Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 10(9): 20084–20112.
DOI: 10.3934/math.2025898
Received: 30 June 2025
Revised: 14 August 2025
Accepted: 25 August 2025
Published: 02 September 2025

Research article

Development of deep learning-based models highlighting the significance of non-manual features in sign language recognition

Maher Jebali[1,*], **Lamia Trabelsi**[1], **Haifa Harrouch**[1], **Rabab Triki**[2] **and Shawky Mohamed**[1]

[1] Computer Science Department, Applied College, University of Ha'il, P.O. Box 2440, Hail City 55476, Saudi Arabia

[2] Management Information Systems Department, Applied College, University of Ha'il, P.O. Box 2440, Hail City 55476, Saudi Arabia

* **Correspondence:** Email: maher.jbeli@gmail.com.

Abstract: The quality of recognition systems for sign language utterances has significantly improved in recent years for the benefit of hearing-impaired people. Nevertheless, research initiatives frequently overlook particular linguistic characteristics of sign languages, such as nonmanual utterances. Nonmanual articulations are an essential element of all sign languages. They encompass not only many elements of facial expression but also ocular gaze, as well as the position of the head and the upper body movements. This study assessed the efficacy of a recognition system utilizing a single video camera about nonmanual features. We presented a two-stage pipeline utilizing 2D body joint locations derived from red, green, blue (RGB) camera data. The initial pipeline examined heteroscedastic head pose network (HHP-net), a technique for calculating head direction from individual frames utilizing a HHP-net to ascertain an individual's head position from a limited number of head keypoints. In the second pipeline, we presented a kinematic hand pose rectification method for enforcing constraints to enhance the realism of hand skeletal representations. Next, we examined spatial-temporal graph convolutional networks and multi-modal long short-term memory to use multi-articulatory information (e.g., body, right hand, and left hand) for the recognition of sign glosses. We trained an spatiotemporal graph convolutional network (ST-GCN) model to learn representations from the upper body and hands. The suggested method was subsequently assessed using two publicly available datasets, the RWTH-PHOENIX-Weather and the Chinese sign language (CSL), featuring a range of nonmanual utterances. By examining several data forms and network characteristics, we identified word segments with 92.8% accuracy from the underlying body joint movement data. The research showed a 17.8% word error rate for whole sentence predictions, a significant improvement from ground truth scores based on labeling that ignored nonmanual content.

Keywords: CNN; CTC; recurrent neural network; sign language recognition; head pose

Mathematics Subject Classification: 37M10

1. Introduction

Automatic sign language recognition (SLR) from videos represents a significant research challenge, attracting substantial interest in recent years, improving accessibility for those who are hearing impaired, and being integrated into sign language (SL) educational applications [1–3]. Nonetheless, SLR is a complex endeavor owing to the numerous closely linked manual and nonmanual modalities, comprising mouthing patterns, body inclination, eye gaze, handshapes, eyebrow action, head orientation, and shoulder movement [4]. These challenges are significantly more complex for SLR in a signer-independent context [5, 6] because of the intrinsic heterogeneity in articulation across signers. Many studies have been undertaken in the field of SLR. A recent study has introduced a novel version of the sign structure to facilitate the perception of subunits akin to phonology in the language that is spoken. The Stokoe model can be utilized to delineate SL constituents, which depend on attributes such as motion, alignment, and shape. The direction of the palm and the finger arrangement can be represented by the hand form, which predominantly conveys a particular significance and is depicted inside the frame. This hand configuration can be either static or dynamic. The fixed position can convey the complete significance of the sign within a unique frame that simultaneously represents ongoing signing [7]. The successive frames must convey the complete significance of the dynamic gesture, referred to as the motion-hold pattern. The correlation between hand movement and positions has been estimated for two-handed signs, whereas other methods are adequate for one-handed signals. Over the past ten years, deep learning (DL) techniques, including 2D and 3D convolutional neural networks (2D-CNNs and 3D-CNNs), have been introduced to address SLR [8–10]. 3D-CNNs immediately correlate the RGB data of a video to a label. Nonetheless, 3D-CNNs typically require a substantial quantity of parameters to achieve dimensional representations, augmenting the model's computational complexity. This is inappropriate in situations where the model must operate directly on mobile devices to prevent the transmission of private information to the internet. Skeleton-based features facilitate the development of less intricate heterogeneous autoregressive (HAR) models, demonstrating resilience to variations in backdrop, illumination, and physical characteristics [11]. CNNs, graph neural networks (GNNs), and transformers have been proposed in the literature for processing skeleton-based information. CNN-based techniques often transform a skeletal sequence into an image, enabling the application of 2D-CNNs for image processing. GNNs encapsulate the spatiotemporal attributes of a skeleton sequence within a graph data structure and manipulate it directly. In contrast, transformers treat a skeleton series as a singular vector. Unlike traditional HAR, the body position joints alone are inadequate for classifying SL glosses. According to [4], facial expressions hold significant importance in SLR, comparable to that of the hands. This study considers both body pose joints, head pose, and detailed hand joints in developing the foundational skeleton graph. We suggest a multimodal SLR architectural design that leverages multi-articulatory spatiotemporal data inherent in both manual and nonmanual elements of SL.

1) Regarding the nonmanual component, we examine methods for effectively and efficiently estimating head posture, proposing a lightweight and flexible neural network that determines head

orientation as a triplet of yaw, pitch, and roll angles. We utilize keypoints derived from 2D human posture estimators as input.

2) We examine an extractor module of spatiotemporal features for manual components, designed to derive visual representations from multiple modalities (left hand, right hand, and upper body) utilizing ST-GCNs and CNNs. This module is employed alongside a temporal modeling component employing multimodal long short-term memory (MM-LSTM) that concurrently acquires knowledge of temporal interactions among the various modalities. Both previously mentioned models are trained independently, and their outputs are integrated using an ensemble module that utilizes the results from the final fully connected layer.

2. Related work

Numerous researchers have been developing SLR systems utilizing various statistical and mathematical models in addition to diverse algorithms for machine learning [12–15]. In [16], the authors discussed a method for recognizing finger spelling that employs principle component analysis (PCA) and red, green, blue – depth (RGB-D) data, utilizing sparse auto-encoder-based methods for selecting features. They evaluated their identification procedure utilizing support vector machines (SVM) on 24 American SL alphabets, achieving a precision of 99.10%. In [17], the authors utilized Kinect to develop SLR systems. The system interprets speeded-up robust features (SURF) descriptors, velocity, and distance according to palm orientation and motion. The system achieved an accuracy of 80% and was evaluated on 34 sign words utilizing SVM. Junfu et al. introduced a multi-stage strategy in the first stage, specifically, a sequence modeling method called connectionist temporal classification (CTC) in [18]. In the subsequent phase, they utilized a feature learning model: a 3D convolutional residual network (3D-ResNet). They collaboratively trained the LSTM model with CTC, employing a gentle dynamic time warping alignment constraint. They employed the RWTH-PHOENIX-Weather and Chinese sign language (CSL) benchmarks to assess their model, attaining word error rates (WER) of 36.7% and 32.7%, respectively.

Koller et al. combined two distinct features in [19], including hidden Markov model (HMM) and CNN sequence characteristics. They subsequently employed a hybrid model based on CNN and HMM for categorization. The assessment of their model encompassed three SL benchmarks, achieving superior precision and decreasing the WER by 20%. Huang et al. presented a systematic literature review model in [20] that incorporates spatiotemporal information and the selection of salient aspects utilizing 3D-CNN and an attention mechanism. Their pattern was assessed using two standard datasets, ChaLearn-14 and CSL, achieving a precision of 95.30% on the ChaLearn benchmark. Pigou et al. suggested a temporal feature based on a pooling identification technique utilizing an SL video dataset in [21]. Sincan et al. devised a hybrid methodology for efficient extraction of features by integrating LSTM, CNN, and a feature pooling technique in [22]. They employed visual geometry group (VGG-16) as a pretrained pattern combined with a CNN-based synchronous approach for the depth and RGB video dataset, attaining 93.15% precision with an Italian SL benchmark. Nonetheless, these systems encounter significant challenges in attaining high-achievement precision and efficiency due to duplicated backgrounds, hand occlusion, fluctuations in light, and hand orientation management. To address the issue, researchers utilized skeletal points from the image rather than directly employing the image's pixels for hand gesture identification. Currently, researchers employed several spatial types of 3D cameras to gather skeletal keypoints from the SL picture data. Researchers have created various applications for extracting skeletal keypoints, including Alphapose, OpenPose, MediaPipe, and MMpose. Shin et al. utilized the geometrical architecture in [23] to derive distance and angular features from the 21 hand keypoints, which were obtained using the MediaPipe framework for the American sign language (ASL) dataset. The primary constraints of the system reliant on hand-crafted features are ineffective features and insufficient generalization. To address this restriction, numerous studies utilized end-to-end DL methods to categorize hand motions based on raw skeletal data [24–26]. These present systems solely account for the spatial information in the frame, neglecting motion and temporal characteristics. In many instances, they are unable to discern intricate connections across the joints.

Yan et al. presented in [27] a graph-based modeling utilizing a GCN for skeleton data analysis. This graph-based methodology has been refined and utilized by numerous additional academics [28–31]. [32] introduced a decoupling GCN to augment the model's capability without elevating the computational expense. A ResNet-based GCN design was introduced in [33] to improve model achievement while reducing computational expense. Nonetheless, the skeleton-based methodology remains little investigated. Furthermore, Al-Hammadi et al. [34] employed a graph convolutional neural network analogous to ST-GCN and MediaPipe [35] to gather hand and body joints for the representation of a signer's skeletal information. While their method successfully collects local information, the lack of appearance data in small segments results in a notable decline in recognition precision when addressing bigger datasets.

Recently, Geetha et al. implemented in [36] an innovative preprocessing method to down-sample streams of video, guaranteeing compatibility with diverse rates of frame across various devices. For the first time, they pretrained the fundamental elements of their network on domain-specific Indian sign language (ISL) data. The CNN is pretrained with ISL word videos, whereas the Transformer is pretrained with Mediapipe pose estimates derived from ISL videos. This initial training adeptly catches the intricacies of hand shapes and body gestures distinctive to ISL, markedly improving sentence identification. Their system attained a WER of 19% on the continuous ISL dataset, illustrating its efficacy for real-time ISL recognition. Guan et al. created a multi-stream keypoint attention network in [37] to show a setup of keypoints produced by a keypoint estimator that is available. To enhance communication across many streams, they explored different approaches, including keypoint merging schemes, head fusion, and self-distillation. The final design is called multi-stream keypoint attention network for sign language recognition (MSKA-SLR), which is improved for sign language translation (SLT) by adding an extra translation network. They conducted extensive trials using established datasets such as Phoenix-2014, Phoenix-2014T, and CSL-Daily to demonstrate the effectiveness of their methods. Yu et al. presented in [38] a dual-stage temporal perception module (DTPM) that combines multi-scale temporal convolutions with transformer-based global modeling to improve temporal feature extraction in continuous sign language recognition (CSLR). This method tackles the difficulties presented by seamless transitions and diverse temporal scales in sign language videos, which conventional fixed receptive field techniques fail to capture well. The DTPM enhances recognition accuracy by hierarchically integrating local and global temporal features. Experimental findings indicate that our strategy surpasses current state-of-the-art models on benchmark CSLR datasets.

This research seeks to improve the precision of word-level SLR by integrating nonmanual components. This would enable our system to identify grammatically important nonmanual manifestations in continuous signing and effectively distinguish between linguistically important expressions that exhibit subtle visual variations.

3. Method

Figure 1 illustrates our proposed method that addresses the problem of CSLR by trying to predict a series of glosses from a set of video frames. The model receives a sequence of RGB video frames from which two categories of features are derived: (1) facial keypoints along with their confidence scores for representing nonmanual components and (2) body and hand keypoints for representing manual articulation. The architecture employs a dual-stream configuration.

The initial stream analyzes face keypoints with the heteroscedastic head pose network (HHP-net) [39]. This network accepts a collection of 2D facial keypoint coordinates and their corresponding detection confidence scores. It produces three continuous values denoting head orientation angles (yaw, pitch, and roll), each with an uncertainty estimate. These attributes encompass nonmanual linguistic elements, including emphasis and prosody.

The secondary stream models manual characteristics and dynamics with spatiotemporal data. The system accepts 2D keypoints from the upper torso and both hands (except the head) and utilizes ST-GCN to analyze spatial and temporal correlations between frames. The resulting features are passed through layers that look at time, an attention system, and an MM-LSTM, which learns how different types of data relate over time. The output is a comprehensive, temporally synchronized depiction of manual features for gloss prediction.

The head posture predictions from the HHP-net are integrated with the MM-LSTM output to enhance the final gloss classification layer. This integration allows the model to utilize both manual articulations and nonmanual cues for enhanced recognition precision. The downstream recognition modules are trained jointly to minimize a combination of CTC loss and distillation loss. However, the overall pipeline is modular, as it relies on externally extracted 2D keypoints that are not optimized during training.

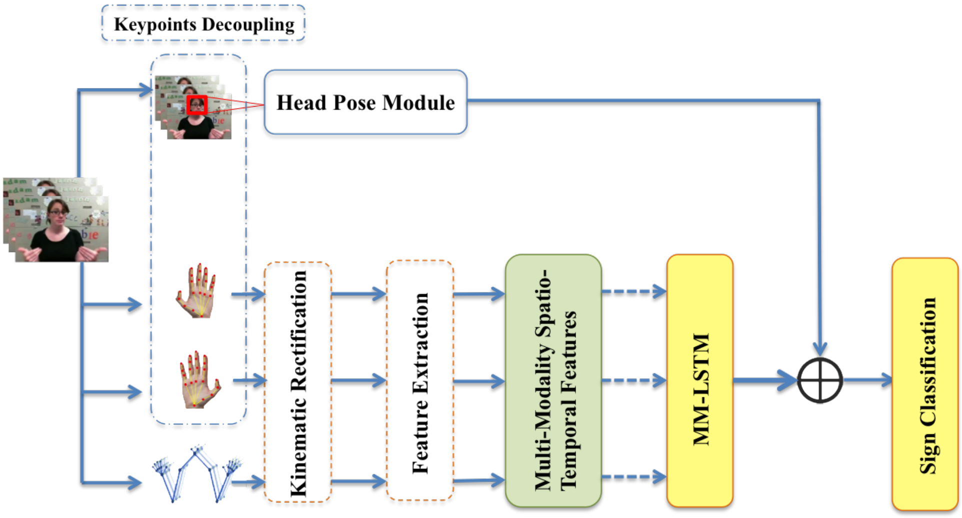


Figure 1. Overview of our proposed multi-modality SLR approach.

3.1. Stream 1: Head pose estimation

3.1.1. Architecture of head pose estimation

Figure 2 presents a summary of our head pose structure. We define the network input as a triplet of vectors $v_1 = [v_1^1, ..., v_1^n]$, $v_2 = [v_2^1, \ldots, v_2^n]$, and $c = [c^1, ..., c^n]$ representing the placements and confidence levels of n keypoints that delineate a face.

Facial landmarks are extracted per-frame using the MediaPipe face mesh detector, which returns a set of 2D facial keypoints together with per-landmark confidence (visibility) scores. In our implementation we use the full face mesh output (468 landmarks); each landmark i in a frame provides a 2D coordinate (x_i, y_i) and a detection confidence score $c_i \in [0, 1]$. The vector $c = [c_1, \ldots, c_n]$ therefore represents the per-landmark confidence values used by the confidence gated unit (CGU) to weigh the contribution of each landmark during head-pose regression.

The entry vectors undergo initial processing in separate streams using 5-channel one-dimensional convolutions, succeeded by a leaky rectified linear unit (ReLU) activation for v_1 and v_2 to mitigate disappearing gradient problems, and a sigmoid activation for the confidence vector c to regulate the influence of varying confidence levels smoothly. The exits of the one-dimensional convolutional layers are flattened to derive v_1^*, v_2^*, and c^* from the separate streams. The vectors are subsequently merged employing element-wise multiplication to yield 2 vectors, $x_1 = v_1^* \bigotimes c^*$ and $x_2 = v_2^* \bigotimes c^*$, in conformity with the principles of the CGU introduced in [40]. The CGU comprises ReLU and sigmoid activation functions. The ReLU and sigmoid functions are employed to the coordinates (v_1^i or v_2^i) and the confidence (c_i), respectively, and their exits are then amplified. The CGU simulates the function of a gate, regulated by confidence, as it yields results close to 0 when confidence is low.

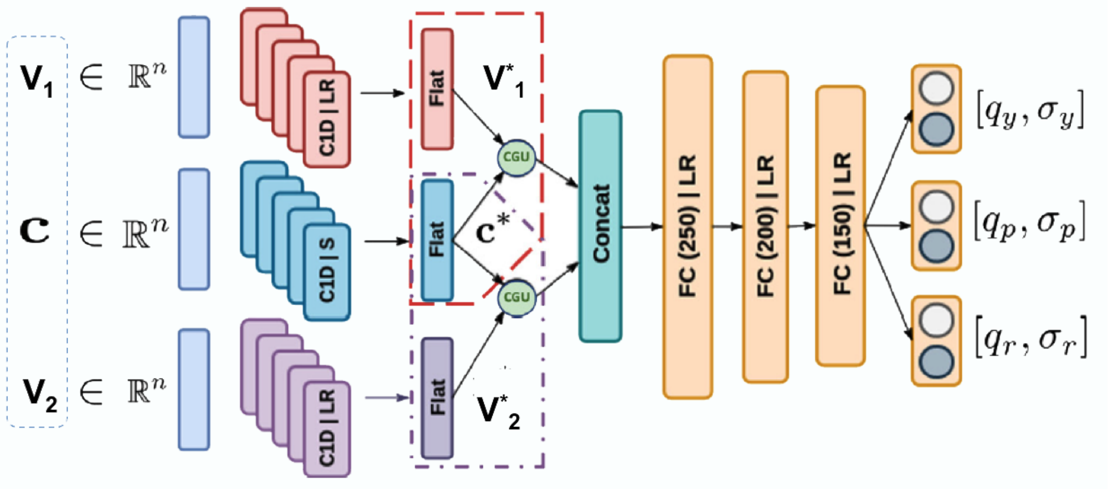


Figure 2. A visual illustration of our head pose structre.

The two gated outputs x_1 and x_2 are merged to form a unique vector, which is then input into the intermediary section of the structure, where a series of 3 fully connected layers with 250, 200, and 150 neurons, respectively, is utilized. Every layer incorporates a LeakyReLU as a nonlinear activation function to prevent vanishing gradients. Three exit layers provide the predicted angles, each linked to its corresponding uncertainty value.

3.1.2. Heteroscedastic neural network multi-assignment loss

We develop a multi-assignment loss function that integrates heteroscedastic aleatory inaccuracy for network training. In contrast to standard neural networks, a heteroscedastic neural network offers an estimation of the inaccuracy associated with each prediction. This is especially beneficial for capturing noise within input observations; in this context, noise pertains to the intrinsic inaccuracies in keypoint localization, which may be influenced by challenging perspectives or occlusions. Certain stances are inherently louder and more susceptible to self-occlusions. This form of inaccuracy can be acquired as a function of the data, resulting in an output that encompasses not solely the 3 angles (yaw, pitch, roll), represented as a vector $q = [q_y, q_p, q_r]$, but also the corresponding inaccuracy values linked to them $\sigma = [\sigma_y, \sigma_p, \sigma_r]$. We will now elucidate the derivation of the multi-assignment loss function, first with a basic heteroscedastic loss formulation.

For the advantage of simplicity, we consider a basic regression issue in which we aim to assess a function $f_\omega : \mathbb{R}^n \leftarrow \mathbb{R}$ such that:

$$y = f_\omega(X) + \varepsilon(X). \tag{3.1}$$

The output constitutes the aggregate of the function $f_\omega(X)$, which is contingent upon specific parameters ω and the input x, and $\varepsilon(X)$, which represents the noise solely dependent on the input x [41]. To measure inaccuracy, we train a model using a training set $X = \{(x_i, y_i)\}_{i=1}^{\iota}$ to estimate the variance and the mean of a purpose distribution through a maximum-likelihood calculation of a neural network. For this objective, we must presume that the mistakes follow a normal distribution, $\varepsilon(X_i) \sim \mathcal{N}(0, (\sigma(x_i)^2))$; therefore, the probability for each point x_i is $p(y_i \| x_i; \omega)$.

Here, $\sigma(x_i)^2$ denotes the variance and y_i represents the mean of this distribution. Therefore, from a constructional perspective, alongside the assessment of the y_i, the heteroscedastic neural network structure is required to be also adjusted to provide an estimate of the variance, which quantifies the inaccuracy linked to the prediction based on the noise present in the training specimens. Observe that inaccuracy is contingent upon the input; for instance, if the noise is uniformly distributed across all input values, the inaccuracy will remain unchanging.

An alternate formulation utilizing the variable transformation $\hat{s} = \log \hat{\sigma}(x_i)^2$ can be employed to mitigate excessive inaccuracies throughout training [42], resulting in the ultimate problem formulation:

$$min_\omega \frac{1}{n} \sum_{i=1}^{\iota} \frac{1}{2} e^{-\hat{s}_i} (y_i - \hat{f}_\omega(x_i))^2 + \frac{1}{2} \hat{s}_i. \tag{3.2}$$

We now delineate the generic formulation of the heteroscedastic loss function pertinent to our case. We expand the model in Eq (3.2) to depict a multi-assignment situation whereby the output comprises three components: $q = [q_y, q_p, q_r]$ and the inaccuracies associated with are $\sigma = [\sigma_y, \sigma_p, \sigma_r]$. We denote the 3 angles: yaw (y), pitch (p), and roll (r). Consequently, the individual assignments within the multi-assignment framework pertain to the distinct assessment of the 3 angles. We assess them by optimizing a distinct function and using their collaborative advantages. The input consists of x_1, x_2, and c, representing the coordinates of the detected facial keypoints and the confidence level in their detection, respectively. We are now able to formulate the multi-assignment heteroscedastic loss function.

This formulation yields a data-driven inaccuracy assessment for every angle, a weight for each sub-loss. Inaccuracy can enhance the robustness of the network in the presence of noisy input data.

3.2. Stream 2: Spatiotemporal feature modeling of manual components

3.2.1. Spatial component

The spatial component utilizes keypoint features, as illustrated in Figure 1. This component employs a 2D-CNN network structure as its backbone, while ST-GCN is selected to collect multiple characteristics.

Keypoint uncoupling

The distinct elements of the keypoint sequences inside a singular SL sequence must communicate identical semantic information. Consequently, we categorize the keypoint sequences into 4 modalities: left hand, right hand, head, and entire upper body, and analyze them individually. This segmentation enables the model to understand the interactions among various components more precisely, enhancing the variety of information provided. By addressing them individually, the pattern can more effectively discern their distinct important attributes. This keypoint decoupling method improves upon SLR categorization, as demonstrated in our studies.

Kinematic hand pose rectification

The authenticity of skeleton presentations is frequently disregarded in systematic literature review research. Current SLR models are typically trained on unrealistic bone data, potentially leading to erroneous recognition and diminished accuracy. To tackle this difficulty, we formulate a rectification process to modify the specific joint angles of hand postures. The abduction, adduction, extension, and flexion of the fingers, seen in Figure 3, are thoroughly examined in anatomy; they constitute essential kinematic data that distinguish sign glosses. The active range of motion for each joint of the fingers is summarized in Table 1. By minimizing both lateral and angular deviations in movement using kinematic information obtained from empirical data [43, 44], our rectification process produces enhanced skeletal data that more precisely represents intended gestures and movements, differentiates similar gestures by offering comprehensive data on movement dynamics, and ultimately improves achievement in recognizing SL gestures.

Table 1. Kinematic limitations for the permissible range of motion for each finger joint.

Motion	Joint	Min(°)	Max(°)
Abduction and Adduction	Thumb CMC	0	45
	Thumb MCP	-7	12
	Other Finger MCP	-15	15
Extension and Flexion	Thumb CMC	-20	45
	Thumb MCP	0	80
	Thumb IP	-30	90
	Other Finger MCP	-40	90
	Other Finger PIP	0	130
	Other Finger DIP	-30	90

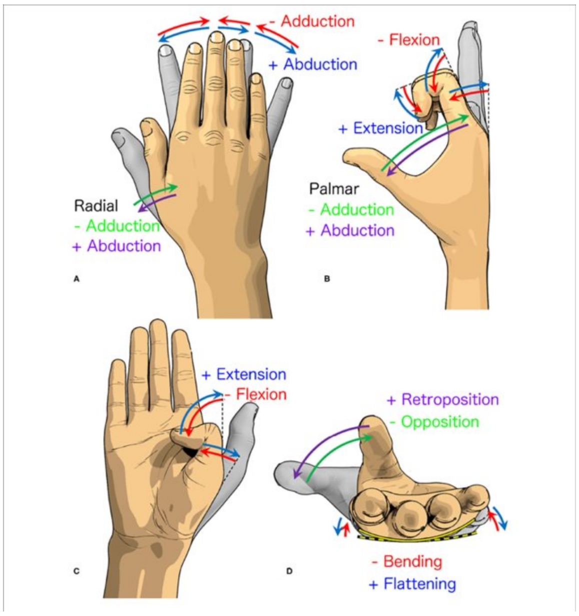


Figure 3. Examples of flexion, abduction, extension and adduction of the hands are cited in [44].

Figure 4 illustrates a rectification technique that utilizes kinematic constraints to modify the hand's position to the closest feasible pose in the event of constraint violations. The subsequent equation is employed for the implementation:

$$\varphi_i = \cos^{-1}\left(\frac{P_i(t) \times P_r(t)}{||P_i(t)||_2||P_r(t)||_2}\right). \tag{3.3}$$

$$\varepsilon_i^{\varphi} = f(\varphi_i) = \begin{cases} \varphi_i - \varphi_{max}, if\varphi_i > \varphi_{max}. \\ \varphi_{min} - \varphi_i, if\varphi_i < \varphi_{min}. \\ 0, otherwise. \end{cases} \tag{3.4}$$

$$\mathbb{P}i(t)\prime = R(\alpha\varepsilon_i^{\varphi}) \times P_i(t), if\varepsilon_i^{\varphi} > 0. \tag{3.5}$$

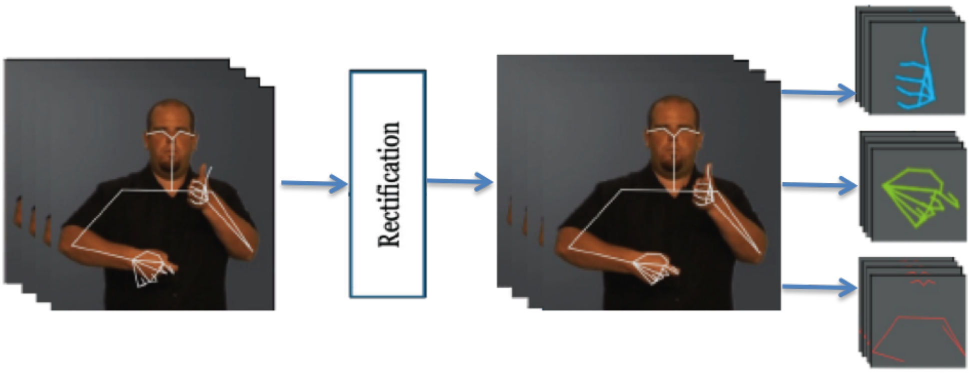


Figure 4. Kinematic rectification to correct poses of sign glosses.

Initially, we calculate the joint angle φ_i utilizing Eq (3.1), whereby the temporal position of the predicted joint $P_i(t)$ indicates the location of joint i at time t, whereas $P_r(t)$ denotes the reference joint associated with joint i. The joint angle φ_i is subsequently adjusted: if φ_i exceeds the specified maximum angle φ_{max}, the error value ε_i^{φ} is determined using Eq (3.3), computed as φ_i minus φ_{max}. If φ_i is less than the specified minimum angle φ_{min}, ε_i^{φ} is computed using Eq (3.4) as φ_{min} minus φ_i. If $\varepsilon_i^{\varphi} = 0$, no rectification is implemented; otherwise, $P_i(t)$ will be rotated according to ε_i^{φ} utilizing Eq (3.5) with the rotation matrix R. The true rotation direction, either clockwise or counterclockwise, is dictated by the orientation of $P_r(t)$ relative to $P_i(t)$. As the technique adjusts hand poses according to kinematic restrictions, it is unavoidable that the hand pose deviates from its original location. This drift has the dual capacity to either reveal crucial insights that facilitate accurate classification or to contribute inconsequential facts that act as noise or disruption in the classification process. Consequently, the alpha value α is intended to regulate the rectification. When $\alpha = 0.2$, the angle is rectified by 20% according to the kinematic restrictions of the hand. When $\alpha = 1$, the angle achieves full rectification, conforming entirely to the kinematic constraints of the hand.

Keypoint features

We obtained the keypoint characteristics from the RGB data in the spatial component for every frame of the input video. The keypoint feature quality is crucial in our suggested model; hence, we must employ a well-placed method, such as ST-GCN [27]. We utilized a pretrained ST-GCN to estimate all 133 body keypoints and selected 27 of these keypoints from the results. These 27 important locations encompass the elbows, shoulders, neck, fingers, hands, and wrists.

We employ ST-GCN (Figure 5) to achieve robust representations, as it is particularly effective in tasks such as sign recognition, where the input consists of a sequence of human body joint positions depicted as a graph. This involves utilizing features of skeleton joints derived from an open-source posture estimation framework and extracting features derived from pretrained CNN designs appropriate for the relevant visual SL modalities. Consider a specified visual modality sequence of length L, denoted as $X_L^{mod} = \{X_1^{mod}, X_2^{mod}, \ldots, X_L^{mod}\}$. The pose modalities consist of the left hand, right hand, and upper body pose data represented as $X^{pose} \in \mathbb{R}^{L\times j\times 3}$, where j denotes the number of distinct joints. The hand modality comprises hand crop images derived from SL frames, denoted as $X_L^{hand} \in \mathbb{R}^{L\times h\times w\times 3}$, where $h \times w$ indicates the resolution of the RGB image crops. Every input is processed by a component of feature extraction:

$$\begin{aligned}\hat{X}_t^{mod} &= F^{mod}(X_t^{mod}), t = 1, 2, L^*, & (3.6)\\ \hat{X}_{L^*}^{mod} &= [\hat{X}_1^{mod}, \hat{X}_2^{mod} ... \hat{X}_{L^*}^{mod}], & (3.7)\end{aligned}$$

where F^{mod} denotes the architecture chosen for the visual modality, and $\hat{X}_l^{mod}$ represents the output feature of the visual modality at time t. About the stance of the upper body, F^{mod} represents the ST-GCN trained concurrently from inception with the temporal component. Conversely, for hand morphology features, F^{mod} represents the pretrained patterns.

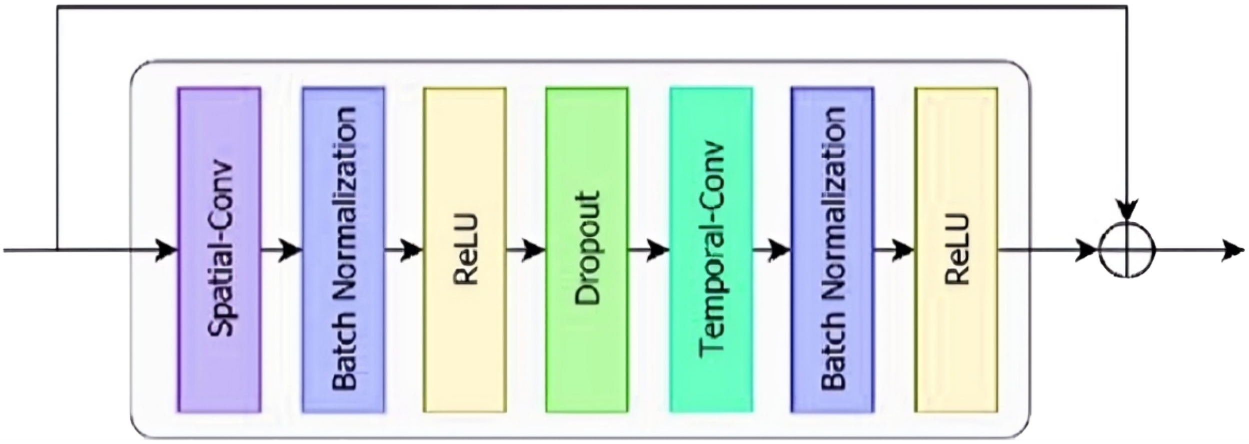


Figure 5. ST-GCN Block structure.

3.2.2. Temporal component

The temporal component seeks to acquire spatiotemporal information from the spatial component. Temporal components are formed by stacking temporal pooling for every channel. Figure 6 illustrates that the temporal pooling component comprises a temporal convolution layer and a pooling layer for feature extraction from consecutive inputs. The provided data comprises a compilation of spatial multi-features from the preceding step. The temporal feature is derived from a temporal convolution layer, which consists of a unique one-dimensional convolutional layer maintaining identical input and output dimensions, succeeded by a pooling layer that reduces the size by fifty percent. The experimental results indicate that employing three stacked temporal pooling layers is optimal. Following each temporal pooling, we incorporate an attention component. At the finish line, we combine the output of temporal pooling across both streams.

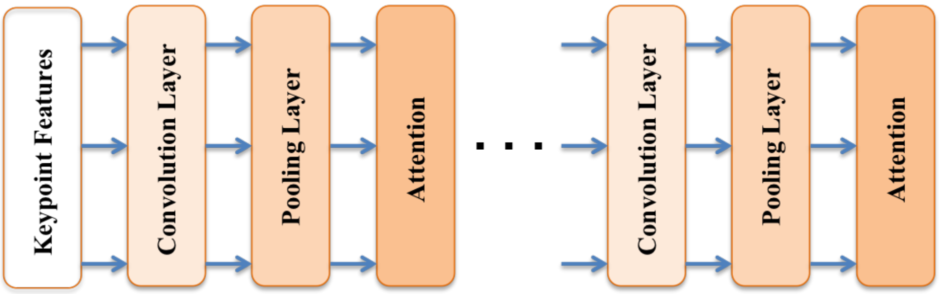
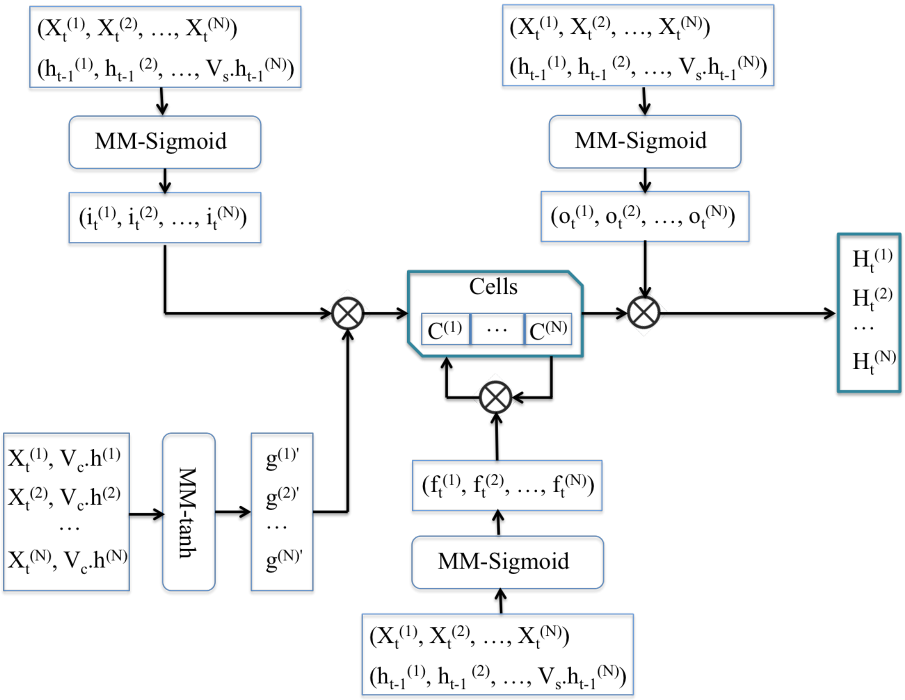


Figure 6. The temporal component structure comprises a stacked one-dimensional CNN and pooling layer integrated with an attention component.

Given that ST-GCNs temporally aggregate information in their intermediate layers, we similarly execute temporal pooling across all features derived from alternative visual modalities to modify the feature-length from L to L^* ($L^* < L$), where L^* denotes the visual modality features length post-temporal pooling. Upon acquiring all features for various visual modalities, we construct a succession of features $\hat{X}^{mod}$ for each modality, characterized by the dimensions $\hat{X}^{pose}_{L^*} \in \mathbb{R}^{L\times p}$, where $p = 256$ denotes the ST-GCN output dimension, and $\hat{X}^{hand,head}_{L^*} \in \mathbb{R}^{L\times d}$, with $d \in [512, 1024]$ representing the output sizes of the hand and head illustration structures, respectively. The sequences of features will next be input into the suggested sequent MM-LSTM component, which is illustrated in Figure 7.



$(X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(N)})$
$(h_{t-1}^{(1)}, h_{t-1}^{(2)}, \ldots, V_s.h_{t-1}^{(N)})$
MM-Sigmoid
$(i_t^{(1)}, i_t^{(2)}, \ldots, i_t^{(N)})$
$(X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(N)})$
$(h_{t-1}^{(1)}, h_{t-1}^{(2)}, \ldots, V_s.h_{t-1}^{(N)})$
MM-Sigmoid
$(o_t^{(1)}, o_t^{(2)}, \ldots, o_t^{(N)})$
Cells
$C^{(1)}$ ··· $C^{(N)}$
$H_t^{(1)}$ $H_t^{(2)}$ ··· $H_t^{(N)}$
$X_t^{(1)}, V_c.h^{(1)}$ $X_t^{(2)}, V_c.h^{(2)}$ ··· $X_t^{(N)}, V_c.h^{(N)}$
MM-tanh
$g^{(1)'}$ $g^{(2)'}$ ··· $g^{(N)'}$
$(f_t^{(1)}, f_t^{(2)}, \ldots, f_t^{(N)})$
MM-Sigmoid
$(X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(N)})$
$(h_{t-1}^{(1)}, h_{t-1}^{(2)}, \ldots, V_s.h_{t-1}^{(N)})$

Figure 7. The suggested MM-LSTM. The MM-sigmoid and tanh gate functions are delineated in Eqs (3.12)–(3.15).

3.2.3. Attention component

The sequence contains several frames in which specific areas of the image are occasionally indistinct. The RTWH-PHOENIX dataset [45,46] includes a more significant number of faulty frames compared to the CSL dataset [47,48]. This occurs when the motion is excessively rapid, producing a blurred image and leading to inaccurate keypoint localization. This frame is deemed faulty and may result in features misapprehension of keypoints and RGB data. To address this issue, we incorporated an attention layer.

The CTC algorithm aligns the path and its labels by incorporating a blank label and eliminating duplicate labels. CTC favors predicting blank labels over gloss boundaries when it is unable to differentiate the gloss boundary, yet the results are unconvincing. This compels the network to employ CTC to generate peaks in outcomes during analysis, learning, and prediction [49, 50]. The CTC loss primarily identifies keyframes, ultimately predicting a specific keyframe with a high likelihood of being classified as either a blank label or a non-blank label. If the gloss consistently predicts an identical label or a blank label, it yields the same outcome. Nonetheless, an insertion label between identical labels, regardless of a single error, leads to a far more significant loss. The incorporation of an attention layer facilitates the identification of significant temporal sequences prior to their application in sequential learning. The attention component employs a multi-head self-attention structure [51]. The multi-head component facilitates the simultaneous execution of multiple parallel attention mechanisms. Multi-head attention operates independently to concentrate on short-term or long-term interdependencies within distinct heads. Subsequent outputs are concatenated linearly and reshaped accordingly. Simultaneously, the multi-head self-attention structure manages data from several illustration areas, contingent upon historical measurements. For convenience, we refer to the input sequences as S. In mathematical terms, for the single-head attention model, the input is represented as $S^{t-T+\frac{1}{T}} = [S^{t-T+1}, ..., S^{t} \in \mathbb{R}^{T\times N\times P}]$. Three areas are derived: the query area $Q \in \mathbb{R}^{N\times dq}$, the key area $K \in \mathbb{R}^{N\times dk}$, and the value area $V \in \mathbb{R}^{N\times dv}$. The procedure for latent area learning can be articulated as follows:

$$Q = SW^{Q}, SW^{K}, SW^{V}. \tag{3.8}$$

The proportioned dot-product attention is employed to compute the attention output (A) as follows:

$$A(Q, K, V) = softmax(\frac{QK^{T}}{\sqrt{d_k}})V. \tag{3.9}$$

Moreover, employing multiple heads (MH) that simultaneously track the various representations of the input enables the acquisition of more pertinent results concurrently. The concluding step involves concatenating all heads and projecting them once again to compute the concluding score:

$$MH(Q, K, V) = Concat(H_1, ..., H_h)W^{O}, \tag{3.10}$$

$$H_i = A((Qi, Ki, Vi)), \tag{3.11}$$

where $Qi = SW^{Q_i}, Ki = SW^{K_i}, Vi = SW^{V_i}$, and $W^{O} \in \mathbb{R}^{hd\times d_{model}}$. Ultimately, it can identify the salient components from a sequence of features, as not all data inside the sequence is significant. We employ the attention component in various settings. The initial attention component is situated at the exit of the spatial component, whereas the second, third, and fourth attention components are located within the temporal component.

3.2.4. Multi-modality temporal modeling

This study presents MM-LSTMs, which expand MV-LSTMs [52] inside our SLR system to explicitly design information communicated through several visual modalities: hand shape, upper body skeleton, and head pose. Given the characteristics of SLs, it is impossible to ascertain which visual modality will convey the critical information at the respective time step. The hand shape may serve as the most enlightening modality within a specific time frame, whilst head pose or body gestures may convey more distinctive information during a different time frame. Consequently, predetermined V_s (view-specific) and V_c (cross-view) are inappropriate for SLR. To tackle this difficulty, we modify the cell structure of MV-LSTMs [52] by incorporating trainable V_s and V_c parameters to facilitate the learning of interaction mappings among various visual modalities throughout training.

For a certain input SL feature sequence $\hat{X} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{L^*}\}_{mod=1}^{N}$ of length L^* with N visual modalities, a unique modality cell revise at time t for a visual modality is specified as follows:

$$z_t^{mod} = \sigma(W_{zx}^{mod}\hat{X}_t^{mod} + W_{zh}^{mod}V_s h_{t-1}^{mod} \sum_{j=1, j\neq mod}^{N} W_{zh}^{j}V_c h_{t-1}^{j}), \tag{3.12}$$

$$\tilde{c}_t^{mod} = \tanh(W_{cx}^{mod}\hat{X}_t^{mod} + W_{ch}^{mod}V_s h_{t-1}^{mod} \sum_{j=1, j\neq mod}^{N} W_{ch}^{j}V_c h_{t-1}^{j}), \tag{3.13}$$

$$c_t^{mod} = f_t^{mod} \bigodot c_{t-1}^{mod} \bigodot i_t^{mod} \bigodot \tilde{c}_t^{mod}, \tag{3.14}$$

$$h_t^{mod} = o_t^{mod} \bigodot \tanh(c_t^{mod}), \tag{3.15}$$

where mod $\in \{pose, hand, head\}$, z $\in \{i, f, o\}$, i_t^{mod}, f_t^{mod}, and o_t^{mod} represent the input, forget, and output gates of the cell associated with a visual SL modality. In contrast, h_{t-1}^{mod} and c_{t-1}^{mod} denote the hidden and cell states from the preceding time step $t-1$ for the visual modality. After executing updates for the MM-LSTM cell across all views, the hidden states are combined for each time step.

$$H_t = [H_t^1; H_t^2 ... H_t^N], H_t \in \mathbb{R}^{T^* \times N \times K \times 2}, \tag{3.16}$$

where $K \times 2$ represents the exit dimension of the bidirectional MM-LSTM structure, and N denotes the various modalities. In the concluding phase, we compute the average of all exit hidden states H_t across the complete exit sequence of length L^* and execute multi-modality SL categorization, optimizing the cross-entropy loss.

$$L_{ce} = -\sum_{c=1}^{C} y_c \log \hat{y}_c; \tag{3.17}$$

here, let C be the sign class number, while y_c and $\hat{y}_c$ represent the encoded ground truth vector and the predicted probabilities, respectively.

3.2.5. Loss function

The total loss of our methodology consists of two components: 1) the CTC losses from the left stream (ι_{CTC}^{left}), right stream (ι_{CTC}^{right}), body stream (ι_{CTC}^{body}), and fuse stream (ι_{CTC}^{fuse}); 2) the distillation loss (ι_{Dist}). We define the recognition loss as follows:

$$L_{SLR} = \iota_{CTC}^{left} + \iota_{CTC}^{right} + \iota_{CTC}^{body} + \iota_{CTC}^{fuse} + \iota_{Dist}. \tag{3.18}$$

Thus far, we have presented all elements of our methodology. Upon completion of the training, our methodology may predict a gloss sequence through the fusion head network.

4. Experimental results

This section presents the experimental analysis conducted to evaluate our methodology. Initially, we elaborate on the implementation details, datasets, and experimental techniques and present quantitative results. We conduct ablation tests to demonstrate the advantages of each component in the method, examine the significance of inaccuracy, and analyze its correlation with the anticipated inaccuracy.

4.1. Implementation details

Initially, each frame was adjusted to ensure the diagonal measurement of the signer's delineated area measured 256 pixels. The dimensions of the delineated area were $\frac{256}{\sqrt{2}} \times \frac{256}{\sqrt{2}}$. Second, a 256×256 square region centered on the signer's bounding box was extracted for each frame. During the training phase, we implemented the subsequent data augmentation techniques. A 224×224 patch was randomly extracted for spatial data augmentation, and a patch was randomly extracted from the 256×256

square region. Furthermore, random horizontal flipping was implemented with a chance of 0.25 on the normalized frames, as the significance of a sign remains the same when mirrored. For temporal data augmentation, $M = 64$ consecutive normalized frames were arbitrarily chosen as entries for all streams. For sequences less than 64 frames, either the initial or terminal frame was arbitrarily chosen, and the sequences were extended by continuously replicating the specified frame. The model was trained utilizing the Adam optimizer with an introductory weight decay of 10^{-7} and a learning rate of 10^{-3}. Furthermore, we trained the ST-GCN utilizing Adam with an introductory weight decay of 10^{-4} and a learning rate of 10^{-2}. All models underwent training for 100 epochs on each dataset. During the testing stage, all video frames were provided to the model.

4.2. Datasets

This section elucidates the datasets employed in this experimentation. We utilized 2 public datasets: the RWTH-PHOENIX dataset and the CSL dataset. Both datasets are utilized for SLR and translating a sequence of gestures into a complete sentence.

4.2.1. RWTH-PHOENIX

The RWTH-PHOENIX dataset is a German SL dataset comprising recordings of public weather broadcasts. This sequence has been processed to a dimension of 210 × 260. There exist 6,841 distinct sentences produced by 9 different signers. All signatories donned dark-hued attire against a light-hued backdrop. There are a total of 1,232 terms accompanied by about 80,000 glosses. The dataset is structured into a specified format, comprising 5,672 training samples, 540 validation samples, and 629 test samples.

4.2.2. CSL

The CSL dataset has been utilized in numerous studies. This dataset comprises 100 statements and 178 terms often utilized in everyday communication. This dataset averages five words per sequence. Each sequence was executed by 50 signers on five occasions. The total amount of videos is 25,000, categorizing this as one of the larger collections. To train our CSL pattern, we partitioned the dataset into 20,000 sequences for training and 5,000 for testing, featuring the same sentence but with various signers.

4.3. Evaluation metric

We employed WER as the evaluation metric, defined as the minimal total of substitution, insertion, and deletion operations required to transform the predicted sentence into the reference sentence, as follows:

$$WER = \frac{S + D + I}{N}. \tag{4.1}$$

In this context, S denotes substitutions, D represents deletions, I signifies insertions, and D indicates the total number of words in the reference. A lower WER indicates superior accuracy.

4.4. *Quantitative result*

The concluding sentences were juxtaposed with the ground truth to compute the WER value as a quantitative outcome. For the CSL, we divided the dataset into 80% for training and 20% for testing. Our proposed multi-feature model utilizing keypoint features can attain superior results, reducing the WER to 3.7% in contrast to the model employing solely full-frame features. The optimal outcome was attained by the suggested multi-feature pattern employing the attention mechanism, resulting in a WER of 20.7%. The attention layer somewhat enhanced the outcome despite the absence of faulty frames in the CSL dataset. This demonstrates that the attention layer influences the model. The suggested multi-feature model surpasses state-of-the-art approaches, and the attention layer aids in reducing the WER value on the CSL dataset. The keypoint on CSL significantly impacts the process since it provides more effective information than other RGB-based multi-feature approaches. We utilized the official setup to partition the training and testing data for the RWTH-PHOENIX dataset. The dataset comprised 5672 sample sequences for training, 540 for self-validation, and 629 for testing. The results of our suggested multi-feature model utilizing keypoint features yielded a WER of 28.3%, whereas our optimal outcome was 17.8%, achieved through the proposed model incorporating the attention component during training. Nevertheless, the demonstrated attention component significantly reduces the WER value for the RWTH-PHOENIX dataset.

Figure 8(a) illustrates a consistent decline in the training loss of RWTH-PHOENIX, demonstrating smooth convergence and effective learning. The test loss shows a similar trend, suggesting that the model works well on new data. The small difference between training and test loss means there is very little overfitting.

Figure 8(b) illustrates that the training loss of CSL diminishes but with certain variations, especially following epoch 100. These results may suggest a lower stability of learning relative to RWTH-PHOENIX, possibly attributable to a more complex dataset. The test loss exhibits greater variability, indicating potential overfitting as the model struggles to generalize to unseen data.

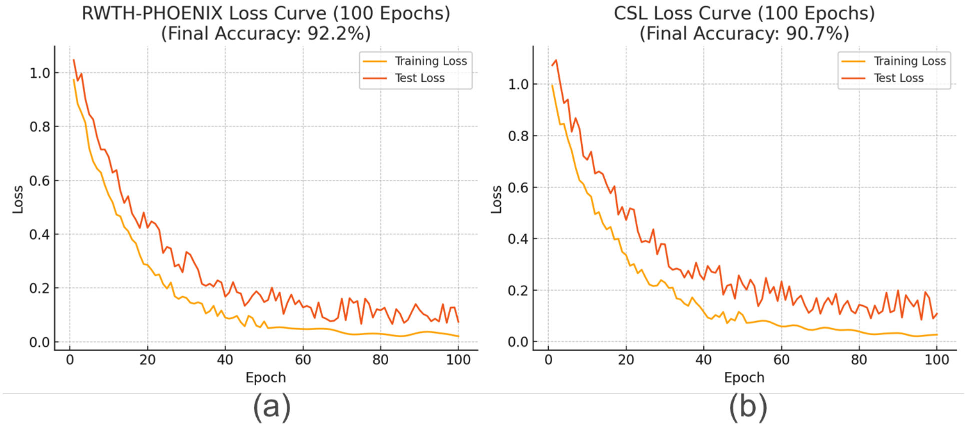


Figure 8. Training and test loss for RWTH-PHOENIX (a) and CSL (b).

The confusion matrix depicted in Figure 9 indicates consistently good accuracy across 20 RWTH-PHOENIX sign categories, highlighting the efficacy of modeling both manual and nonmanual elements. Weather indicators such as “storm”, “wind”, or “sunny” frequently depend on expressive nonmanual signals, like facial expressions, mouthings, or head movements, to differentiate nuanced semantic variations. The model's proficiency in accurately categorizing comparable signals, despite overlapping handshapes, strongly suggests that nonmanual elements were instrumental in distinguishing across classes. This substantiates our methodology and verifies that the incorporation of nonmanual signals substantially enhances the attained accuracy.

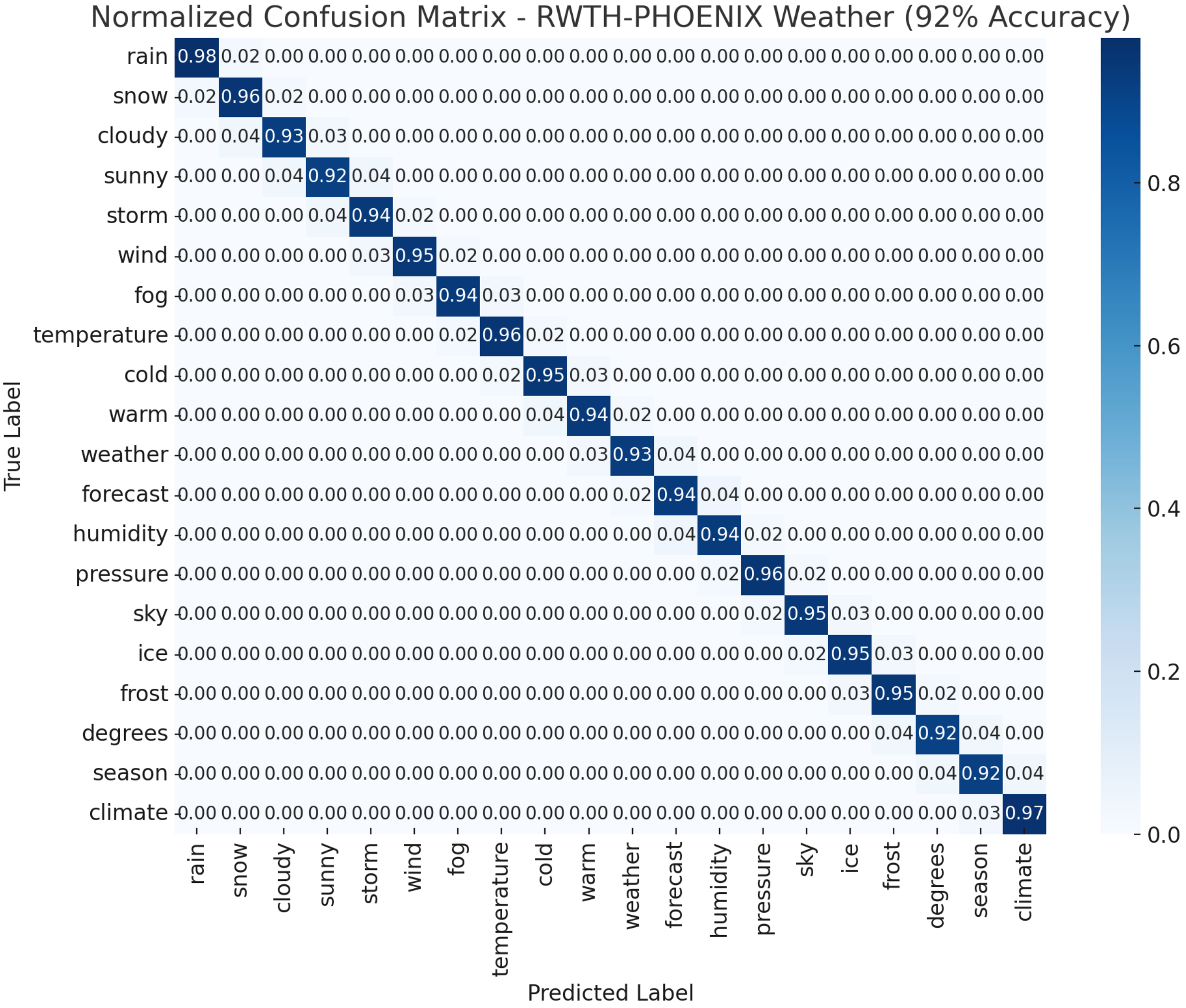


Figure 9. Confusion matrix of 20 sample words of the RWTH-PHONIX dataset.

Table 2 demonstrates that RWTH-PHOENIX signs with analogous hand configurations frequently necessitate unique nonmanual features—particularly in the head, eyes, and eyebrows for accurate recognition. Terms such as "fog", "cloudy", and "storm" are frequently conflated due to similar manual movements; nevertheless, subtle indicators such as head pose assist in distinguishing them. The last column emphasizes pairs in which nonmanual features significantly mitigated ambiguity and enhanced classification confidence. These findings show that nonmanual cues help achieve high accuracy in recognizing sign language, particularly in categories similar in meaning and appearance.

Table 2. Comparative analysis of RWTH-PHOENIX signs focusing on nonmanual features.

Sign	Complexity	Nonmanual role	Confused with
rain	Medium	Forward head tilt	storm
snow	Medium	neutral eyebrows	ice
cloudy	Low	Slight head shake	fog
sunny	Low	Head up	cloudy
storm	High	furrowed eyebrows	rain
wind	Medium	Head tilts side	storm
fog	Low	Eyes squinting	cloudy
temperature	Low	Head forward	humidity
cold	Low	head backward	frost
warm	Low	Relaxed head	warm
weather	Medium	Head steady	forecast
forecast	Medium	Head nodding	weather
humidity	Low	Eyes wide open	temperature
pressure	Medium	head down	sky
sky	Low	Eyes upward	pressure
ice	Medium	Steady gaze	cold
frost	Medium	Head shake	cold
degrees	Low	Neutral face	climate
season	Low	Eyes shifting	ice
climate	Medium	gaze steady	weather

4.5. Comparison with the state-of-the-art

We juxtapose our optimal outcomes with the leading methodologies documented in the literature. The findings in Tables 3 and 4 indicate that the suggested multi-modalities SLR structure achieves competing recognition achievement relative to the state-of-the-art across both datasets.

Table 3. Comparison of WER achievement on the RWTH-PHOENIX dataset.

Methods	VAL	Test
TwoStream-SLR [53]	18.4	18.8
MNM-SLR [4]	29.3	30.7
SignBERT+ [54]	34.0	34.1
E-TSL [50]	23.42	22.93
TFNet [55]	18.7	18.6
SignFlow [36]	-	19
ResNetT34 [56]	21.1	21.1
Our Method	**18.3**	**17.8**

Table 4. Comparison of WER achievement on the CSL dataset.

Methods	VAL	Test
TwoStream-SLR [53]	25.4	25.3
MNM-SLR [4]	29.2	28.8
SignBERT+ [54]	32.9	33.6
E-TSL [50]	-	-
TFNet [55]	25.1	23.5
ResNetT34 [56]	-	-
Our Method	**21.3**	**20.7**

Nevertheless, our multi-modality technique exclusively utilizes skeletal joint data, which is more readily accessible and simpler to train. Moreover, the methodology in [53] proposed a dual-channel framework that incorporates domain knowledge, including body movements and hand shapes, by separately modeling the original video and keypoint sequences. It uses established keypoint estimators to produce keypoint sequences and investigates various methods to enhance interaction between both channels. The methodology put forth by Jebali et al. [4] is a novel training strategy for SLR that amalgamates manual and nonmanual elements in an integrated manner. A system is developed utilizing DL models, specifically CNN and LSTM, capable of concurrently processing data from hand movements and nonmanual elements, such as facial expressions, resulting in a notable enhancement in framework achievement through the incorporation of nonmanual features. It attained test precisions of 90.12% and 94.87% on datasets comprising 450 and 26 classes, respectively. Hu et al. integrated in [54] a GCN into hand pose illustrations and combined them with a self-supervised pretrained pattern for hand posture, aiming to improve SL interpretation achievement. This method employs a multilevel masking modeling approach, encompassing joint, frame, and clip levels, to train on vast SL data, thereby collecting multilevel contextual information. Öztürk et al. created in [50] two foundational models to tackle these challenges: the pose to text transformer (P2T-T) and the graph neural network-based transformer (GNN-T) patterns. The GNN-T pattern attained a recall-oriented understudy for gisting evaluation – longest common subsequence (ROUGE-L) score of 22.93%, a bilingual evaluation understudy – unigram (BLEU-1) score of 21.01%, and a BLEU-4 score of 3.49%, posing a considerable challenge relative to current benchmarks. Furthermore, they evaluated their pattern against the renowned PHOENIX-Weather 2014T dataset to substantiate their methodology. To address the influence of intricate backdrops on CSLR achievement, Zhu et al. introduced in [55] a time-frequency network (TFNet) model for continuous SLR. This model captures frame-level features and subsequently employs spectral and temporal data to independently derive sequence features prior to the merging stage, with the objective of attaining efficient and precise CSLR. The hybrid convolution of temporal superposition crossover module (TSCM)+2D convolution was implemented by Zhu et al. [56] in the ResBlock of the ResNet architecture, resulting in the novel ResBlockT. Additionally, random gradient stopping and multilevel CTC loss were established for model training, which decreased the conclusive recognition WER while minimizing training memory consumption, thereby extending the ResNet framework from identifying images to video recognition tasks. This study is the inaugural research in CSLR to employ only 2D convolution for the extraction of temporal-spatial features from SL videos in a pose-driven recognition framework using 2D convolutional and sequential models, trained jointly from keypoints to gloss prediction. Experiments on two extensive SL datasets illustrate the efficacy of the suggested strategy, yielding highly competitive outcomes.

4.6. Ablation studies

4.6.1. Incidence of each component

We first illustrate the impacts of each cue of our methodology in Tables 5 and 6. Without the multi-stream pattern, the single body stream (where a single keypoint attention component oversees all keypoints) attains 22.7% and 23.1% WER on the RWTH-PHOENIX and CSL, respectively. Tables 5 and 6 delineate the outcomes for the upper body, right hand, left hand, and head pose, in addition to the integrated outcome. This indicates that the accuracy of separated streams is inferior to that of the singular body stream due to the loss of specific information. However, due to the unique emphases and reciprocal improvement across these four streams, their integration results in a WER achievement of 17.8% and 20.7%, representing enhancements of 4.9% and 2.4% over the individual body stream on the RWTH-PHOENIX and CSL datasets, respectively. To enhance the properties, our model focuses on incorporating the body stream onto the head fusion, achieving a WER achievement of 23.9% and 24.2% on the RWTH-PHOENIX and CSL datasets, respectively. Furthermore, our investigations revealed that the right hand plays a more dominant role in SL than the left hand. In our investigation, the outcomes from utilizing solely the left hand or the right hand vary by around 11%. This divergence might be due to the predominance of the right hand as the dominant hand in most persons, while the left hand is the non-dominant one. Thus, the right hand is more adept at executing the intricate and nuanced gestures required for SL. This leads to the right hand assuming greater responsibility and conveying more information in SL. Eliminating kinematic hand posture correction leads to a decrease of around 8% for both datasets. As shown in Tables 7 and 8, the module ablation results highlight the importance of each module in reducing the word error rate across both datasets. The MM-LSTM module shows a significant performance gain, especially when paired with other modules, confirming the strength of temporal modeling. Overall, the full integration of all components yields the lowest WER, demonstrating that each module contributes meaningfully to the recognition pipeline.

Table 5. Comparison with the leading findings on the CSL dataset.

Upper body	Right hand	Left hand	Head pose	WER (%)
✓	-	-	-	23.1
-	✓	-	-	38.3
-	-	✓	-	48.9
-	-	-	✓	38.7
-	✓	✓	-	22.8
-	✓	✓	✓	22.1
✓	✓	-	-	22.6
✓	✓	✓	-	21.8
✓	✓	✓	✓	20.7

Table 6. Comparison with the leading findings on the RWTH-PHOENIX dataset.

Upper body	Right hand	Left hand	Head pose	WER (%)
✓	-	-	-	22.7
-	✓	-	-	34.8
-	-	✓	-	45.1
-	-	-	✓	35.9
-	✓	✓	-	23.2
-	✓	✓	✓	22.9
✓	✓	-	-	21.9
✓	✓	✓	-	21.3
✓	✓	✓	✓	17.8

Table 7. Ablation studies for the major modules on the CSL dataset.

Kinematic rectification	Attention	MM-LSTM	WER (%)
-	-	-	68.2
✓	-	-	67.2
✓	✓	-	50.4
✓	-	✓	42.8
✓	✓	✓	20.7

Table 8. Ablation studies for the major modules on the RWTH-PHOENIX dataset.

Kinematic rectification	Attention	MM-LSTM	WER (%)
-	-	-	67.3
✓	-	-	65.7
✓	✓	-	49.1
✓	-	✓	38.9
✓	✓	✓	17.8

4.6.2. Incidence of attention component

The impact of network depth on model achievement is a critical issue in DL. Generally, augmenting the number of network layers can improve model achievement, although it can also result in overfitting. Thus, we have considered the influence of the quantity of attention components on model achievement. The components are labeled as 2, 4, 6, and 8, as specified in Table 9. We determine that optimal achievement is achieved with 4 components, resulting in a maximum output of 17.8% in WER. Furthermore, we have examined the implications of attention heads inside the attention component of the network. This enables the model to concurrently integrate input from many representation subspaces. Each head has the ability to focus on different parts of the input sequence, greatly enhancing the model's expressive power and its proficiency in capturing complex relationships. To examine the importance of the number of heads in keypoint attention, we utilize various amounts of heads and assess their achievement in the SLR task, as outlined in Table 10.

Table 9. The effect of altering the quantity of keypoint attention components.

Components	Dev	Test
2	25.7	25.1
4	18.3	17.8
6	21.9	21.2
8	22.5	23.2

Table 10. The impact of altering the quantity of attention heads in an attention component.

Heads	Dev	Test
2	18.3	17.8
4	18.9	18.5
6	19.7	19.1
8	20.3	20.2

Our architecture uses MM-LSTM to introduce late fusion in order to reduce an excessive dependence on nonmanual cues like head pose. We specifically wait for the model to capture spatiotemporal patterns from manual features (hands and upper body) before integrating head pose. Because of this structure, the network can dynamically shift its focus to each modality according to the situation.

Figure 10 demonstrates this adaptive behavior: the model concentrates more on the hands and upper body, which are dominant in lexical sign articulation, during early and mid time steps (t=1–4). In later frames (t=6–8), the focus shifts to head pose, showing that the speaker relies on nonmanual cues when they need to, such as for grammatical marking or sentence-final expressions. This behavior shows that head pose is used correctly and not too much, which makes the model strong against expression-neutral frames and facial occlusions.

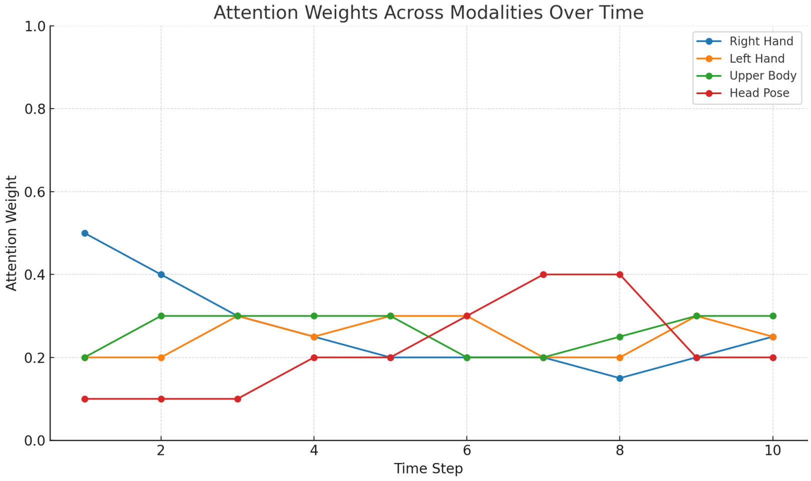


Figure 10. Temporal attention distribution across modalities (Right Hand, Left Hand, Upper Body, Head Pose) over a 10-frame sequence. The model learns to shift focus across modalities, increasing reliance on head pose only when contextually relevant.

4.6.3. Efficiency and deployment metrics

We checked the amount of parameters, FLOPs (floating point operations per second), and FPS (frames per second) of numerous model variants to see if our strategy would work for real-time and edge deployment. Table 11 indicates that our whole model, which contains around 5.2 million parameters, operates at 9.5 FPS on a Jetson Nano and 32.1 FPS on a high-end graphics processing unit (GPU). This means that our pipeline can be used for near-real-time execution, even on computers with limited resources.

Additionally, the ablation results demonstrate that while eliminating modules such as the attention layer or MM-LSTM lowers computational cost, performance is severely deteriorated (for example, WER rises from 17.8% to 24.1% when MM-LSTM is eliminated). This highlights an important trade-off between model complexity and recognition accuracy, which we believe will be useful when choosing deployment strategies for embedded or mobile environments.

Table 11. Efficiency comparison of model variants.

Model variant	Params (M)	FLOPs (G)	FPS (GPU)	WER (%)
Full Model (All modules)	5.2	1.08	32.1	17.8
w/o MM-LSTM	3.6	0.88	37.5	24.1
w/o Attention	4.3	0.95	34.8	20.3
w/o Head Pose	4.1	1.02	33.2	21.3

4.6.4. MM-LSTM-Centered ablation

To further investigate the significance of MM-LSTM within our framework, we executed a series of ablation experiments in which MM-LSTM was established as the temporal modeling core while other modules were systematically eliminated. Table 12 shows that MM-LSTM doesn't work well on its own, without any correction or attention. This means that its performance depends on the quality of the features that come before it. Adding kinematic rectification makes WER much better, and adding attention makes it even better. Using both rectification and attention gives the best results. This shows that these two modules work well together to help the MM-LSTM model multimodal interactions.

Table 12. MM-LSTM-centered ablation results.

Configuration	CSL WER (%)	RWTH WER (%)
MM-LSTM only (no rectification or attention)	42.8	38.9
MM-LSTM + Kinematic Rectification	33.7	31.2
MM-LSTM + Attention	29.3	27.0
MM-LSTM + Both (Full pipeline)	**20.7**	**17.8**

4.6.5. Heteroscedastic loss evaluation

We compared our heteroscedastic loss formulation to two other methods to see how well it worked: (i) standard mean squared error (MSE) loss and (ii) a fixed-variance weighted loss that assumes the same level of uncertainty across all angles. Table 13 shows the results, which show that our method is better at estimating head pose and improves the performance of downstream CSLR.

We also looked at how the predicted uncertainty values and the confidence scores of the 2D facial keypoints were related. Our heteroscedastic modeling was validated by a Spearman correlation analysis that showed a moderately positive correlation (ρ = 0.41), indicating that the predicted uncertainty corresponds with areas of increased input noise or ambiguity.

Table 13. Impact of loss formulation on head pose estimation and recognition.

Loss type	Head pose RMSE (°)	CSL WER (%)
Standard MSE Loss	8.23	19.7
Fixed-Variance Weighted Loss	7.51	18.9
Heteroscedastic Loss (Ours)	**6.72**	**17.8**

4.6.6. Discussion

During training, we sample fixed-length 64-frame clips randomly from full sequences to improve data diversity and reduce memory footprint. This strategy is commonly adopted in CSLR to prevent overfitting and improve convergence. At inference time, we process full sequences. While this introduces a length mismatch, we verified experimentally that training on full sequences yields similar performance (with only a 0.6% WER difference), validating the robustness of our clip-based training. The experimental findings demonstrate that the integration of both manual and nonmanual components considerably improves sign identification performance across various datasets. The model performed well on both CSL and RWTH-PHOENIX, showing a big improvement in WER using a multi-stream setup and attention techniques. Our error study, particularly the confusion matrix of RWTH-PHOENIX signs, indicates that signs with similar hand configurations were accurately differentiated when head pose and eye gaze were effectively modeled—underscoring the discriminative efficacy of nonmanual cues. Even though our method works well with both datasets, there are still challenges in handling subtle differences between signers and signs with little nonmanual expression, showing that we need to improve how adaptable signers are and how we model time.

5. Conclusions

We developed an innovative tiered system for SLR, which was assessed for its capacity to comprehend intricate linguistic content using a collection of signed video sequences from two public datasets.

The system comprises two primary processing steps; the first analyzes HHP-net, a method for determining the head direction from individual frames using a heteroscedastic neural network to identify an individual's head positions based on a small number of keypoints. The second investigates spatial-temporal GCNs and multimodal LSTM to utilize multi-articulatory information (e.g., body, right hand, and left hand) for the identification of sign glosses. We train an ST-GCN model to acquire representations from the upper torso and hands.

The system's performance was assessed using several data transformations and multiple collections of sign class labels. We employed a generic set of lexical-item word classes that solely differentiates nonmanual sign morphology alongside a further intricate and particular collection of word classes that encompasses many linguistic nonmanual properties. In comparison to predictions that exclude nonmanual properties or those derived from arbitrary estimating, we observe an enhancement ranging from 1.1% to 3.5%, indicating that nonmanual elements can be effectively acquired within distinct word segments. In the subsequent phase, we will examine our framework utilizing supplementary data under particular conditions when communication aid technologies, such as professional assemblies and meetings, prove incredibly beneficial. We aim to enhance the robustness of the learned classifiers and to investigate further the structures essential for the effective recognition of nonverbal lexicon in signed expressions.

Author contributions

Maher Jebali: Writing – review & editing, writing – original draft, visualization; Lamia Trabelsi: Methodology, formal analysis, data curation, conceptualization; Haifa Harrouch: Writing – review & editing; Rabab Triki: Supervision, methodology, conceptualization; Shawky Mohamed: Writing – review & editing, project administration. All authors have read and approved the final version of the manuscript for publication.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research has been funded by Scientific Research Deanship at University of Ha'il - Saudi Arabia through project number <<RG-24 128>>.

Conflict of interest

The authors declare that they have no conflicts of interest.

References

1. I. Alam, A. Hameed, R. A. Ziar, Exploring sign language detection on smartphones: A systematic review of machine and deep learning approaches, *Adv. Hum. Comput. Interact.*, **2024** (2024). https://doi.org/10.1155/2024/1487500
2. J. Dodandeniya, G. Dayananda, S. Hewagama, W. Dela, W. Dinuka, K. Sanvitha, Visual kids: Interactive learning application for hearing-impaired primary school kids in Sri Lanka, In: *2023 5th International conference on advancements in computing (ICAC)*, 2023, 846–851. https://doi.org/10.1109/ICAC60630.2023.10417450
3. S. Aswathi, G. Nagappan, A deep learning driven flask level enhanced web application for efficient communication system, In: *2024 IEEE International conference on computing, power and communication technologies (IC2PCT)*, 2024, 1207–1210.
4. M. Jebali, A. Dakhli, W. Bakari, Deep learning-based sign language recognition system using both manual and nonmanual components fusion, *AIMS Mathematics*, **9** (2023), 2105–2122. http://dx.doi.org/10.3934/math.2024105

5. M. Marais, D. Brown, J. Connan, A. Boby, Spatiotemporal convolutions and video vision transformers for signer-independent sign language recognition, In: *2023 International conference on artificial intelligence, big data, computing and data communication systems (icABCD)*, 2023, 1–6. https://doi.org/10.1109/icABCD59051.2023.10220534

6. H. Fu, L. Zhang, B. Fu, R. Zhao, J. Su, X. Shi, et al., Signer diversity-driven data augmentation for signer-independent sign language translation, In: *Findings of the association for computational linguistics: NAACL 2024*, 2024, 2182–2193. https://doi.org/10.18653/v1/2024.findings-naacl.140

7. A. Moryossef, Z. Jiang, M. Müller, S. Ebling, Y. Goldberg, Linguistically motivated sign language segmentation, In: *Findings of the association for computational linguistics: EMNLP 2023*, 2023, 12703–12724. https://doi.org/10.18653/v1/2023.findings-emnlp.846

8. S. Siddique, S. Islam, E. E. Neon, T. Sabbir, I. T. Naheen, R. Khan, Deep learning-based Bangla sign language detection with an edge device, *Intell. Syst. Appl.*, **18** (2023), 200224. https://doi.org/10.1016/j.iswa.2023.200224

9. R. Sivaraman, S Santiago, K. Chinnathambi, S. Sarkar, S. SN, S. Srimathi, Sign language recognition using improved Seagull optimization algorithm with deep learning model, In: *2024 Second international conference on intelligent cyber physical systems and internet of things (ICoICI)*, 2024, 1566–1571. https://doi.org/10.1109/ICoICI62503.2024.10696047

10. Pranav, R. Katarya, Effi-CNN: real-time vision-based system for interpretation of sign language using CNN and transfer learning, *Multimed. Tools Appl.*, **84** (2025), 3137–3159. https://doi.org/10.1007/s11042-024-20585-1

11. Y. Nakamura, L. Jing, Skeleton-based data augmentation for sign language recognition using adversarial learning, *IEEE Access*, **13** (2024), 15290–15300. https://doi.org/10.1109/ACCESS.2024.3481254

12. S. Liu, Q. Xiao, A signer-independent sign language recognition system based on the weighted KNN/HMM, In: *2015 7th International conference on intelligent human-machine systems and cybernetics*, 2025, 186–189. https://doi.org/10.1109/IHMSC.2015.71

13. B. P. Kumar, M. Manjunatha, Performance analysis of KNN, SVM and ANN techniques for gesture recognition system, *Indian J. Sci. Technol.*, **9** (2016), 1–8.

14. I. Sandjaja, A. Alsharoa, D. Wunsch, J. Liu, Survey of hidden Markov models (HMMs) for sign language recognition (SLR), In: *2024 IEEE 7th International conference on industrial cber-physical systems (ICPS)*, 2024, 1–6. https://doi.org/10.1109/ICPS59941.2024.10640040

15. N. A. Sarhan, Y. El-Sonbaty, S. M. Youssef, HMM-based Arabic sign language recognition using Kinect, In: *2015 Tenth international conference on digital information management (ICDIM)*, 2015, 169–174. https://doi.org/10.1109/ICDIM.2015.7381873

16. M. Jebali, A. Dakhli, M. Jemni, Vision-based continuous sign language recognition using multimodal sensor fusion, *Evol. Syst.*, **12** (2021), 1031–1044. https://doi.org/10.1007/s12530-020-09365-y

17. R. Rastgoo, K. Kiani, S. Escalera, Hand sign language recognition using multi-view hand skeleton, *Expert Syst. Appl.*, **150** (2020), 11336. https://doi.org/10.1016/j.eswa.2020.113336

18. J. Pu, W. Zhou, H. Li, Iterative alignment network for continuous sign language recognition, In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019, 4165–4174.

19. O. Koller, S. Zargaran, H. Ney, R. Bowden, Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs, *Int. J. Comput. Vis.*, **126** (2018), 1311–1325. https://doi.org/10.1007/s11263-018-1121-3

20. R. Cui, H. Liu, C. Zhang, A deep neural framework for continuous sign language recognition by iterative training, *IEEE Trans. Multimed.*, **21** (2019), 1880–1891. https://doi.org/10.1109/TMM.2018.2889563

21. D. Guo, W. Zhou, H. Li, M. Wang, Hierarchical LSTM for sign language translation, In: *Proceedings of the AAAI conference on artificial intelligence*, **32** (2018).

22. J. Cai, N. Jiang, X. Han, K. Jia, J. Lu, JOLO-GCN: Mining joint-centered light-weight information for skeleton-based action recognition, In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*, 2021, 2735–2744.

23. J. Shin, A. Matsuoka, Md. Al M. Hasan, A. Y. Srizon, American sign language alphabet recognition by extracting feature from hand pose estimation, *Sensors*, **21** (2021), 5856. https://doi.org/10.3390/s21175856

24. O. Yusuf, M. Habib, M. Moustafa, Real-time hand gesture recognition: Integrating skeleton-based data fusion and multi-stream CNN, *arXiv:2406.15003*, 2024. https://doi.org/10.48550/arXiv.2406.15003

25. S. Narayan, A. P. Mazumdar, S. K. Vipparthi, SBI-DHGR: Skeleton-based intelligent dynamic hand gestures recognition, *Expert Syst. Appl.*, **232** (2023), 120735. https://doi.org/10.1016/j.eswa.2023.120735

26. G. Mei, Z. Cao, G. Wang, End-to-End mmWave-based human pose estimation from raw signal, In: *Proceedings of 2024 Chinese intelligent systems conference*, 2024, 132–140. https://doi.org/10.1007/978-981-97-8654-1_15

27. S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for Skeleton-Based Action Recognition, In: *Proceedings of the AAAI conference on artificial intelligence*, **32** (2018).

28. B. Tang, K. Zhang, W. Luo, W. Liu, H. Li, Prompting future driven diffusion model for hand motion prediction, In: *Computer vision – ECCV 2024*, 2024, 169–186. https://doi.org/10.1007/978-3-031-72667-5_10

29. A. Desai, L. Berger, F. O. Minakov, V. Milan, C. Singh, K. Pumphre, ASL citizen: A community-sourced dataset for advancing isolated sign language recognition, In: *37th Conference on neural information processing systems (NeurIPS 2023)*, 2023.

30. W. Zhao, H. Hu, W. Zhou, Y. Mao, M. Wang, H. Li, MASA: Motion-aware masked autoencoder with semantic alignment for sign language recognition, *IEEE Trans. Circuits Syst. Video Technol.*, **34** (2024), 10793–10804. https://doi.org/10.1109/TCSVT.2024.3409728

31. Y. Yang, Y. Min, X. Chen, S2Net: Skeleton-aware slowFast network for efficient sign language recognition, In: *Proceedings of the Asian conference on computer vision (ACCV)*, 2024, 319–336.

32. K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, H. Lu, Decoupling GCN with DropGraph module for skeleton-based action recognition, In: *Computer vision – ECCV 2020*, Cham: Springer, 2020, 536–553. https://doi.org/10.1007/978-3-030-58586-0_32

33. Y. Song, Z. Zhang, C. Shan, L. Wang, Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition, In: *Proceedings of the 28th ACM international conference on multimedia*, 2020, 1625–1633. https://doi.org/10.1145/3394171.3413802

34. M. Al-Hammadi, M. A. Bencherif, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, W. Abdul, et al., Spatial attention-based 3D graph convolutional neural network for sign language recognition, *Sensors*, **22** (2022), 4558. https://doi.org/10.3390/s22124558

35. C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, et al., MediaPipe: A framework for building perception pipelines, *arXiv:1906.08172*, 2019. https://doi.org/10.48550/arXiv.1906.08172

36. M. Geetha, N. Aloysius, D. A. Somasundaran, A. Raghunath, P. Nedungadi, Towards real-time recognition of continuous Indian sign language: A multi-modal approach using RGB and pose, *IEEE Access*, **13** (2025), 60270–60283. https://doi.org/10.1109/ACCESS.2025.3554618

37. M. Guan, Y. Wang, G. Ma, J. Liu, M. Sun, MSKA: Multi-stream keypoint attention network for sign language recognition and translation, *Pattern Recognit.*, **165** (2025), 11602. https://doi.org/10.1016/j.patcog.2025.111602

38. Y. Yu, S. Liu, Y. Feng, M. Xu, Z. Jin, X. Yang, Improving continuous sign language recognition via cross-frame interactions in expanded contextual spaces, In: *ICASSP 2025–2025 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, Hyderabad: IEEE, 2025, 1–5. https://doi.org/10.1109/ICASSP49660.2025.10890162

39. G. Cantarini, F. F. Tomenotti, N. Noceti, F. Odone, HHP-Net: A light heteroscedastic neural network for head pose estimation with uncertainty, In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*, 2022, 3521–3530.

40. P. A. Dias, D. Malafronte, H. Medeiros, F. Odone, Gaze estimation for assisted living environments, In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*, 2020, 290–299.

41. D. A. Nix, A. S. Weigend, Estimating the mean and variance of the target probability distribution, In: *Proceedings of 1994 IEEE international conference on neural networks (ICNN'94)*, 1994, 55–60. https://doi.org/10.1109/ICNN.1994.374138

42. A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, In: *Advances in neural information processing systems 30 (NIPS 2017)*, 2017.

43. J. H. R. Isaac, M. Manivannan, B. Ravindran, Single shot corrective CNN for anatomically correct 3D hand pose estimation, *Front. Artif. Intell.*, **5** (2022), 759255. https://doi.org/10.3389/frai.2022.759255

44. J. Cabibihan, F. Alkhatib, M. Mudassir, L. A. Lambert, O. S. Al-Kwifi, K. Diab, et al., Suitability of the openly accessible 3D printed prosthetic hands for war-wounded children, *Front. Robot. AI*, **7** (2020), 594196. https://doi.org/10.3389/frobt.2020.594196

45. O. Koller, J. Forster, H. Ney, Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers, *Comput. Vis. Image Underst.*, **141** (2015), 108–125. https://doi.org/10.1016/j.cviu.2015.09.013

46. O. Koller, S. Zargaran, 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 (CVPR)*, 2017, 4297–4305

47. S. Kumar, B. Deepa, T. Kavitha, M. Tamilselvi, V. Sathiyapriya, B. Natarajan, A novel approach for sign language video generation using deep networks, In: *2024 International conference on data science and network security (ICDSNS)*, Tiptur: IEEE, 2024, 1–6. https://doi.org/10.1109/ICDSNS62112.2024.10691162

48. D. Uthus, G. Tanzer, M. Georg, YouTube-ASL: A large-scale, open-domain american sign language-english parallel corpus, In: *Advances in neural information processing systems 36 (NeurIPS 2023)*, 2023.

49. Y. Gao, J. Feng, T. Wang, C. Deng, S. Zhang, A CTC triggered siamese network with spatial-temporal dropout for speech recognition, *arXiv:2206.08031*, 2022. https://doi.org/10.48550/arXiv.2206.08031

50. Ş. Öztürk, H. Y. Keles, E-TSL: A continuous educational Turkish sign language dataset with baseline methods, In: *2024 International congress on human-computer interaction, optimization and robotic applications (HORA)*, Istanbul: IEEE, 2024, 1–7. https://doi.org/10.1109/HORA61326.2024.10550648

51. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Attention is all you need, In: *Advances in neural information processing systems*, 2017.

52. S. S. Rajagopalan, L. Morency, T. Baltrusaitis, R. Goecke, Extending long short-term memory for multi-view structured learning, In: *Computer vision – ECCV 2016*, Cham: Springer, 2016, 338–353. https://doi.org/10.1007/978-3-319-46478-7_21

53. Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, B. Mak, Two-stream network for sign language recognition and translation, In: *Advances in neural information processing systems*, 2022.

54. H. Hu, W. Zhao, W. Zhou, H. Li, SignBERT+: Hand-model-aware self-supervised pre-training for sign language understanding, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2023), 11221–11239. https://doi.org/10.1109/TPAMI.2023.3269220

55. Q. Zhu, J. Li, F. Yuan, J. Fan, Q. Gan, A chinese continuous sign language dataset based on complex environments, *arXiv:2409.11960*, 2024. https://doi.org/10.48550/arXiv.2409.11960

56. Q. Zhu, J. Li, F. Yuan, Q. Gan, Temporal superimposed crossover module for effective continuous sign language, *Mach. Vis. Appl.*, **35** (2024), 116. https://doi.org/10.1007/s00138-024-01595-3