wu, Q. Yu, L. Wan, T. Fang, J. Zhang, Q. Li, and Y. Yuan, "A lightweight network-based sign language robot with facial mirroring and speech system," Expert Systems with Applications, vol. 262, p. 125492, 2025.

[134] M. Mnif, S. Sahnoun, M. Kaaniche, B. B. Atitallah, A. Fakhfakh, and O. Kanoun, "Ultra-fast edge computing approach for hand gesture classification based on eit measurements," in 2024 IEEE International Symposium on Robotic and Sensors Environments (ROSE). IEEE, 2024, pp. 1–7.

[135] O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4297–4305.

[136] Y. L. Gweth, C. Plahl, and H. Ney, "Enhanced continuous sign language recognition using pca and neural network features," in 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2012, pp. 55–60.

[137] N. Cihan Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Subunets: End-to-end hand shape and continuous sign language recognition," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 3056–3065.

[138] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7361–7369.

[139] X. Pei, D. Guo, and Y. Zhao, "Continuous sign language recognition based on pseudo-supervised learning," in Proceedings of the 2nd Workshop on Multimedia for Accessible Human Computer Interfaces, 2019, pp. 33–39.

[140] H. Zhou, W. Zhou, and H. Li, "Dynamic pseudo label decoding for continuous sign language recognition," in 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019, pp. 1282–1287.

[141] C. Wei, W. Zhou, J. Pu, and H. Li, "Deep grammatical multi-classifier for continuous sign language recognition," in 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM). IEEE, 2019, pp. 435–442.

[142] K. L. Cheng, Z. Yang, Q. Chen, and Y.-W. Tai, "Fully convolutional networks for continuous sign language recognition," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV. Springer, 2020, pp. 697–714.

[143] I. Papastratis, K. Dimitropoulos, and P. Daras, "Continuous sign language recognition through a context-aware generative adversarial network," Sensors, vol. 21, no. 7, p. 2437, 2021.

[144] A. Hao, Y. Min, and X. Chen, "Self-mutual distillation learning for continuous sign language recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11 303–11 312.

[145] Q. Zhu, J. Li, F. Yuan, and Q. Gan, "Multiscale temporal network for continuous sign language recognition," Journal of Electronic Imaging, vol. 33, no. 2, pp. 023 059–023 059, 2024.

[146] L. Hu, L. Gao, Z. Liu, and W. Feng, "Self-emphasizing network for continuous sign language recognition," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 1, 2023, pp. 854–862.

[147] J. Zheng, Y. Wang, C. Tan, S. Li, G. Wang, J. Xia, Y. Chen, and S. Z. Li, "Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 23 141–23 150.

[148] Z. Cui, W. Zhang, Z. Li, and Z. Wang, "Spatial–temporal transformer for end-to-end sign language recognition," Complex & Intelligent Systems, vol. 9, no. 4, pp. 4645–4656, 2023.

[149] L. Hu, L. Gao, Z. Liu, and W. Feng, "Scalable frame resolution for efficient continuous sign language recognition," Pattern Recognition, vol. 145, p. 109903, 2024.

[150] R. Zuo and B. Mak, "Improving continuous sign language recognition with consistency constraints and signer removal," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 20, no. 6, pp. 1–25, 2024.

[151] S. Danish, A. Khan, L. M. Dang, M. Alonazi, S. Alanazi, H.-K. Song, and H. Moon, "Metaverse applications in bioinformatics: A machine learning framework for the discrimination of anti-cancer peptides," Information, vol. 15, no. 1, p. 48, 2024.

[152] H. Brock, I. Farag, and K. Nakadai, "Recognition of non-manual content in continuous japanese sign language," Sensors, vol. 20, no. 19, p. 5621, 2020.

[153] J. Forster, C. Oberdörfer, O. Koller, and H. Ney, "Modality combination techniques for continuous sign language recognition," in Pattern Recognition and Image Analysis: 6th Iberian Conference, IbPRIA 2013, Funchal, Madeira, Portugal, June 5-7, 2013. Proceedings 6. Springer, 2013, pp. 89–99.

[154] F. Wei and Y. Chen, "Improving continuous sign language recognition with cross-lingual signs," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 23 612–23 621.

[155] R. Li and L. Meng, "Multi-view spatial-temporal network for continuous sign language recognition," arXiv preprint arXiv:2204.08747, Apr. 2022, available: https://arxiv.org/abs/2204.08747.

[156] U. Von Agris, J. Zieren, U. Canzler, B. Bauer, and K.-F. Kraiss, "Recent developments in visual sign language recognition," Universal Access in the Information Society, vol. 6, pp. 323–362, 2008.

[157] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need. advances in neural information processing systems," Advances in neural information processing systems, vol. 30, no. 2017, 2017.

[158] L. Guo, W. Xue, Q. Guo, B. Liu, K. Zhang, T. Yuan, and S. Chen, "Distilling cross-temporal contexts for continuous sign language recognition," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 10 771–10 780.

[159] "Local context-aware self-attention for continuous sign language recognition" $\mathcal{O}$

[160] Z. Huang, W. Xue, Y. Zhou, J. Sun, Y. Wu, T. Yuan, and S. Chen, "Dual-stage temporal perception network for continuous sign language recognition," The Visual Computer, pp. 1–16, 2024.

[161] S. Alyami, H. Luqman, and M. Hammoudeh, "Isolated arabic sign language recognition using a transformer-based model and landmark keypoints," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 23, no. 1, pp. 1–19, 2024.

[162] Z. Liu, J. Wu, Z. Shen, X. Chen, Q. Wu, Z. Gui, L. Senhadji, and H. Shu, "Improving end-to-end sign language translation with adaptive video representation enhanced transformer," IEEE Transactions on Circuits and Systems for Video Technology, 2024.

[163] Y. Du, T. Peng, and X. Hu, "Beyond granularity: Enhancing continuous sign language recognition with granularity-aware feature fusion and attention optimization," Applied Sciences, vol. 14, no. 19, p. 8937, 2024.

[164] H. Ranjbar and A. Taheri, "Continuous sign language recognition using intra-inter gloss attention," arXiv preprint arXiv:2406.18333, 2024.

[165] M. Mnif, S. Sahnoun, M. Kaaniche, B. B. Atitallah, A. Fakhfakh, and O. Kanoun, "Ultra-fast edge computing approach for hand gesture classification based on eit measurements," in 2024 IEEE International Symposium on Robotic and Sensors Environments (ROSE), 2024, pp. 1–7.

[166] B. Bauer, H. Hienz, and K.-F. Kraiss, "Video-based continuous sign language recognition using statistical methods," in Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, vol. 2. IEEE, 2000, pp. 463–466.

[167] Q. Yuan, W. Geo, H. Yao, and C. Wang, "Recognition of strong and weak connection models in continuous sign language," in 2002 International Conference on Pattern Recognition, vol. 1. IEEE, 2002, pp. 75–78.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

[168] R. Yang, S. Sarkar, and B. Loeding, "Enhanced level building algorithm for the movement epenthesis problem in sign language recognition," in 2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007, pp. 1–8.
[169] ——, "Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming," IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 3, pp. 462–477, 2009.
[170] B. Bauer and K. Karl-Friedrich, "Towards an automatic sign language recognition system using subunits," in Gesture and Sign Language in Human-Computer Interaction: International Gesture Workshop, GW 2001 London, UK, April 18–20, 2001 Revised Papers. Springer, 2002, pp. 64–75.
[171] P. Buehler, A. Zisserman, and M. Everingham, "Learning sign language by watching tv (using weakly aligned subtitles)," in 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 2961–2968.
[172] T.-Y. Pan, L.-Y. Lo, C.-W. Yeh, J.-W. Li, H.-T. Liu, and M.-C. Hu, "Real-time sign language recognition in complex background scene based on a hierarchical clustering classification method," in 2016 IEEE second international conference on multimedia big data (BigMM). IEEE, 2016, pp. 64–67.
[173] K. Tripathi and N. B. G. Nandi, "Continuous indian sign language gesture recognition and sentence formation," Procedia Computer Science, vol. 54, pp. 523–531, 2015.
[174] J. Zhang, W. Zhou, and H. Li, "A threshold-based hmm-dtw approach for continuous sign language recognition," in Proceedings of international conference on internet multimedia computing and service, 2014, pp. 237–240.
[175] ——, "A new system for chinese sign language recognition," in 2015 IEEE China summit and international conference on signal and information processing (ChinaSIP). IEEE, 2015, pp. 534–538.
[176] R. Elakkiya and K. Selvamani, "Subunit sign modeling framework for continuous sign language recognition," Computers & Electrical Engineering, vol. 74, pp. 379–390, 2019.
[177] M. F. Tolba, A. Samir, and M. Aboul-Ela, "Arabic sign language continuous sentences recognition using pcnn and graph matching," Neural Computing and Applications, vol. 23, pp. 999–1010, 2013.
[178] S.-H. Yu, C.-L. Huang, S.-C. Hsu, H.-W. Lin, and H.-W. Wang, "Vision-based continuous sign language recognition using product hmm," in The first Asian conference on pattern recognition. IEEE, 2011, pp. 510–514.
[179] S. Sarkar, B. Loeding, R. Yang, S. Nayak, and A. Parashar, "Segmentation-robust representations, matching, and modeling for sign language," in CVPR 2011 Workshops. IEEE, 2011, pp. 13–19.
[180] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos, "Hand tracking and affine shape-appearance handshape sub-units in continuous sign language recognition," in Trends and Topics in Computer Vision: ECCV 2010 Workshops, Heraklion, Crete, Greece, September 10-11, 2010, Revised Selected Papers, Part I 11. Springer, 2012, pp. 258–272.
[181] D. Kelly, J. Reilly Delannoy, J. McDonald, and C. Markham, "A framework for continuous multimodal sign language recognition," in Proceedings of the 2009 International Conference on Multimodal Interfaces, 2009, pp. 351–358.
[182] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney, "Speech recognition techniques for a sign language recognition system," hand, vol. 60, p. 80, 2007.
[183] P. N. Vassilia and K. G. Margaritis, "Multimodal continuous recognition system for greek sign language using various grammars," in Advances in Artificial Intelligence: 4th Hellenic Conference on AI, SETN 2006, Heraklion, Crete, Greece, May 18–20, 2006. Proceedings. Springer, 2006, pp. 584–587.
[184] C. Vogler and D. Metaxas, "A framework for recognizing the simultaneous aspects of american sign language," Computer Vision and Image Understanding, vol. 81, no. 3, pp. 358–384, 2001.
[185] A. Choudhury, A. Kumar Talukdar, M. Kamal Bhuyan, and K. Kumar Sarma, "Movement epenthesis detection for continuous sign language recognition," Journal of Intelligent Systems, vol. 26, no. 3, pp. 471–481, 2017.
[186] I. Infantino, R. Rizzo, and S. Gaglio, "A framework for sign language sentence recognition by commonsense context," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 37, no. 5, pp. 1034–1039, 2007.
[187] D. A. Kumar, A. S. C. S. Sastry, P. V. V. Kishore, and E. K. Kumar, "Indian sign language recognition using graph matching on 3d motion captured signs," Multimedia Tools and Applications, vol. 77, no. 24, pp. 32 063–32 091, 2018.
[188] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," Expert Systems with Applications, vol. 164, p. 113794, 2021.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3554046

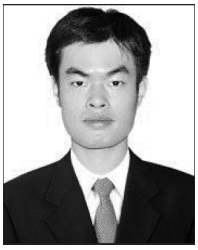
ASMA KHAN received her Bachelor's degree from Islamia University Peshawar, Pakistan. She is currently pursuing a Master's degree at Sejong University, Korea, where she is also a research assistant at the CVPR Lab. Her research interests include image processing, generative AI, medical image analysis, and computer vision.

SEYONG JIN received a B.Eng. degree in Industrial Engineering from Induk University, South Korea, in 2022. He is currently a master's student in the Department of Artificial Intelligence at the CVPR (Computer Vision and Pattern Recognition) Lab, Sejong University, South Korea, where he is conducting research in computer vision. His research interests include computer vision, generative AI, and data analysis.

GEON-HEE LEE received his bachelor's degree in software application and his double major in virtual reality from Kangnam University in 2024. He's currently pursuing his master's degree in artificial intelligence from Sejong University in 2024. His research interests include virtual reality, human-computer interaction, convolutional neural networks, and data analysis.

GUL E ARZU received a B.S. Degree from National Textile University in Software Engineering, Pakistan, in 2024. She is currently pursuing a Master's degree in the Department of Computer Science and Engineering at the Korean University of Sejong. Her research interests include AI and Computer Vision.

TAN N. NGUYEN received the M.E. degree from the Ho Chi Minh City University of Technology (HCMUT), Vietnam, and the Ph.D. degree from Sejong University, South Korea, in 2019. He is currently an Assistant Professor with the Department of Architectural Engineering, Sejong University. His research interests include developing numerical methods, application robust methods to model structure considering modern materials, and new theoretical models. In addition, he also investigates highly nonlinear problems, instability of structures, and applies deep learning to structural analysis.



L. MINH DANG received a B.S. degree in information systems from the University of Information Technology, VNU HCMC, Vietnam, in 2016. He is currently pursuing a PhD degree in computer science at Sejong University, Seoul, South Korea. In 2017, he joined the Computer Vision Pattern Recognition Laboratory. His current research interests include computer vision, natural language processing, and artificial intelligence.

WOONG CHOI received the B.E. and M.E. degrees in control and instrumentation engineering from Chosun University, Gwangju, in 1998 and 2000, respectively, and the Ph.D. degree in intelligent systems science from Tokyo Institute of Technology, Tokyo, Japan, in 2005. From 2005 to 2010, he was with Ritsumeikan University. From 2010 to 2022, he was with the National Institute of Technology, Gunma College. He is currently an Associate Professor with the Division of ICT Convergence Engineering, at Kangnam University. His research interests include virtual reality, human-computer interaction, and related fields.

HYEONJOON MOON received the B.S. degree in electronics and computer engineering from Korea University, in 1990, and the M.S. and Ph.D. degrees in electrical and computer engineering from the State University of New York at Buffalo, in 1992 and 1999, respectively. From January 1996 to October 1999, he was a Senior Researcher at the Electro-Optics/Infrared Image Processing Branch at the U.S. Army Research Laboratory (ARL), Adelphi, MD, USA. He developed a face recognition system evaluation methodology based on the Face Recognition Technology (FERET) Program. From November 1999 to February 2003, he was a Principal Research Scientist at Viisage Technology, Littleton, MA, USA. His main interest in research and development is real-time facial recognition systems for access control, surveillance, and big database applications. He has an extensive background in still image and real-time video-based computer vision and pattern recognition. Since March 2004, he has been with the Department of Computer Science and Engineering, Sejong University, where he is currently a Professor and the Chairperson. His current research interests include image processing, biometrics, artificial intelligence, and machine learning.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/