SignEvaluator: A Gesture and Sentence Characteristic-Based Sign Language Quality Assessment System

Zhiwen Zheng, Qingshan Wang, *Member, IEEE*, Qi Wang, and Dazhu Deng

***Abstract*—Sign language is a basic form of communication for hearing-impaired individuals. An evaluation of the quality of sign language gestures helps improve the efficiency of sign language learning. This article proposes SignEvaluator, a sign language quality assessment system with a movement quality feature extractor and assessment generator. In the former, three quality measures are proposed for gestures and sentences. The trajectory of the palm is mapped onto position space with kernel density estimation. For finger movements, the instantaneous energy and curvature of the gesture signals are extracted with Bézier curves. Meanwhile, the performer's familiarity with gestures is indicated by the movement fluency metric of sentences. In the assessment generator, the final assessment results are calculated by combining the weights of different quality metrics and the confidence of different gesture levels. The results indicate that SignEvaluator obtained an F1-score of 0.89 for 702 sentences collected from 20 performers.**

***Index Terms*—Assessment generator, finger flexibility, movement quality feature extractor, palm trajectory, sign language quality assessment (SLQA).**

I. Introduction

SIGN language quality assessment (SLQA) is the process of evaluating a performer's proficiency in executing sign language gestures. It is of significance in several domains, including sign language translation, teaching [1], [2], [3], and virtual reality.

SLQA is a subset of action quality assessment (AQA), a field that has gained substantial attention in recent years [4]. AQA can be categorized into two groups based on the equipment employed: vision-based methods and wearable device-based methods. Vision-based methods are usually used in sports [5], [6], [7], healthcare, and daily living skills quality assessment [8]. The vision-based methods encompasses two phases: feature extraction and evaluation. AQA experiments were the first to use manual feature extraction techniques in the feature extraction phase [9]. With the advent of deep learning, feature extraction in AQA studies has also incorporated neural networks like long short-term memory (LSTM) networks [10] and 3-D convolution (C3D) neural networks [6], [11], [12]. These networks have proven to be effective in extracting features for AQA tasks. In the evaluation phase, existing AQA studies focus on three categories depending on the objectives of the assessment: regression scoring, grading, and pairwise sorting. Regression scoring [6] is usually used in sporting events, where referees assign scores to movements as ground truth. Models are trained on these videos to predict the scores of movements.

The development of wearable sensor technology has led to an increase in the usage of portable, affordable, and accurate wearable sensor devices in AQA research [13]. The signals collected in wearable device-based approaches include bioelectric and motion signals. Bioelectric signals generated by nerve cells, which control movements during exercise, are frequently characterized by a weak bio-current [14]. Commonly used bioelectric signals in AQA include electroencephalography (EEG) signals [14] and surface electromyography (sEMG) signals [15], [16]. The motion signals [17] collected in AQA include acceleration, angular velocity, and quadrature motion in gyroscopes, which to monitor rigid body movements.

However, the majority of the aforementioned approaches focus on coarse-grained actions. This article studies the quality assessment of sign language, which are fine-grained movements. The sign gesture quality is examined based on three gesture levels: amateur, skilled, and professional. The evaluation of sign language quality faces several challenges, which are primarily constrained by the intrinsic characteristics of sign language. First, compared to coarse-grained movements, sign language is a sort of fine-grained movement with smaller motion ranges and faster movement speeds, which leads to greater difficulty in evaluating movement quality. Moreover, variations in execution habits among different performers introduce additional inconsistencies, even among individuals with comparable skill levels.

To achieve an accurate assessment of the sign gesture quality, three quality metrics in sign language gestures and sentences are proposed to address the question. In terms of gestures, the palm position and finger flexibility are designed as quality metrics. The palm trajectory is calculated and mapped to position space with kernel density estimation, which gives insight into the movement deviation in the palm position metric. In the finger

Received 7 November 2024; revised 12 March 2025; accepted 14 March 2025. Date of publication 4 April 2025; date of current version 27 June 2025. This article was recommended by Associate Editor F. Scotti. *(Corresponding author: Qingshan Wang.)*

Zhiwen Zheng, Qingshan Wang, and Qi Wang are with the School of Mathematics, Hefei University of Technology, Hefei, Anhui 230601, China (e-mail: zhiwen.zheng@hdu.edu.cn; qswang@hfut.edu.cn; wangq@hfut.edu.cn).

Dazhu Deng is with Hefei Special Education Center, Hefei, Anhui 230041, China (e-mail: ddz5201@sina.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2025.3552476.

Digital Object Identifier 10.1109/THMS.2025.3552476

flexibility metric, the movement quality feature extracted from the instantaneous energy (IE) and curvature signal of the Bézier curve is associated with finger movements. In terms of sentences, the movement fluency metric is represented by texture features of the signal extracted through a modified Laplace operator and a constructed one-dimensional convolution (C1D) neural network. It shows the performer's familiarity with gestures. The final assessment results are obtained by combining the extracted quality features. The assessment using different metrics is considered, and the final gesture assessment results are generated by calculating different levels of confidence and information entropy between the quality metrics.

This study proposes SignEvaluator, an SLQA system with a movement quality feature extractor and an assessment generator. SignEvaluator evaluates the level of sign language execution of the performer based on the gestures performed by the performer.

Our main contributions are summarized as follows.

1) We proposed three quality metrics in terms of sign gestures and sentences in the movement quality feature extractor and extracted the corresponding movement quality features.
2) We developed an assessment generator to produce assessment results by combining the extracted movement quality features. In this procedure, Gaussian mixture distribution is first used to reconstruct the movement quality gestures. Then, the different levels of confidence and information entropy of the assessment results under different metrics are calculated to obtain the final assessment results.
3) We constructed a real SLQA dataset with 42 451 samples and assessed SignEvaluator's SLQA performance on this dataset. The experimental results present an F1-score of 0.89 in SignEvaluator.

The rest of this article is organized as follows. Related works on AQA are introduced in Section II. In Section III, SignEvaluator is proposed to realize SLQA. The model's performance is evaluated in Section IV. Finally, Section V concludes this article.

II. Related Works

Existing studies on AQA focused on *vision-based methods* and *wearable device-based methods*.

A. Vision-Based Methods

Vision-based approaches are widely adopted in sporting events [18], [19], [20] and daily living skills assessment [8]. The AQA process includes feature extraction and evaluation phase.

In the initial phase, the manual feature extraction method was first applied to AQA research. Wang et al. [8] focused on AQA and accurately assess the performance of complex movements in areas such as sports and medical procedures. Xu et al. [21] proposed a fine-grained diving action video dataset: FineDiving for evaluating action execution quality. Zhang et al. [22] divided the video into different segments and realized AQA by learning the changing relationship of actions between different segments through the temporal attention mechanism.

Meanwhile, Tang et al. [12] proposed a graph embedding unit that performs parallel convolution operations on RGB video signals in both channel and temporal domains. This lets complementary features be extracted across time and modalities. Recently, feature extraction has been used to neural networks, including the LSTM [10], [19], and C3D [6], [11] are applied for feature extraction. Xu et al. [10] proposed a deep architecture with a self-attentive LSTM and a multiscale convolution skip LSTM to extract local and global features of movement sequences, respectively. Wang et al. [11] proposed a tube self-attention network for AQA, which can generate extensive spatio-temporal contextual information on motion sequences by utilizing sparse feature interactions. In the evaluation phase, there are three types of AQA tasks: regression scoring, grading, and pairwise sorting. Gedamu et al. [23] proposed a fine-grained spatio-temporal parsing network composed of intrasequence action parsing module and spatio-temporal parsing module to evaluate small actions in videos. Zhu et al. [24] solved a variety of human-centered video tasks by learning human motion representations from large-scale and heterogeneous data resources. Parmer et al. [6] proposed a C3D-LSTM structure for movement feature selection and generating evaluation scores for sports motion. To address the data uncertainty present in the AQA dataset, Zhang et al. [25] developed a distributed autoencoder, which encodes videos as distributions and samples scores using reparameterization techniques.

To solve the difficulty of collecting high-quality action data and the diversity of specific actions or skill levels lead to the challenge of data scarcity in AQA research works, InceptionNet [26] are applied to capture the features of different levels through its multiscale convolution kernel structure, thereby improving the generalization ability of the model. Kothadiya et al. [27] enhanced InceptionV4 by optimizing backpropagation with uniform connections, and proposed an ensemble learning framework of different convolutional neural networks to further improve the recognition accuracy and robustness of model.

Vision-based methods are cost-effective and convenient for data collection. However, the assessment results obtained by these methods are sensitive to environmental factors such as camera angles and lighting conditions.

B. Wearable Device-Based Methods

Wearable devices have gained popularity in AQA due to their portability and low signal acquisition costs. These devices can collect signals that can be broadly categorized into two main categories: bioelectric signals and motion signals [28].

The former consists of EEG [14] signals and sEMG signals [15]. EEG signals provide detailed information about the activities of brain nerve cells on the surface of the cerebral cortex or scalp. This noninvasive technique captures the variations in electrical waves generated during brain activity. Vishnupriya et al. [14] studied the effects of magnitude-based weight trimming techniques on a motion classification task, completing the movement assessment with EEG signals. In contrast to EEG signal acquisition, which requires individuals to implant EEG sensor devices on the scalp, sEMG sensors can be worn anywhere on the body [29]. Lara et al. [16] gathered sEMG signals from a user's hand generated during movement to assess their

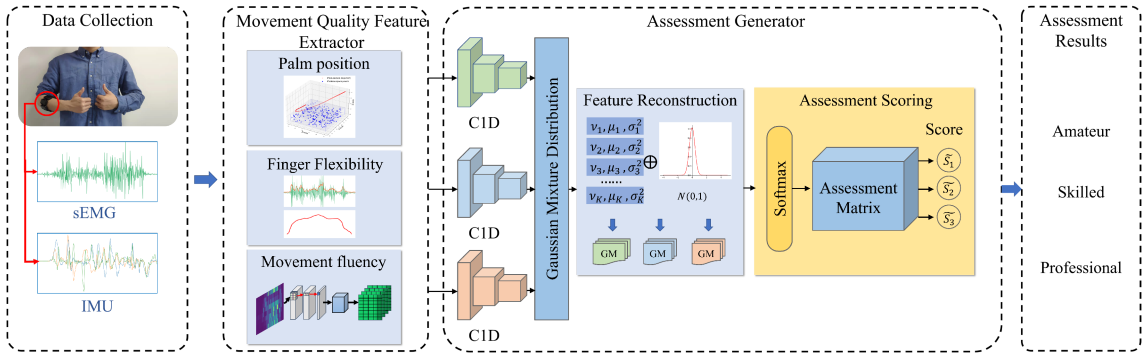

Fig. 1. System overview of SignEvaluator.

hand dexterity by accurately predicting 10 common hand gestures. Motion sensors are used in AQA studies to record changes in body positions during motion extraction. Dutta et al. [17] developed a glove equipped with six flexible sensors, three force sensors, and a motion processing unit to assess stroke patients' grasping ability and level of recovery.

In summary, approaches based on wearable devices have several benefits for gesture recognition, including portability, easy signal acquisition, and steady signal transmission. The bioelectric and motion sensors are combined to implement SLQA in this study.

III. Approach

As shown in Fig. 1, SignEvaluator consists of two modules: a movement quality feature extractor and an assessment generator.

A. Data Collection

In our investigation, data was gathered using a sensor bracelet. The bracelet included two types of sensors: eight sEMG sensors and an inertial measurement unit (IMU) sensor. The bracelet is worn on the upper part of the dominant hand (typically the right hand) during gesture execution. The signal sampling frequency of the sEMG sensor and IMU sensor are 200 Hz and 50 Hz, respectively. It is worth noting that the finger movement drives changes in arm muscles, causing changes in sEMG signals. IMU signals consist of three parts: 4-D gyroscopic quadrature signals, 3-D *X*–*Y*–*Z* axis acceleration signals, and 3-D *X*–*Y*–*Z* axis angular velocity signals.

B. Movement Quality Feature Extractor

In this section, three quality metrics are proposed, and movement quality features are extracted from sEMG and IMU signals based on the three quality metrics.

1) Quality Metrics:

Metric 1 (M_1). *Palm position:* Amateur sign language performers tend to place their palms higher than professional and skilled performers during gesture execution.

The metric focuses on quantifying the level of variation in motion when assessing sign gestures. To ensure that the gestures are executed accurately, amateur signers tend to instinctively raise their palms to positions where they have better visibility. Fig. 2 shows an amateur and skilled/professional performer's execution of the gesture "in."

Fig. 2. Metric 1 reflects the gesture "in" of different sign language gesture performers. (a) Amateur. (b) Skilled/Professional.

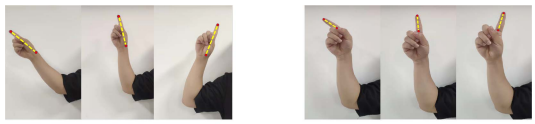

Fig. 3. Metric 2 reflects the gesture "what" of different sign language gesture performers. (a) Amateur/Skilled. (b) Professional.

Metric 2 (M_2). *Finger flexibility:* Professional performers have more flexible fingers and possess the ability to execute all finger movements within gestures more proficientl, in comparison to both amateur and skilled performers.

The professional performers, who have extensive experience in executing gestures, demonstrate greater flexibility in their finger joints, leading to more precise and fluid finger movements. Fig. 3 illustrates an amateur/skilled performer and professional performer's execution of the gesture "what."

Metric 3 (M_3). *Movement fluency:* Professional and skilled sign language performers have smooth transitions between gestures and imperceptible changes in the action magnitude of a sentence. In contrast, amateur performers complete gestures in an incoherent manner, and there is an apparent change in their magnitude of movements.

Assessed in terms of sentence execution, movement fluency, and unapparent action magnitude change indicate that the performer is more familiar with sign gestures, i.e., a higher level of execution. Fig. 4 illustrates an amateur performer and a skilled/professional performer performing the sign language gesture "everywhere."

(a) (b)

Fig. 4. Metric three reflects the gesture "everywhere" of different sign language gesture performers. (a) Amateur. (b) Skilled/Professional.

2) Feature Extraction: The movement quality feature is retrieved from the collected signals based on the proposed quality metrics. Before feature extraction, a Butterworth low-pass filter is used to filter out the high-frequency machine noise from sEMG and IMU signals, and 0 is used to fill signals at a fixed length. For example, in our work, the length of filled sEMG signals is $4T$, the length of filled IMU signals is T, and $T = 1000$ in our work.

Feature based on Metric 1: Sign language performers with different levels of gesture execution have inconsistent palm positions during gesture execution, which is revealed in the collected IMU signals.

The relative position feature description method is proposed for extracting palm position characteristics. This method aims to eliminate the deviation angle caused by wearing and map the collected signals to the position range space, which is taken as the gesture position feature. The method contains two steps.

In the first step, the deviation angle in the collected signal is eliminated using the rotation operator, and the palm motion trajectory is calculated. During gesture execution, a deviation angle θ_t between the bracelet coordinate system and the earth-fixed coordinate system occurs at t moment due to the bracelet-wearing position.

Suppose $\text{acc}_{x,t}$ is the tth frame X-axis acceleration signals, and the t moment deviation angle $\theta_t = [\theta_{x,t}, \theta_{y,t}, \theta_{z,t}]$. According to the Euler angle-quaternion conversion formula, the tth frame deviation angle θ_t can be calculated from the tth frame quaternion $Q_t = [q_{1,t}, q_{2,t}, q_{3,t}, q_{4,t}]$ as follows:

$$\begin{bmatrix}\theta_{x,t}\\ \theta_{y,t}\\ \theta_{z,t}\end{bmatrix} = \begin{bmatrix}\arctan(\frac{2(q_{0,t}q_{1,t}+q_{2,t}q_{3,t})}{1-2(q_{1,t}^2+q_{2,t}^2)})\\ \arcsin(2(q_{0,t}q_{2,t}-q_{3,t}q_{1,t}))\\ \arctan(\frac{2(q_{0,t}q_{3,t}+q_{1,t}q_{2,t})}{1-2(q_{2,t}^2+q_{3,t}^2)})\end{bmatrix}. \tag{1}$$

In t moment, the real X-axis acceleration signals $\overline{\text{acc}_{x,t}}$ can be obtained with the rotational inverse process Rot:

$$\overline{\text{acc}_{x,t}} = \text{acc}_{x,t}\text{Rot}_t^{-1} \tag{2}$$

where Rot_t can be calculated with θ_t as follows:

$$\text{Rot}_t = \begin{bmatrix}c_y c_x & s_z s_y c_x - c_z s_x & c_z s_y c_x + s_z s_x\\ c_y s_x & s_z s_y s_x + c_z c_x & c_z s_y s_x - s_z c_x\\ -s_y & s_z c_y & c_z c_y\end{bmatrix} \tag{3}$$

where, $c_x = \cos\theta_{x,t}, c_y = \cos\theta_{y,t}, c_z = \cos\theta_{z,t}, s_x = \sin\theta_{x,t}$, $s_y = \sin\theta_{y,t}$, and $s_z = \sin\theta_{z,t}$.

Moreover, the real position on the X-axis $d_{x,t}$ at t moment is calculated

$$d_{x,t} = \sum_{j=1}^{t}\left(\Delta t\sum_{i=1}^{j}(\overline{\text{acc}_{x,i}})\Delta t\right) \tag{4}$$

where Δt is the sampling interval of IMU signals, and $\Delta t = 0.02$. The real palm positions on the X-axis d_x can be represented as $d_x = \{d_{x,t}|1 \le t \le T\}$. It is necessary to obtain Y-axis real positions d_y and Z-axis real positions d_z in the same manner, and the palm trajectory is subsequently acquired during gesture execution.

In the second step, the calculated palm trajectory is mapped to nontemporal position space. The distributions of the positions chosen to correspond to sign gesture movements are first estimated. Suppose that the set of positions associated with the sign gesture is P_a. Through calculating the standard deviation of two adjacent positions, whether the position in P_a can be determined

$$\begin{cases}\frac{1}{K}\sum_{k=1}^{K}\sqrt{s_{k,4t}^2 - s_{k,4(t-1)}^2} \ge D, (d_{x,t}, d_{y,t}, d_{z,t}) \in P_a\\ \frac{1}{K}\sum_{k=1}^{K}\sqrt{s_{k,4t}^2 - s_{k,4(t-1)}^2} < D, (d_{x,t}, d_{y,t}, d_{z,t}) \notin P_a\end{cases} \tag{5}$$

where $s_{k,t}$ represents the kth channel sEMG signal at t moment, $1 < t \le T$. K is the number of sEMG signal channels. The threshold D is an empirical value obtained by experiments and $D = 3.2$ in our study.

To map the temporal gesture motion trajectory to nontemporal position space, P_a is represented as a mixture distribution $p(\varepsilon|\delta)$

$$p(\varepsilon|\delta) = \sum_{i=1}^{M}\upsilon_i y(\varepsilon|\mu_i, \sigma_i^2) \tag{6}$$

where ε is a random variable and represents the element in set P_a. υ_i, μ_i, and σ_i are the weight, mean, and variance of each distribution, respectively, and $\delta = \{\upsilon_i, \mu_i, \sigma_i^2\}$. M represents the freedom degree, and $M = 50$. The distribution $p(\varepsilon|\delta)$ indicates the change in palm position range during gesture execution. As one of the nonparametric estimation methods, kernel density estimation (KDE) can accurately estimate the distribution characteristics of samples. In our study, KDE is applied to estimate δ. The standard KDE equation is

$$f(\psi) = 1/nh\sum_{i=1}^{n}K((\psi - \psi_i)/h) \tag{7}$$

where $K(\cdot)$ is the kernel function, ψ represents the position samples and $K(\psi) \ge 0, \int K(\psi)d\psi = 1$. In this study, the Gaussian function is chosen as the kernel function. n is the number of samples and h is the bandwidth that can be obtained through Silverman's method [30]. With KDE estimation, the spatial position distribution of the performer's hands, i.e., (6), can be obtained.

Then, the extracted movement quality features are then less affected by the sentence when the palm trajectory is mapped to a position space. At t moment, the X-axis coordinates $d'_{x,t}$ of the points in position space corresponding to $d_{x,t}$ in palm trajectory

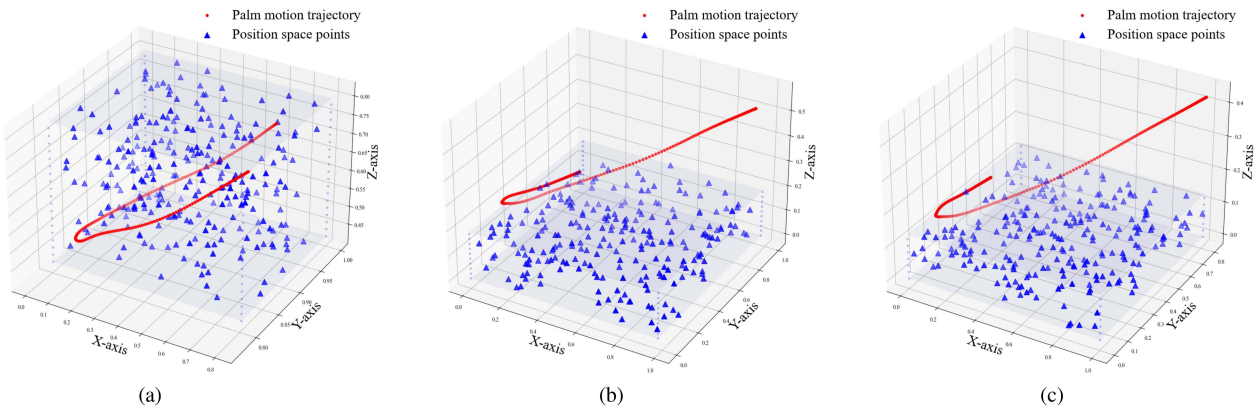

Fig. 5. Palm positions of an amateur performer, skilled and professional performer during gesture execution. (a) Amateur. (b) Skilled. (c) Professional.

are calculated based on the mixture distribution $p(\varepsilon|\delta)$

$$d'_{x,t} = \sum_{i=1}^{M} \upsilon_i(\mu_i\gamma + \sigma_i) \tag{8}$$

where the random variable γ follows $N(0,1)$ distribution. Similarly, the same method is used to determine the *Y*-axis and *Z*-axis coordinates of the points in position space $d'_{y,t}$ and $d'_{z,t}$, respectively. The position on *X*-axis can be represented as $d'_x = \{d'_{x,t}|1 \le t \le T\}$. Position (d'_x, d'_y, d'_z) are taken as the *palm position feature* corresponding to metric 1 for a gesture sample, with a size of $3 \times T$.

Fig. 5 shows the palm motion trajectory and position space points of an amateur, skilled, and professional performer. It indicates that when the performer becomes more proficient in sign gesture execution, the palm position space is lower in points.

Feature based on Metric 2: The finger flexibility metric is indicated by the signal strength and coherence of the sEMG signals. In terms of signal strength, sEMG signals with more dramatic fluctuations indicate strenuous finger joints activity. Thus, IE was chosen as the unit of measurement to express signal variation. The IE of the kth channel sEMG signal at t moment $s_{k,t}$ can be calculated as follows:

$$ie_{k,t} = 1/h\sqrt{\sum\nolimits_{i=t}^{t+h} s_{k,i}^2} \tag{9}$$

where $1 \le k \le 8$ for sEMG signals. h is the sliding window width and an empirical value, $h = 12$ in our work. Based on (9), the feature $\varpi = [ie_{k,t}]_{8\times 4T}$ of sEMG signals can be obtained. Fig. 6 shows a channel sEMG signal and corresponding IE collected from an amateur/skilled and a professional performer. The results show that IE is higher when there are strong signal fluctuations and lower IE when there are weak signal fluctuations. The IE of the professional performer is larger than that of the amateur performer.

For the latter, amateur/skilled performers frequently overlook finger movements. Professional signers' sEMG signals are locally smooth, whereas amateur and proficient performers' signals lack that. In our study, the discrete sEMG signals are transformed into a locally continuous curve, and the radian of each segment curve is calculated as part of the finger flexibility feature. The specific steps are described as follows.

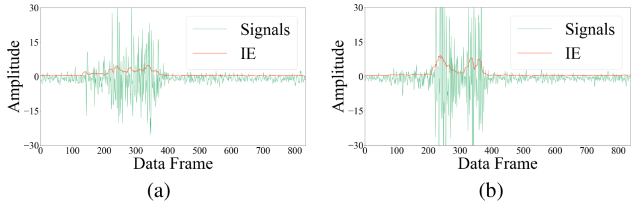

Fig. 6. The sEMG signals and corresponding IE on the gesture 'what' of an amateur/skilled and professional performer. (a) Amateur/Skilled. (b) Professional.

First, the sEMG signals are split into segments with a length of 10. Suppose that the lth $(1 \le l \le \lfloor 4T/c \rfloor)$ segment of $kath$ $(1 \le k \le 8)$ channel sEMG signals represented as $\{s_{k,c(l-1)}, s_{k,c(l-1)+1}, \ldots, s_{k,cl}\}$, where $s_{k,cl}$ represents the cth frame kth channel sEMG signals on lth segment. c is the size of sEMG signal segment. c is an empirical value, and $c = 10$.

The Bézier curve [31] is then calculated to fit each segment of the sEMG signal. As a local fit curve, each data point is treated as a control point in the fitting process. In contrast to other fitting methods, Bézier curve fitting closely encapsulates the influence of each data point on the entire curve. The composition of the Bézier fit curve of the lth segment of kth channel sEMG signals at mth frame can be represented as follows:

$$B_{k,m} = \sum_{i=0}^{c} \binom{c}{i} s_{k,lc-c+i} \cdot (1-m)^{n-i} \cdot m^i \tag{10}$$

where $0 \le m \le c$.

Last, the curvature of each segment of the Bézier fit curve is calculated to evaluate the smoothness of signals and is considered as one of the features of these signals. As shown in Fig. 7, to quantify the curvature at each Bézier fit curve frame, the external circle $O_{k,m}$ that passes through the ends of Bézier fit curve $(0, B_{k,0})$ and $(c, B_{k,c})$ at kth channel is developed to replace the curvature circle at point $(k, m, B_{k,m})$. The curvature $\rho_{k,m}$ at $(k, m, B_{k,m})$ is calculated after determining the radius of circle O_m. The curvature $\rho_{k,m}$ of point $(k, m, B_{k,m})$ can be

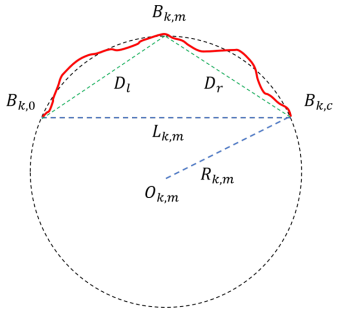

Fig. 7. Diagram of an external circle of a Bézier curve.

calculated as follows:

$$\rho_{k,m} = 1/R_{k,m} = 4S_{\triangle B_{k,0}B_{k,m}B_{k,c}}/L_{k,m}D_lD_r \tag{11}$$

where $S_{\triangle B_{k,0}B_{k,m}B_{k,c}}$ represents the area of triangle formed by the three points $(0, B_{k,0})$, $(m, B_{k,m})$, and $(c, B_{k,c})$, which can be calculated with $L_{k,m}$, D_l, and D_r. Thus, the curvature feature $\Omega = [\rho_{k,t}]_{8\times 4T}$ of sEMG signals is obtained.

The normalized IE of sEMG signals ϖ' and normalized curvature feature Ω' are combined together as the *finger flexibility feature* of a gesture sample $[\varpi', \Omega']^T$.

Feature based on Metric 3: As shown in Fig. 8, spectrograms of sEMG signals obtained from amateur performers show an ostensible textural feature, which is associated with a discontinuity in the signals. In contrast, the sEMG signals collected from skilled/professional performers exhibit a smoother transition without any apparent discontinuities or textural features.

Based on this phenomenon, a Laplace operator-based C1D (Laplace-C1D) neural network is constructed to extract the textural properties of sEMG signals. The Laplace operator captures the local changes in gradient of signals, which can be represented as follows:

$$\Delta f = \bigtriangledown^2 f = \partial^2 f/\partial x^2 \tag{12}$$

where x is an independent variable and f is a function of x. To demonstrate the local gradient characteristics of sEMG signals under different receptive fields, the Lagrange mean value theorem is combined to improve the computation of the Laplace operator as follows:

$$\begin{aligned}\partial^2 f/\partial x^2 &= 1/k[f'(x+k) - f'(x)] \\ &= 1/k^2[f(x+k) + f(x-k) - 2f(x)]\end{aligned} \tag{13}$$

where x stands for the object that the Laplace operator will be computing. Parameter k controls the size of sensation field for feature extraction, and k is assigned 1, 3, and 5 in our work.

The framework of the Laplace-C1D neural network is shown in Fig. 9. The backbone of the Laplace-C1D neural network is three C1D layers with the Laplace operator, and $k = 1$, $k = 3$, and $k = 5$ for the three layers, respectively. To maintain the signal's scale invariance, padding is performed before each C1D operation. Later, the normalization layer normalizes the extracted texture features. It is worth noting that since the convolution kernels of the Laplace-C1D neural network are initialized as different Laplace operators, the Laplace-C1D neural network does not need to be trained. Through the Laplace-C1D neural network, the *movement fluency feature* β of a sign language sentence sample can be obtained, and the size of β is $8 \times 4T$.

C. Assessment Generator

The assessment generator combines the gesture features extracted based on the quality metrics to generate quality assessment results. This section introduces the structure of assessment generator and training process with the developed loss function.

1) Module Structure: As shown in Fig. 10, the assessment generator consists of a feature learning module, a Gaussian mixture module and an assessment scoring module. The inputs of assessment generator are palm position feature (d'_x, d'_y, d'_z), finger flexibility feature $[\varpi', \Omega']^T$, and movement fluency feature β.

To improve model's generalization, the Gaussian mixture module reconstructs movement quality features through Gaussian mixture distribution. The distribution of h_i can be defined as follows:

$$P(h_i|\varphi_i) = \sum_{k=1}^{K} w_{i,k}\phi(h_i|\varphi_{i,k}) \tag{14}$$

where $w_{i,k}(i = 1, 2, 3)$ is the weight of kth Gaussian mixture distribution of ith gesture features. K represents the degree of freedom in a Gaussian mixture distribution, which is an empirical value that decreases as the batch size increases during the training process. In our work, K is set to 200, while the batch size is 1024. $\varphi_i = \{\varphi_{i,k}|1 \le k \le K\}$. $\phi(h_i|\varphi_{i,k})$ is the kth Gaussian distribution of h_i, which is defined as follows:

$$\phi(h_i|\varphi_{i,k}) = \frac{1}{\sqrt{2\pi}\sigma_{i,k}} \exp\left(-\frac{(h_i - \mu_{i,k})^2}{2\sigma_{i,k}^2}\right) \tag{15}$$

where $\mu_{i,k}$ and $\sigma_{i,k}^2$ are the mean and variance, $\varphi_{i,k} = (\mu_{i,k}, \sigma_{i,k}^2)$. As shown in Fig. 10, the learned movement quality feature h_i is reconstructed as $\overline{h_i}$, where $\overline{h_i} = \{\overline{h_{i,n}}|1 \le n \le N\}(i = 1, 2, 3)$, this process is shown as follows:

$$\overline{h_{i,n}} = \sum_{k=1}^{K} w_{i,k}(\mu_{i,k} \cdot \Gamma + \sigma_{i,k}) \tag{16}$$

where Γ is a matrix containing a random number that conforms to the $N(0, 1)$ distribution, and the size of Γ is $K \times 1$.

The assessment scoring module combines the weight of quality metrics and the confidence of gesture levels to generate assessment results. Suppose that $p_{i,j}(1 \le i \le 3, 1 \le j \le 3)$ represents the probability of a sample belonging to ith gesture level evaluated by jth quality metric.

First, the information entropy of jth metric of different gesture level H_j is calculated as follows:

$$H_j = -1/\ln 3 \sum_{i=1}^{3} p_{i,j} \cdot \ln p_{i,j}. \tag{17}$$

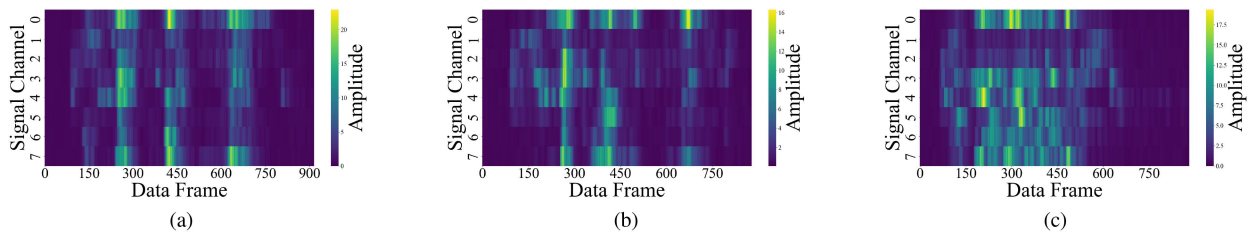

Fig. 8. Raw sEMG signals of a gesture collected by different level performers. (a) Amateur. (b) Skilled. (c) Professional.

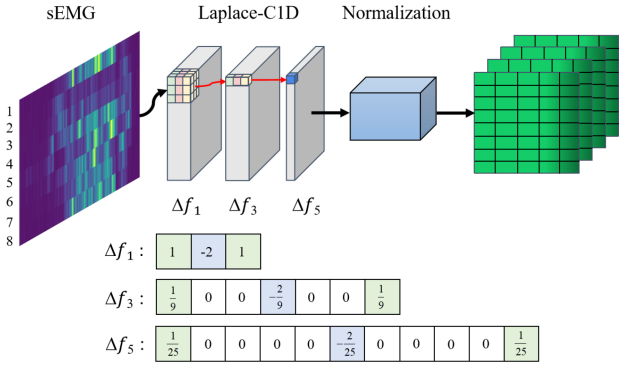

Fig. 9. Framework of Laplace-C1D neural network.

The weight of jth quality metric ω_j is defined as follows:

$$\omega_j = (1 - H_j) / \sum_{k=1}^{3} (1 - H_k). \tag{18}$$

The higher ω_j, the more influence jth quality metric has on the final assessment results.

The classification error e_i for ith gesture level is calculated as follows:

$$e_i = \sum_{j=1}^{3} [1 - \overline{p_{i,j}} \ln(p_{i,j})] \tag{19}$$

where $\overline{p_{i,j}}$ is ground truth of the sign language gesture sample. The confidence of ith gesture level can be defined based on (19)

$$\eta_i = 1/3 \ln((1 - e_i)/e_i). \tag{20}$$

With the decrease of e_i, η_i rises. η_i reflects the confidence of model to ith gesture level, which is beneficial to improve the model's assessment performance.

Moreover, based on (18) and (20), the assessment score S_i of ith gesture level is calculated by multiplying the weighted sum of quality metrics $\omega_j, p_{i,j} (1 \le j \le 3)$ with the confidence η_i

$$S_i = \eta_i \sum_{j=1}^{3} \omega_j \cdot p_{i,j}. \tag{21}$$

Through normalization, the final assessment score $\widetilde{S}_i$ can be obtained. The category with the highest final assessment score is the assessment level. An executor's proficiency in sign language can be assessed through three critical the position of the palm, the dexterity of the fingers, and the fluency of movements. For instance, to an amateur performer, SignEvaluator might recommend enhancing their sign language skills by refining gesture positions, engaging in finger flexibility exercises, and focusing on the transitions between gestures within sentences.

2) Module Training: The training process of the assessment generator involves developing movement quality features and evaluating performers based on the quality of their sign language gesture sample. First, the cosine similarity is applied to define the distance $L_i (i = 1, 2, 3)$ between the reconstructed features $\overline{h_i} (i = 1, 2, 3)$ and feature learning module learned features $h_i (i = 1, 2, 3)$

$$L_i = -\frac{\sum_{n=1}^{N} h_{i,n} \times \overline{h_{i,n}}}{\sqrt{\sum_{n=1}^{N} (h_{i,n})^2} \times \sqrt{\sum_{n=1}^{N} (\overline{h_{i,n}})^2}} \tag{22}$$

where and N represents the number of samples. The assessment generator is then trained to provide better assessment results by building a cross-entropy loss function L_4

$$L_4 = -\sum_{n=1}^{N} Y_n^* \log(Y_n) \tag{23}$$

where Y_n^* and Y_n represent the nth ground truth and the assessment result of nth sign language signal sample obtained by module, respectively. The assessment generator is trained by minimizing the joint loss function L, which can be calculated

$$L = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3 + \lambda_4 L_4 \tag{24}$$

where $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$.

IV. Performance Evaluation

In this section, the performance of SignEvaluator[1] is evaluated. The experiments are conducted using a MYO bracelet and a PC. The PC consists of an Intel Core i9-10900 K CPU and an Nvidia GeForce RTX 3090 GPU with 24 GB graphics memory. The operating system and deep learning framework of the PC is Ubuntu 18.04 and torch-GPU (version 1.10.1, Python 3.7.11).

A. Dataset Description

To evaluate the effectiveness of SignEvaluator, we selected 702 commonly used sign language sentences, covering various domains of everyday life such as financial services, healthcare, education, and retail. Each sentence consists of 3–10 gestures.

[1][Online]. Available: https://codeocean.com/capsule/2741663/tree

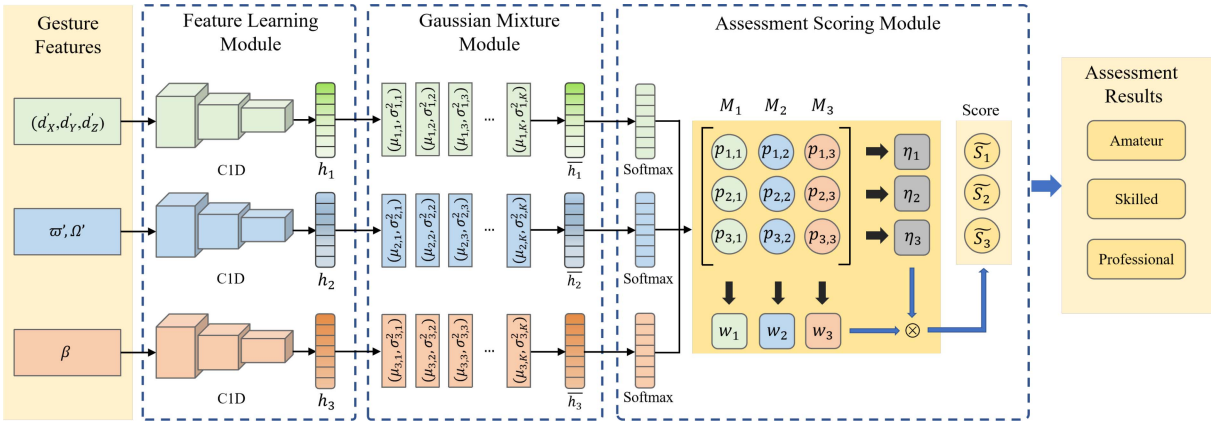

Fig. 10. Structure diagram of the assessment generator.

A total of 20 volunteers from various age groups were invited to participate in the sample collection. These volunteers included sign language users of varying proficiency levels. Their sign language proficiency was thoroughly assessed by experts, and they were categorized into three skill levels: amateur, skilled, and professional. The amateur group consisted primarily of university students who underwent three months of sign language training prior to participating in the study to ensure their accurate use of sign language. The skilled group was composed of hearing impairments who had received education in special schools and frequently used sign language in their daily lives. The professional group comprised experienced sign language experts with extensive understanding and practice in the application of sign language. Notably, all participants were right-handed, a condition maintained for experimental consistency.

During dataset acquisition, we applied the MYO bracelet to capture sign language movements. The MYO sensor contains 8 sEMG sensors and an IMU sensor. Before the experiment, the sensors went through a rigorous calibration process to ensure the accuracy of the data. We collected 42 451 sign language gesture samples over a period of six months. A portion of the samples was randomly selected as the training set, while the remaining 30% was used as the test set. It is worth noting that both the training and test sets were composed of sign language signal samples, with the test set additionally including data from users not present in the training set.

B. Experimental Setup

In the experiments, precision rate p, recall rate r, and F1-score f are selected as metrics to evaluate the assessment performance of SignEvaluator. They can be calculated as follows:

$$\begin{cases} p = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \\ r = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \\ f = 2\cdot p\cdot r/(p+r) \end{cases} \tag{25}$$

where TP denotes the true positive, FP is the false positive, and FN is the false negative. The optimal combination of these hyperparameter is $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.3$, and $\lambda_4 = 0.5$ through experiments.

TABLE I
STATISTICAL DATASETS

Groups	Number of volunteers	Sample size
Amateur group	9	11, 879
Skilled group	7	15, 645
Professional group	4	14, 935

TABLE II
IMPACT OF SIGNAL PROCESSING APPROACH IN MOVEMENT QUALITY FEATURE EXTRACTION (# A : AMATEUR, # S : SKILLED, AND # P PROFESSIONAL)

F1-score	# A	# S	# P	Average
SignEvaluator	0.88	0.87	0.92	0.89
Without deviation angle elimination	0.84	0.80	0.85	0.83
Without Laplace operator	0.84	0.82	0.86	0.84
Replace Laplace-C1D with a CNN	0.83	0.82	0.81	0.82

Generally, for a signer, his proficiency of sign language sentences is not likely to exhibit significant improvements over a brief period. Moreover, there is a considerable disparity in the skill levels among performers, which makes the execution of sign language sentences a more accurate mirror of their proficiency. Consequently, we have enlisted the specialists to undertake a comprehensive assessment of the signers. According to different levels of sign language executors, the sign language sentence samples are divided into three different parts: amateur, skilled, and professional. The F1-score for each of the three proficiency levels are calculated.

C. Component Evaluation

To evaluate the performance of SignEvaluator, we assess its effectiveness by removing or replacing specific components within the system.

Impact of deviation angle θ elimination: In this experiment, the palm position feature (d'_x, d'_y, d'_z) is calculated by *X*/*Y*/*Z*-axis acceleration signals $\mathrm{acc}_x, \mathrm{acc}_y, \mathrm{acc}_z$ without eliminating the deviation angle θ. Table II presents the assessment performance of SignEvaluator. It can be seen that the deviation angle θ elimination operation can raise the F1-score of SignEvaluator by 7.23%.

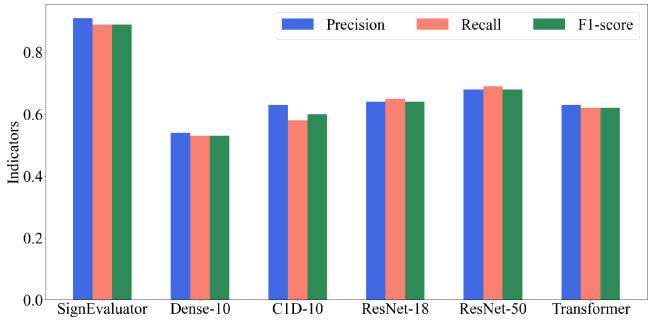

Fig. 11. SLQA performance comparison with baseline model.

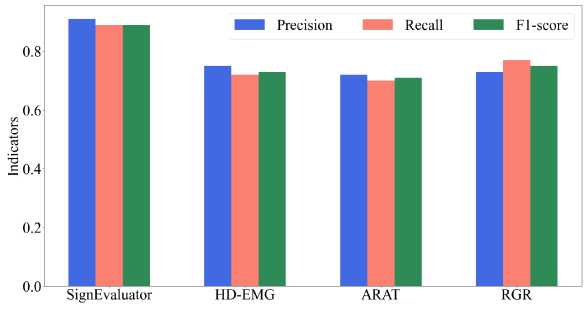

Fig. 12. SLQA performance comparison with SOTA.

Impact of Laplace operator: In this experiment, the convolution kernels of the Laplace operator-based C1D neural network are replaced with randomly initialized convolution kernels. The new convolution kernels are generated with random numbers between 0 and 1. It is important to note that SignEvaluator's training phase does not make use of the modified Laplace operator-based C1D neural network. The content of convolution kernels for training SignEvaluator is randomly transformed 5 times, and the F1-score of the model is calculated on an average of 5 times. The assessment performance of SignEvaluator is shown in Table II, which indicates that the Laplace operator improves the F1-score of SignEvaluator by 5.95%.

Impact of Laplace-C1D module: In this experiment, the performance of Laplace-C1D model are evaluated. The Laplace-C1D model are replaced with a CNN model, which contains three convolutional layers with kernel sizes of 5×5, 3×3, and 2×2, respectively. It is worth noting that the size of the signal features extracted by CNN model and Laplace-C1D model remains consistent. The performance of SignEvaluator is presented in Table II, which indicates that the Laplace-C1D module improves the F1-score of SignEvaluator by 8.54%.

D. Comparison With Existing Methods

Comparison with baseline models: SignEvaluator is compared with several baseline classification models, including the Dense-10 [32], C1D-10 [33], ResNet-18 [32], ResNet-50 [32], and Transformer [34] models. All models use an Adam optimizer with an initial learning rate of 0.001, a batch size of 64, and employ early stopping based on validation loss. Notably, the different modality signals are concatenated via a three-layer fully connected network before classification using a softmax activation function. Fig. 11 demonstrates the comparison results on the test set, which show that SignEvaluator performs better than the four baseline models in the SLQA task.

Comparison with the state of the art (SOTA): An experiment comparing SignEvaluator with SOTA is conducted. In the experiments, HD-EMG [16], ARAT [17], and RGR [19] are adopted as SOTA methods. A comparative study is conducted to evaluate the performance of SignEvaluator. The inputs of HD-EMG [16] and ARAT [17] are sEMG and IMU signal samples, respectively. The input of RGR [19] contains both sEMG and IMU signal samples.

For HD-EMG [16], the time-domain and temporal-spatial sEMG features were extracted in sEMG signal samples to train support vector machine learning classifiers. The IMU signal samples are handled in this experiment using the data preprocessing methods, feature extraction methods, and feature selection methods from ARAT [17] are applied to handle the IMU signal samples. And the optimally adjusted support vector classifier is used to classify different quality sign gesture samples. In RGR [19], a Siamese architecture was applied to measure the distance between two actions. In this experiment, we re-implement this architecture and conduct SLQA.

Fig. 12 shows the experimental results, which demonstrated that the F1-score of SignEvaluator is 17.98% and 20.22% higher than HD-EMG [16] and ARAT [17], respectively. The reason is that the comparison studies included only single-modal signals, resulting in limited data availability, and the proposed method addresses the sign language quality traits by incorporating targeted evaluation metrics and feature extraction. Moreover, the F1-score of SignEvaluator is 15.73% higher than that of RGR [19]. This difference can be attributed to the incorporation of Gaussian reconstruction in the assessment generator of SignEvaluator, which enhances the model's generalization performance.

TABLE III
THE RESULTS OF INDEPENDENT JUDGMENT ARE COMPARED WITH SOTA METHODS

Levels	Volunteers (No.)	Predict			
		SignEvaluator	HD-EMG	ARAT	RGR
# A	1	0.947	0.817	0.810	0.800
	2	0.930	0.773	0.750	0.765
	3	0.930	0.780	0.750	0.758
	4	0.919	0.765	0.734	0.727
	5	0.901	0.757	0.726	0.742
# S	6	0.864	0.750	0.734	0.750
	7	0.883	0.734	0.684	0.693
	8	0.895	0.684	0.667	0.693
# P	9	0.907	0.830	0.898	0.802
	10	0.930	0.817	0.810	0.824

A : Amateur, # S : Skilled, and # P Professional.

E. Independent Judgment

To validate the cross-user generalizability of SignEvaluator, 10 additional participants are recruited, including 5 university students with normal hearing (3 men and 2 women), 3 hearing-impaired people who are proficient in sign language (1 man and 2 women), and 2 sign language specialists (1 man and 1 woman) with more than 10 years of teaching experience. The participants are instructed to execute 20 randomly selected sign language

sentences. Each sentence is performed 5 times, and 1,000 test samples are collected. SignEvaluator is evaluated on these samples. As shown in Table III, SignEvaluator achieved superior F1-scores, consistently outperforming sEMG-based (HD-EMG [16]), IMU-based (ARAT [17]), and multimodal-fusion (RGR [19]) baselines by 8.7%–22.3% margins, confirming its robustness across diverse proficiency levels and execution styles.

V. Conclusion

Sign language is a basic form of communication for hearing-impaired individuals. This article proposes SignEvaluator, a system for evaluating the quality of sign language that includes an evaluation generator and an extractor of movement quality features. In the movement quality feature extractor, three quality metrics are proposed for sign language gestures and sentences. The palm position metric and finger flexibility metric are designed to evaluate gestures in sign language. The palm trajectory is mapped to position space with kernel density estimation to indicate movement deviation. The IE and Bézier curvature of gesture signals are extracted to represent finger movements. The movement fluency metric for sign language sentences uses a C1D network and a modified Laplace operator to indicate the performer's familiarity with gestures. In the assessment generator, movement quality features are reconstructed using Gaussian mixture distribution to enhance model generalization. The final assessment results are calculated by combining different levels of confidence and information entropy under different metrics. The results indicate that SignEvaluator obtained an F1-score of 0.89 for 702 sentences collected from 20 performers. As part of future research, we will quantify a performer as a score based on a series of sign language sentences.

References

[1] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 505–523, Apr.–Jun. 2021.

[2] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li, "Improving sign language translation with monolingual data by sign back-translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1316–1325.

[3] J.-H. Pan, J. Gao, and W.-S. Zheng, "Adaptive action assessment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8779–8795, Dec. 2022, doi: 10.1109/TPAMI.2021.3126534.

[4] T. Ogasawara, H. Fukamachi, K. Aoyagi, S. Kumano, H. Togo, and K. Oka, "Archery skill assessment using an acceleration sensor," *IEEE Trans. Hum.-Mach. Syst.*, vol. 51, no. 3, pp. 221–228, Jun. 2021.

[5] S.-J. Zhang, J.-H. Pan, J. Gao, and W.-S. Zheng, "Semi-supervised action quality assessment with self-supervised segment feature recovery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6017–6028, Sep. 2022, doi: 10.1109/TCSVT.2022.3143549.

[6] P. Parmar and B. Morris, "Action quality assessment across multiple actions," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2019, pp. 1468–1476.

[7] T. Nagai, S. Takeda, M. Matsumura, S. Shimizu, and S. Yamamoto, "Action quality assessment with ignoring scene context," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 1189–1193.

[8] S. Wang et al., "A survey of video-based action quality assessment," in *Proc. Int. Conf. Netw. Syst. AI*, 2021, pp. 1–9.

[9] A. Calado, P. Roselli, V. Errico, N. Magrofuoco, J. Vanderdonckt, and G. Saggio, "A geometric model-based approach to hand gesture recognition," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 52, no. 10, pp. 6151–6161, Oct. 2022.

[10] C. Xu, Y. Fu, B. Zhang, Z. Chen, Y.-G. Jiang, and X. Xue, "Learning to score figure skating sport videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4578–4590, Dec. 2020.

[11] S. Wang, D. Yang, P. Zhai, C. Chen, and L. Zhang, "TSA-NET: Tube self-attention network for action quality assessment," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 4902–4910.

[12] S. Tang, D. Guo, R. Hong, and M. Wang, "Graph-based multimodal sequential embedding for sign language translation," *IEEE Trans. Multimedia*, vol. 24, pp. 4433–4445, 2022, doi: 10.1109/TMM.2021.3117124.

[13] Z. Wang et al., "Hear sign language: A real-time end-to-end sign language recognition system," *IEEE Trans. Mobile Comput.*, vol. 21, no. 7, pp. 2398–2410, Jul. 2022.

[14] V. R., N. Robinson, R. Reddy M., and C. Guan, "Performance evaluation of compressed deep CNN for motor imagery classification using EEG," in *Proc. 43 rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2021, pp. 795–799.

[15] Y. Liu, X. Li, L. Yang, and H. Yu, "A transformer-based gesture prediction model via semg sensor for human-robot interaction," *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, Art. no. 2510615, doi: 10.1109/TIM.2024.3373045.

[16] J. E. Lara, L. K. Cheng, and N. Paskaranandavadivel, "Muscle-specific high-density electromyography arrays for hand gesture classification," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 5, pp. 1758–1766, May 2022.

[17] D. Dutta, S. Aruchamy, S. Mandal, and S. Sen, "Poststroke grasp ability assessment using an intelligent data glove based on action research arm test: Development, algorithms, and experiments," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 2, pp. 945–954, 2022.

[18] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Group-aware contrastive regression for action quality assessment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7899–7908.

[19] H. Jain, G. Harit, and A. Sharma, "Action quality assessment using siamese network-based deep metric learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2260–2273, Jun. 2021.

[20] H.-Y. Li, Q. Lei, H.-B. Zhang, and J.-X. Du, "Skeleton based action quality assessment of figure skating videos," in *Proc. 11th Int. Conf. Inf. Technol. Med. Educ.*, 2021, pp. 196–200.

[21] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, and J. Lu, "Finediving: A fine-grained dataset for procedure-aware action quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2949–2958.

[22] Y. Zhang, W. Xiong, and S. Mi, "Learning time-aware features for action quality assessment," *Pattern Recognit. Lett.*, vol. 158, pp. 104–110, 2022.

[23] K. Gedamu, Y. Ji, Y. Yang, J. Shao, and H. T. Shen, "Fine-grained spatio-temporal parsing network for action quality assessment," *IEEE Trans. Image Process.*, vol. 32, pp. 6386–6400, 2023.

[24] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "MotionBERT: A unified perspective on learning human motion representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 15085–15099.

[25] B. Zhang et al., "Auto-encoding score distribution regression for action quality assessment," *Neural Comput. Appl.*, vol. 36, pp. 929–942, 2024, doi: 10.1007/s00521-023-09068-w.

[26] L. Zou, M. S. Munir, Y. K. Tun, S. S. Hassan, P. S. Aung, and C. S. Hong, "When hierarchical federated learning meets stochastic game: Toward an intelligent UAV charging in urban prosumers," *IEEE Internet Things J.*, vol. 10, no. 12, pp. 10438–10461, Jun. 2023.

[27] D. R. Kothadiya, C. M. Bhatt, H. Kharwa, and F. Albu, "Hybrid inception-net based enhanced architecture for isolated sign language recognition," *IEEE Access*, vol. 12, pp. 90889–90899, 2024.

[28] J. He, H. He, A. Pradhan, and N. Jiang, "Low-latency gesture recognition from spatial filtering of single-element ultrasound signals," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 2517909, doi: 10.1109/TIM.2023.3271722.

[29] T.-Y. Pan, W.-L. Tsai, C.-Y. Chang, C.-W. Yeh, and M.-C. Hu, "A hierarchical hand gesture recognition framework for sports referee training-based EMG and accelerometer sensors," *IEEE Trans. Cybern.*, vol. 52, no. 5, pp. 3172–3183, May 2022.

[30] A. R. Forrest, "Interactive interpolation and approximation by Bézier polynomials," *Comput. J.*, vol. 15, no. 1, pp. 71–79, 1972.

[31] G. G. Lorentz, "Bernstein polynomials," 2nd Ed., Chelsea Publishing Co., New York, 1986.

[32] Y. Huang, X. Li, Z. Du, and H. Shen, "Spatiotemporal enhancement and interlevel fusion network for remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5609414.

[33] X. Han, F. Lu, J. Yin, G. Tian, and J. Liu, "Sign language recognition based on R(2+1)D with spatial–temporal–channel attention," *IEEE Trans. Hum.-Mach. Syst.*, vol. 52, no. 4, pp. 687–698, Aug. 2022.

[34] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.