Sensors and Materials, Vol. 38, No. 5 (2026) 2723–2738
MYU Tokyo

S & M 4465

Taiwanese Sign Language Recognition and Natural Sentence Generation System Based on Spatiotemporal Graph Convolutional Networks and Distilled Bidirectional Encoder Representations from Transformers

Neng-Sheng Pai, Li-An Weng, Pi-Yun Chen,* and Lian-Sheng Hong

Department of Electrical Engineering, National Chin-Yi University of Technology, Taichung 41170, Taiwan (ROC)

(Received December 22, 2025; accepted April 14, 2026)

Keywords: sign language recognition, natural sentence generation, MediaPipe, ST-GCN, DistilBERT

We present a Taiwanese Sign Language (TSL) recognition and natural sentence generation system that focuses on continuous sign language recognition, in contrast to most existing approaches that primarily address isolated sign recognition. The proposed system integrates a spatiotemporal graph convolutional network (ST-GCN) with a distilled bidirectional encoder representations from transformers (DistilBERT)-based language generation model, with the aim of reducing communication barriers for the deaf and hard-of-hearing community. First, a camera sensor is used to capture sign language videos. MediaPipe is then utilized to extract human body key points from sign language video sequences. These spatiotemporal key point representations are subsequently processed by the ST-GCN model to perform sign recognition. Finally, the recognized sign sequences are translated into fluent and natural sentences using a fine-tuned DistilBERT model. Experimental evaluations are conducted on a self-collected dataset consisting of 42 classes of TSL videos, along with a frame sampling analysis. The results indicate that uniformly sampling video sequences to 70 frames yields the best recognition performance for the ST-GCN model. For sentence generation, 24 predefined Chinese sentence templates are employed to fine-tune the DistilBERT model. Experimental results indicate that the proposed method can achieve accurate and natural sentence generation under low-resource training conditions. Overall, the proposed system exhibits strong performance in terms of lightweight model architecture, robust gesture recognition accuracy, and natural language generation quality, thereby validating its effectiveness and feasibility for continuous sign language translation and language generation tasks.

1. Introduction

For deaf and hard-of-hearing individuals, sign language serves as a primary means of communication. According to statistics reported by the World Health Organization (WHO),[1] approximately 1.5 billion people worldwide suffer from hearing loss, accounting for nearly 20%

*Corresponding author: e-mail: chenby@ncut.edu.tw
https://doi.org/10.18494/SAM6144

ISSN 0914-4935

of the global population. In Taiwan, data published by the Ministry of Health and Welfare[2] indicate that the number of individuals with hearing impairments increased from 124825 in 2020 to 138966 in 2024, representing an approximately 11% growth over five years. This trend highlights the continuous rise in the hearing-impaired population and the corresponding increase in communication-related demands.

However, most people in society do not possess sign language proficiency, which often results in significant communication barriers and social exclusion for deaf and hard-of-hearing individuals in daily life, education, and employment. Moreover, sign language differs substantially from spoken and written languages in terms of grammatical structure and word order, making sign language translation a highly demanded yet challenging research problem.

To address this issue, we propose a Taiwanese Sign Language (TSL) translation system. The proposed system first uses a camera sensor to capture sign language videos. It then utilizes the Holistic and Hands modules of MediaPipe[3] to perform human pose estimation (HPE) and constructs customized upper-body key point coordinate sequences. These spatiotemporal representations are fed into a spatiotemporal graph convolutional network (ST-GCN)[4] for sign gesture recognition, where the outputs correspond to sign-level vocabulary tokens. Subsequently, the recognized tokens are translated into grammatically correct and semantically complete Chinese sentences using a natural language model based on distilled bidirectional encoder representations from transformers (DistilBERT).[5] Finally, a graphical user interface (GUI) is designed to provide users with intuitive interaction and clear visualization of the translation results.

2. Related Work

Here, we review related work on human pose estimation, action recognition, and natural language models, with a focus on the techniques directly applied in the proposed system.

2.1 Human pose estimation

The aim of HPE is to locate human body key points from visual input. Traditional methods, such as part-based models and depth sensors (e.g., Microsoft Kinect),[6] were limited by occlusion problems and high computational cost. Later, deep learning-based methods such as DeepPose[7] and OpenPose[8] achieved significant improvements but often required heavy computation, making them less suitable for real-time applications.

To address these limitations, in this study, we adopt MediaPipe, a lightweight framework designed for real-time HPE. MediaPipe applies a single-stage convolutional neural network (CNN)[9] architecture to directly predict all body joints without multistage refinement, which reduces latency and resource consumption. It provides modules for upper-body and hand landmark detection, enabling the efficient extraction of key points required for sign language recognition.

2.2 Action recognition

Early approaches to action recognition relied on handcrafted features such as optical flow or dynamic time warping, which are effective only in simple scenarios. With the advent of deep learning, CNNs were used for extracting spatial features, while recurrent neural networks (RNNs)[10] were applied to capture temporal dependences. Later, 3D-CNN[11] integrated both spatial and temporal modeling but still lacked efficiency in representing the structural properties of human motion.

To overcome these challenges, skeleton-based action recognition has gained attention. By representing the human body as a graph structure of joints and bones, a graph neural network (GNN)[12] can capture spatiotemporal relationships more effectively. In particular, the ST-GCN was proposed as the first GNN-based framework for skeleton-based action recognition. ST-GCN integrates spatial graph convolution with temporal convolution, making it well suited for capturing sign language movements in both time and space.

2.3 Natural language models

Natural language models are designed to generate semantically coherent and grammatically correct sentences. Earlier models, such as statistical n-gram[13] and RNN-based methods, were limited in handling long-range dependences. The introduction of the Transformer architecture[14] enabled parallel processing and global context modeling through the self-attention mechanism, which became the foundation for advanced models.

Among them, bidirectional encoder representations from transformers (BERT)[15] achieved state-of-the-art performance in language understanding tasks by applying bidirectional encoding. However, its large model size restricts real-time and resource-constrained applications. To address this, DistilBERT was developed through knowledge distillation as a lighter and faster version of BERT while retaining most of its accuracy. Owing to its efficiency and strong sentence generation capability, DistilBERT is adopted in this study for translating recognized sign tokens into fluent Chinese sentences.

3. Methodology

In the proposed system, skeleton-based action recognition is integrated with natural language generation to achieve TSL translation. The overall framework consists of three main stages: key point extraction, action recognition, and sentence generation (Fig. 1).

3.1 System framework

The workflow of the proposed system is shown in Fig. 1. Input sign language videos are processed using MediaPipe Holistic and Hands modules to extract body and hand landmarks. The extracted key points form temporally ordered sequences that represent continuous sign movements. These sequences are then input into the ST-GCN model for gesture recognition,

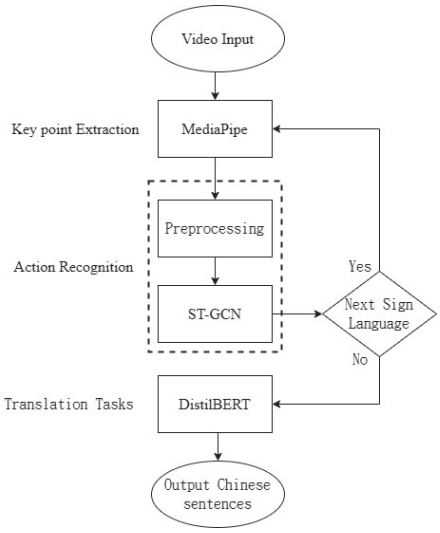


Fig. 1. System architecture of the proposed TSL translation system.

which outputs a sequence of sign tokens. Finally, a fine-tuned DistilBERT model translates the recognized tokens into fluent Chinese sentences.

3.2 Data collection and temporal alignment

Since no large-scale TSL dataset is publicly available, a custom dataset of 42 sign classes (Table 1) was recorded using a Logitech C310 webcam (Fig. 2). Each class contains 100 video samples, with a duration of 1–3 s per sign.

To standardize input lengths, each video was decomposed into frame sequences. If the number of frames exceeded the target length, sequences were cropped (Fig. 3); if fewer, frames were duplicated (padding) (Fig. 4) to reach the fixed length. This ensured temporal alignment and stable training. The selection of frames during cropping or duplication followed the sampling strategy defined in Eq. (1), which distributes frames uniformly across the video to preserve motion information.

$$F_i = \left\lfloor \frac{N}{T} \cdot i \right\rfloor \tag{1}$$

Here, N is the total number of frames in the original video, T is the target frame length, i is the index of the i-th sampled frame (i = 1, ..., T), and F_i represents the index of the corresponding frame selected from the original video. This strategy guarantees that the temporal structure of the action sequence is retained while adjusting videos to a fixed length. The overall temporal alignment strategy is summarized in Table 2.

3.3 Key point extraction with MediaPipe

In this study, we primarily employed the Holistic and Hands modules of MediaPipe. As shown in Fig. 5,[3] the Holistic module is capable of predicting face, hand, and body landmarks

Table 1
The custom TSL dataset.

English	Chinese	English	Chinese	English	Chinese
go	去	continue	持續	likewise	也是
excuse_me	不好意思	days	好多天	we	我們
bathroom	廁所	wait_a_minute	稍等一下	together	一起
where	哪裡	help	幫忙	can	可以
thank_you	謝謝	need	需要	can_not	不可以
you_are_welcome	不客氣	please	請	mood	情緒
hello	你好	ask	詢問	poorly	不好
name	名字	this	這是	moved	難過
unknow	不知道	feeling	心情	why	為什麼
happy	高興	today	今天	exam	考試
know	知道	play	玩	I	我
you	你	photography	攝影	mom	母親
weather	天氣	interest	興趣	scolding	責罵
rain	雨	Taiwan	臺灣	corn	玉米

Fig. 2. (Color online) Logitech C310 webcam.

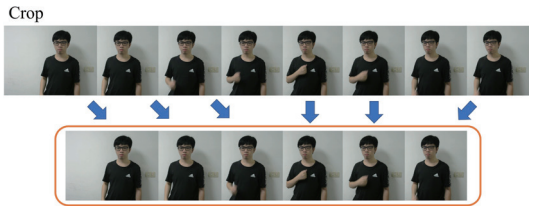


Fig. 3. (Color online) Example of frame cropping for temporal alignment.

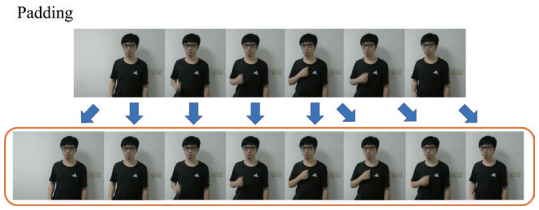


Fig. 4. (Color online) Example of frame padding for temporal alignment.

Table 2
Temporal alignment strategy for sign language videos.

Case	Method applied
Frames > target length	Cropping of redundant frames
Frames < target length	Frame duplication (padding)
Frames = target length	No modification required

simultaneously, making it suitable for full-body pose analysis. In Fig. 6,(3) the Hands module, in contrast, focuses specifically on the hand region and provides more stable tracking of finger articulations. Since the hand landmarks estimated by the Holistic module are relatively coarse, the Holistic and Hands modules are integrated to enhance the granularity of hand motion details.

To focus on the most relevant regions for sign language recognition, the integrated landmark set was refined to retain only the upper-body landmarks (0–14 from Holistic) and the 21 landmarks from each hand (Hands module). To achieve skeleton fusion, the root landmark

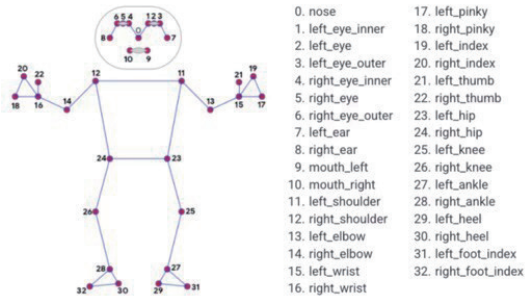


Fig. 5. (Color online) Landmark positions predicted by the Holistic module.[3]

Fig. 6. (Color online) Landmark positions predicted by the Hands module.[3]

(index 0) of each hand in the Hands module was aligned to overwrite the left-hand (index 15) and right-hand (index 16) wrist landmarks of the Holistic module. This ensured seamless connectivity between body and hand nodes.

Because both the Holistic and Hands modules index landmarks starting from 0, a re-indexing process was required to establish a unified numbering system. The final indexing scheme is as follows:

- 0–14: Upper-body landmarks (holistic)
- 15–35: Left-hand landmarks (hands)
- 36–56: Right-hand landmarks (hands)

The final integrated landmark configuration is summarized in Table 3, and the fused skeletal structure is illustrated in Fig. 7.

This integration strategy provides a balance between global posture awareness (upper-body landmarks) and fine-grained hand articulation (detailed hand landmarks). The fused and re-indexed landmark set forms the graph nodes for the ST-GCN model, with edges defined in accordance with anatomical connectivity across the body and hands.

3.4 Action recognition with ST-GCN

The ST-GCN was employed to recognize sign gestures. The model architecture is illustrated in Fig. 8.[4] By modeling both spatial relationships (between joints) and temporal dynamics (across frames), ST-GCN effectively captured the structural and dynamic characteristics of sign language movements.

ST-GCN models a skeleton sequence as a spatiotemporal graph. Each frame is represented as a spatial graph, where nodes correspond to human joints and edges represent the natural connections of the human body. In addition, temporal connections are constructed by linking the same joint across consecutive frames, thus forming a spatiotemporal graph structure.

Formally, a skeleton sequence can be defined as a graph $G = (V, E)$, where V is the set of nodes (joints), and E is the set of edges. The graph convolution operation at node v_{ti} is given by

$$f_{out}\left(v_{ti}\right) = \sum_{v_{tj} \in B\left(v_{ti}\right)} \frac{1}{Z_{ti}\left(v_{tj}\right)} f_{in}\left(v_{tj}\right) \cdot w\left(l_{ti}\left(v_{tj}\right)\right), \tag{2}$$

Table 3
Landmark configuration of the integrated Holistic + Hands modules.

Region	Source module	Index range	Landmarks	Description
Upper body	Holistic	0–14	15	Head, torso, and upper-limb joints
Left hand	Hands	15–35	21	Finger joints and palm key points
Right hand	Hands	36–56	21	Finger joints and palm key points
Total	—	0–56	57	Unified skeletal graph nodes

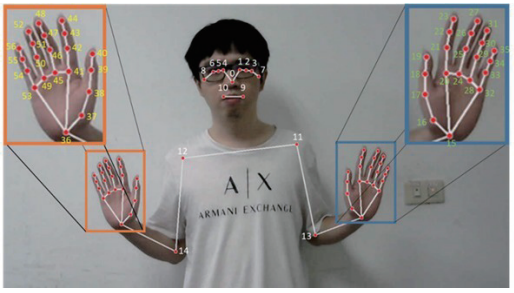


Fig. 7. (Color online) Integration of Holistic and Hands modules for skeleton construction.

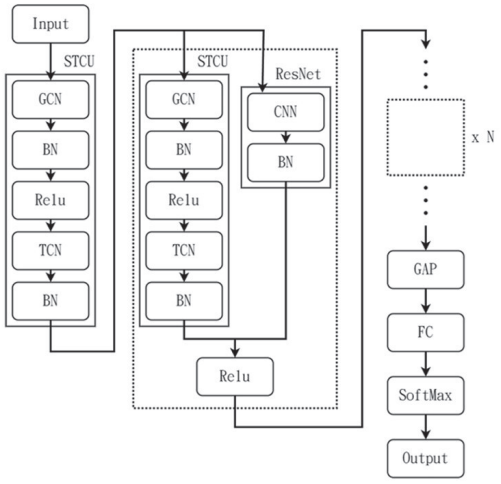


Fig. 8. Network structure of the ST-GCN model.[4]

where f_{in} and f_{out} are the input and output feature maps, t denotes the temporal index in the input skeleton sequence, $B(v_{ti})$ denotes the set of neighbors of node v_{ti}, Z_{ti} is a normalization factor, w is the weight function, and $l_{ti}(v_{tj})$ assigns neighboring nodes to a subset using partitioning strategy.

To simultaneously capture temporal dynamics, ST-GCN extends graph convolution over multiple frames. The operation can be expressed in matrix form as

$$f_{out} = \sum_k \Lambda_k^{-\frac{1}{2}} A_k \Lambda_k^{-\frac{1}{2}} f_{in} w_k \ , \tag{3}$$

where A_k is the adjacency matrix for the k-th subset, Λ_k is the degree matrix, and w_k is the learnable weight matrix. By stacking multiple spatiotemporal graph convolution layers, ST-GCN

is able to effectively learn both the spatial dependences of body joints and the temporal variations of motion sequences, which is crucial for recognizing sign language gestures. The training configuration, including loss function, optimizer, and batch size, is summarized in Table 4.

3.5 Sentence generation with DistilBERT

While ST-GCN outputs sequences of sign tokens, these do not directly correspond to fluent Chinese sentences owing to grammatical and structural differences. To address this, a fine-tuned DistilBERT model was adopted for sentence generation. Using 24 manually designed Chinese sentence templates, the model was trained to map token sequences into grammatically correct and semantically coherent sentences. The training parameters for DistilBERT are listed in Table 5.

4. Experiments

4.1 Experimental environment

The experiments were conducted on a desktop computer equipped with an Intel Core i7-13700 CPU, 32 GB of RAM, and an NVIDIA RTX 4090 GPU. The specifications of the software environment are listed in Table 6. The actual photography environment is shown in Fig. 9.

3.2 Experimental metrics

In this study, four evaluation metrics were used to assess the model performance, namely, *accuracy*, *F*1 *score*, *recall*, and *precision*. These metrics provide a detailed analysis of the

Table 4
Training parameters for the ST-GCN model.

Parameter	Value
Batch size	32
Activation function	ReLU
Dropout	0.1
Loss function	Cross-entropy
Optimizer	AdamW

Table 5
Training parameters for the DistilBERT model.

Parameter	Value
Batch size	8
Transformer encoder	6
Self-attention head	12
Activation function	GELU
Dropout	0.1
Loss function	Cross-entropy
Optimizer	AdamW

Table 6
Experimental environment.

Software	Version
Python	3. 9
OpenCV	4. 5
Mediapipe	0. 10. 0
PyTorch	1. 13
Torchvision	0. 14
Hugging face transformers	4. 26
Numpy	1. 21
Matplotlib	3. 9. 4

(a) (b)

Fig. 9. (Color online) The actual photography environment is shown. (a) Video camera and tripod. (b) Background environment in the shot.

model's ability to correctly predict labels in the dataset and are commonly used to evaluate classification models. Each metric is described as follows.

- *Accuracy*:

Accuracy is commonly used to measure the correctness of the model's predictions. It is defined as the ratio of correctly predicted samples to the total number of samples, usually expressed as a percentage. The formula is given as [16]

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

where true positive (*TP*) denotes the number of samples correctly predicted as positive, true negative (*TN*) denotes the number of samples correctly predicted as negative, false positive (*FP*) denotes the number of negative samples incorrectly predicted as positive, and false negative (*FN*) denotes the number of positive samples incorrectly predicted as negative.

- *Recall*:

Recall measures the ability of the model to correctly identify positive samples. A higher *recall* indicates that the model can detect more true positives while reducing false negatives. The formula is given as

$$Recall = \frac{TP}{TP + FN}. \tag{5}$$

- *Precision*:

Precision evaluates the proportion of correctly predicted positive samples among all predicted positives, i.e., the ratio of true positives in the predicted positive results. The formula is given as

$$Precision = \frac{TP}{TP + FP}. \tag{6}$$

- *F*1-*score*:

The *F*1 *score* is the harmonic mean of *precision* and *recall*, which provides a balanced evaluation of model performance. The *F*1 *score* ranges from 0 to 1, with higher values indicating better performance. The formula is given as

$$F1\,score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{7}$$

4.3 Recognition performance of ST-GCN

The integrated MediaPipe configuration imposes strict constraints, such as supporting only a single user and limiting the prediction range to the upper-body region. These restrictions make most publicly available action recognition datasets unsuitable for evaluation. Therefore, from the UCF-101 dataset, we selected 15 action classes[17] that meet these conditions. The dataset details are summarized in Table 7. For each class, 100 video samples were used, resulting in a total of 1500 videos, which were split into 1200 training samples and 300 validation samples at a ratio of 4:1.

After iterative training, the best performance was achieved with an *accuracy* of 90.66%, as illustrated in Fig. 10 (confusion matrix). The confusion matrix clearly shows the distribution of classification results across the selected classes. In addition, the training process is depicted in Fig. 11, which shows the *accuracy* and *loss* curves, confirming convergence of the ST-GCN model.

Table 7
Selected action classes from UCF-101 dataset for evaluation.

Dataset	UCF-101
Class	Archery, BodyWeightSquats, BoxingPunchingBag, CleanAndJerk, GolfSwing, JugglingBalls, JumpingJack, PlayingCello, PlayingDaf,PlayingDhol, PlayingFlute, PlayingGuitar, PlayingTabla, PlayingViolin, YoYo

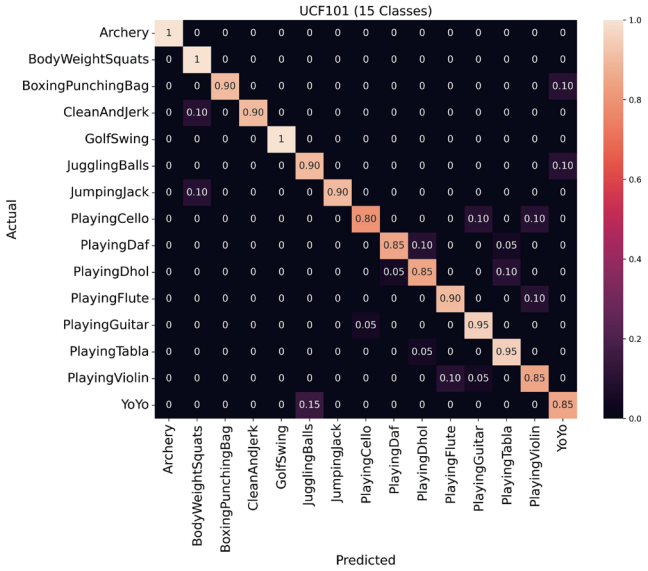
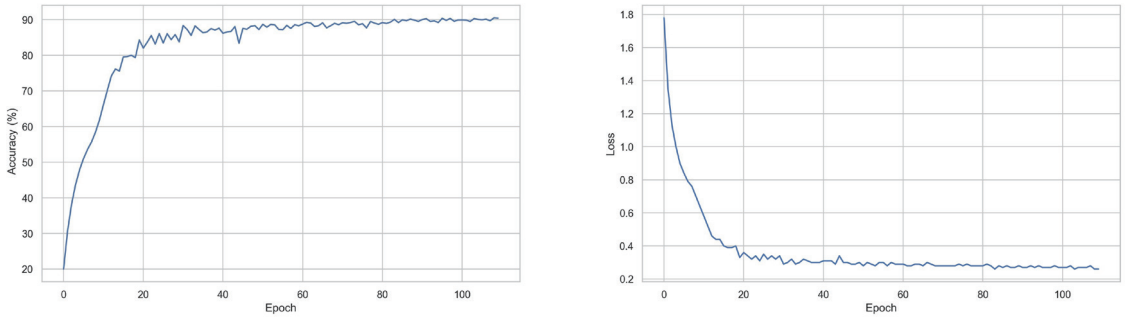


Fig. 10. (Color online) Confusion matrix of ST-GCN on UCF-101 evaluation set.



(a) (b)

Fig. 11. (Color online) *Accuracy* and *loss* curves of ST-GCN on UCF-101. (a) *Accuracy* curves. (b) *Loss* curves.

Although the achieved *accuracy* of 90.66% is not the highest compared with those in other studies using UCF-101, this result demonstrates the feasibility of combining MediaPipe with ST-GCN for action recognition. The lower *accuracy* can be attributed to inconsistencies within the dataset. Since UCF-101 is compiled from videos collected online, variations in recording style and camera viewpoint often exist within the same class. As shown in Fig. 12, the Archery category includes samples captured from multiple angles, making it more difficult for the model to focus on a single motion pattern.

The proposed ST-GCN model was trained and tested on the custom TSL dataset. The recognition results under different frame lengths are shown in Table 8. The model achieved its highest *accuracy* of 97.5% at 70-frame inputs, while shorter sequences (30–40 frames) yielded lower recognition rates owing to insufficient temporal information. Longer sequences (≥80 frames) slightly reduced *accuracy*, likely because of redundant or noisy frames introduced during alignment.

The confusion matrix of ST-GCN on the test set is illustrated in Fig. 13, showing that most classes achieved high recognition rates. However, some visually similar signs, such as “this” and “go”, exhibited partial misclassification owing to overlapping hand trajectories.

To further evaluate the training dynamics, Figs. 14 and 15 present the validation *accuracy* and *loss* curves across epochs. The curves indicate stable convergence, with the validation *accuracy* consistently improving while the *loss* gradually decreases.

Overall, the results confirm that the ST-GCN model effectively captures both spatial joint dependences and temporal motion patterns, yielding robust recognition performance across different sign categories. The high *F*1-*scores* (>95% for most configurations) also suggest that the model maintains a good balance between *precision* and *recall*, ensuring reliable recognition in practical applications.

4.4 Sentence generation performance of DistilBERT

The DistilBERT model was evaluated on its ability to generate fluent Chinese sentences from recognized sign tokens. Experimental results showed that the model achieved an *accuracy* of 100% on the evaluation dataset. This indicates that every input sequence of sign tokens was successfully translated into the correct Chinese sentence.

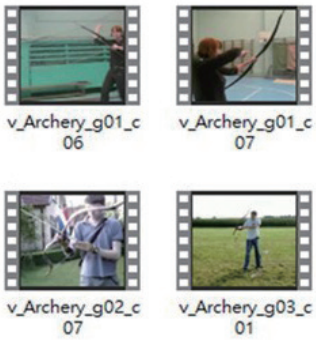


Fig. 12. (Color online) Variations in recording style and viewpoints within the Archery class in UCF-101.

Table 8
Recognition performance of ST-GCN for different frame lengths.

Frame length	*Accuracy* (%)	*Precision* (%)	*Recall* (%)	*F*1-*score* (%)	*Loss*
30 frames	82.3	80.5	81.0	80.7	0.36
40 frames	89.1	87.2	88.5	87.8	0.27
50 frames	95.0	94.6	94.9	94.7	0.21
60 frames	96.7	96.5	96.8	96.6	0.20
70 frames	97.5	97.3	97.4	97.3	0.18
80 frames	96.8	96.6	96.7	96.6	0.19
90 frames	95.5	95.2	95.3	95.2	0.22

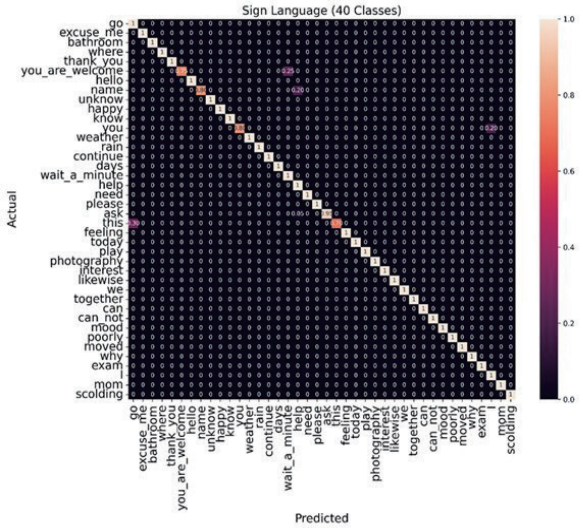
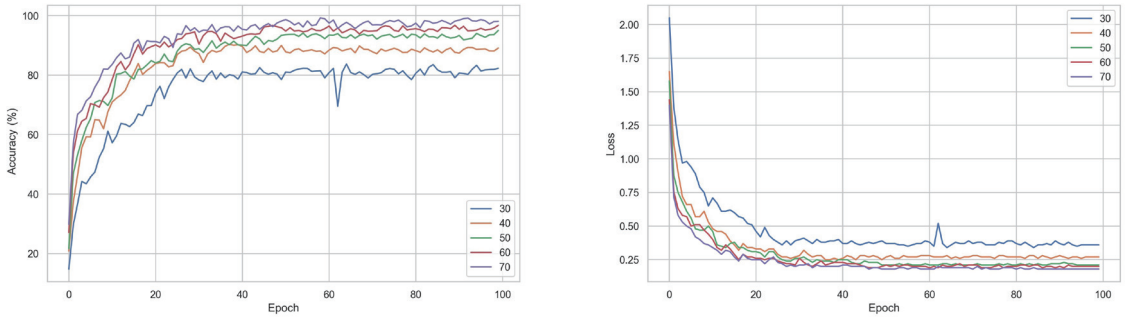


Fig. 13. (Color online) Confusion matrix of ST-GCN on 70-frame dataset.

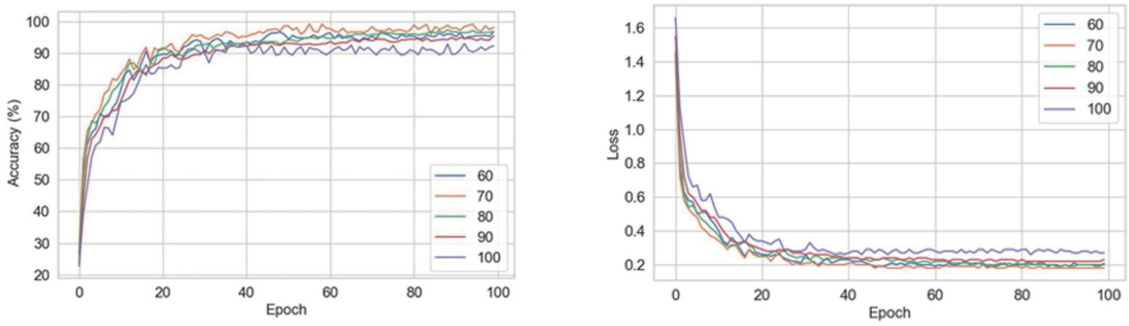
To train the model effectively, 24 Chinese characters in sentence samples were designed to represent typical conversational expressions in TSL. These sentence templates were used as ground-truth outputs in the sequence-to-sequence generation task.

In Fig. 16, the classification results of DistilBERT are presented. The figure demonstrates how the model assigns the correct sign token at each decoding step, ensuring that the generated sequence precisely matches the intended sentence. This confirms the effectiveness of the model



(a) (b)

Fig. 14. (Color online) *Accuracy* and *loss* curves of ST-GCN with input sequences of 30–70 frames. (a) *Accuracy* curve. (b) *Loss* curve.



(a) (b)

Fig. 15. (Color online) *Accuracy* and *loss* curves of ST-GCN with input sequences of 60–100 frames. (a) *Accuracy* curve. (b) *Loss* curve.

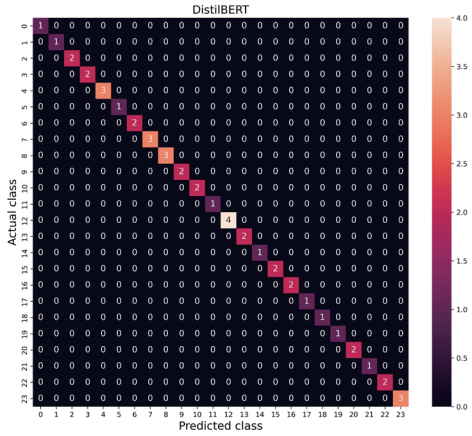


Fig. 16. (Color online) DistilBERT output matrix for sentence generation.

in producing coherent and accurate translations. The full list of sentence samples is provided in Table 9.

4.5 Graphical user interface implementation

To demonstrate practical applicability, a GUI was developed (Fig. 17). The interface allows users to input sign language videos, process them with the proposed system, and display both the recognized tokens and the generated Chinese sentences.

Table 9
Sentence samples used for training DistilBERT (24 templates).

Number	Chinese sentence (English sentence)	Number	Chinese sentence (English sentence)	Number	Chinese sentence (English sentence)
0	不好意思 (Excuse me)	8	你要去哪裡 (Where are you going)	16	我考試沒考好 (I didn't do well on the exam)
1	不客氣 (You're welcome)	9	天氣怎麼樣 (How's the weather)	17	為什麼 (Why)
2	今天玩得很開心 (I had a great time today)	10	廁所在哪裡 (Where is the restroom)	18	稍等一下 (Please wait a moment)
3	你叫什麼名字 (What's your name)	11	很高興認識你 (Nice to meet you)	19	請問 (May I ask)
4	你好 (Hello)	12	我也是 (Me too)	20	謝謝 (Thank you)
5	你今天心情如何 (How are you feeling today)	13	我們可不可以一起攝影 (Can we take photos together)	21	這是什麼 (What is this)
6	我被媽媽罵了 (My mom scolded me)	14	我心情不好 (I'm in a bad mood)	22	雨下了好幾天 (It's been raining for days)
7	你的興趣是什麼 (What are your interests)	15	我的興趣是攝影 (My hobby is photography)	23	請幫幫我 (Please help me)

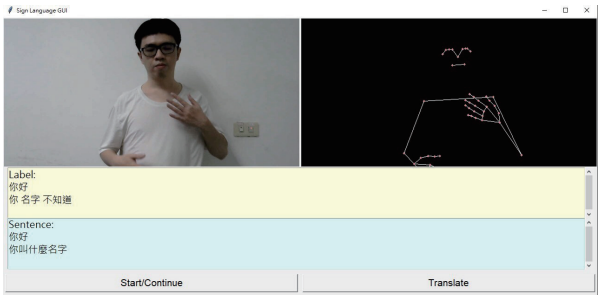


Fig. 17. (Color online) GUI interface.

The overall GUI workflow is shown in Fig. 18. It consists of the following four major steps.

I. Video Input: The user uploads prerecorded videos.
II. Key Point Extraction: The input video is processed by the MediaPipe Holistic and Hands modules to extract skeletal key points.
III. Action Recognition: The extracted key points are input into the ST-GCN model to recognize sign tokens.
IV. Sentence Generation and Display: The recognized tokens are passed to DistilBERT, which generates a fluent Chinese sentence. Both tokens and final sentences are displayed on the GUI.

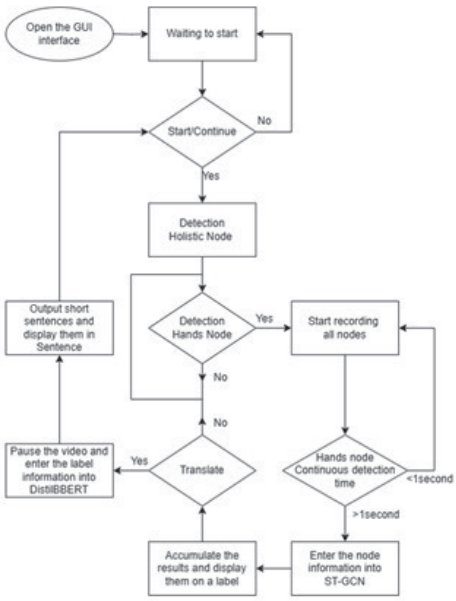


Fig. 18. (Color online) GUI workflow.

5. Conclusions

In this study, we proposed a TSL translation system that integrates skeleton-based action recognition with natural sentence generation. The system combines MediaPipe for multimodal key point extraction, ST-GCN for gesture recognition, and a fine-tuned DistilBERT for sentence generation. A custom TSL dataset consisting of 42 sign categories was recorded using a camera sensor, and temporal alignment techniques were applied to standardize video inputs.

Experimental results demonstrated that ST-GCN effectively recognized sign gestures with high *accuracy*, achieving a peak performance of 97.5% for 70-frame input sequences. Furthermore, the DistilBERT model successfully transformed recognized tokens into fluent Chinese sentences using 24 predefined training templates. The integration of recognition and generation resulted in natural and meaningful translations, confirming the feasibility of the proposed approach for TSL comprehension.

References

1 World Health Organization: https://www.who.int/health-topics/hearing-loss#tab=tab_2 (accessed August 2024).
2 Ministry of Health and Welfare: https://dep.mohw.gov.tw/DOS/cp-5224-62359-113.html (accessed August 2024).
3 Google MediaPipe Solutions Guide: https://ai.google.dev/edge/mediapipe/solutions/guide (accessed August 2024).
4 S. Yan, Y. Xiong, and D. Lin: Proc. AAAI Conf. Artificial Intelligence (AAAI, 2018) 7444–7452 https://doi.org/10.1609/aaai.v32i1.12328
5 V. Sanh, L. Debut, J. Chaumond, and T. Wolf: arXiv (2019) 1. https://doi.org/10.48550/arXiv.1910.01108
6 J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, and R. Moore: Proc. 2011 IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2011) 1297–1304. https://doi.org/10.1109/CVPR.2011.5995316

7 A. Toshev and C. Szegedy: Proc. 2014 IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2014) 1653–1660. https://doi.org/10.1109/CVPR.2014.214
8 Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh: IEEE Trans. Pattern Anal. Mach. Intell. **43** (2021) 172. https://doi.org/10.1109/TPAMI.2019.2929257
9 Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner: Proc. IEEE. **86** (IEEE, 1998) 2278. https://doi.org/10.1109/5.726791
10 J. L. Elman: Cognit. Sci. **14** (1990) 179. https://doi.org/10.1207/s15516709cog1402_1
11 S. Ji, W. Xu, M. Yang, and K. Yu: IEEE Trans. Pattern Anal. Mach. Intell. **35** (2013) 221. https://doi.org/10.1109/TPAMI.2012.59
12 F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini: IEEE Trans. Neural Networks **20** (2009) 61. https://doi.org/10.1109/TNN.2008.2005605
13 R. Kneser and H. Ney: 1995 Int. Conf. Acoustics, Speech, and Signal Processing **1** (IEEE, 1995) 181. https://doi.org/10.1109/ICASSP.1995.479394
14 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin: Proc. 31st Conf. Neural Information Processing Systems (NIPS, 2017) 1–15. https://doi.org/10.48550/arXiv.1706.03762
15 J. Devlin, M. W. Chang, K. Lee, and K. Toutanova: Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies **1** (2019) 4171. https://aclanthology.org/N19-1423/
16 D. M. W. Powers: arXiv 2 (2011) 37. https://doi.org/10.48550/arXiv.2010.16061
17 UCF101 - Action Recognition Data Set: https://www.crcv.ucf.edu/data/UCF101.php (accessed August 2024).