

Sign Language Recognition Methods: Applications and Advances of Deep Learning Technology

Junan Jiang *

School of GuangDong, GuangZhou HuaLi Collage, GuangZhou, China

* Corresponding Author Email: yulianavjvdv@gmail.com

Abstract. In the past few years, the field of sign language recognition (SLR) has attracted widespread attention. Deep learning models demonstrate their significant application potential with their outstanding performance. This paper reviews the application and the advancement of deep learning in the field of SLR, focusing on the application of models, such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory Networks. (LSTM), and Transformers in SLR. This paper systematically analyzes the key step of SLR, including data collection, preprocessing, model training, and performance evaluation. On this basis, this paper discusses the potential of MediaPipe's hand keypoint detection and the integration of CNN, 3D-CNN, and 2D-CNN fusion models, RNN-Transducer and H2SNet frameworks, as well as Transformer-based models, in improving the accuracy of SLR. The aim is to promote the development of accessible communication technologies and enhance the quality of life for those who are hearing-impaired community. This effort also seeks to provide a theoretical foundation and practical guidance for future technological advancements and societal inclusivity.

Keywords: Sign Language Recognition; CNN; RNN; Transformer Models.

1. Introduction

According to the World Health Organization (WHO), there are approximately 466 million people worldwide with hearing disabilities. Hearing loss affects approximately 466 million people worldwide, as per statistics from the World Health Organization (WHO). Hearing-impaired often face communication barriers in their daily lives. Sign language (SL), It is one of the main ways that people who are hard of hearing communicate. It is a visual language system that facilitates interaction through body language, facial emotions, and gestures [1]. Even though SL is a crucial communication tool for those who are hard of hearing, there still remains a significant barrier to communication when interacting with people who do not understand SL. This phenomenon not only affects the integration of hearing-impaired individuals in society, education, and employment but also has a detrimental impact on their quality of life. With the rapid advancement of computer technology and artificial intelligence in recent years, automatic SLR technology has provided a potential solution to this issue. SLR technology analyzes the hand movements and facial expressions of the target subject and other visual information, transforming it into text or speech. In this way, it can reduce the cost of communication between hearing-impaired individuals and hearing people., facilitating their interaction.

Traditional SLR algorithms primarily rely on manual feature extraction and gesture classification, such as edge detection, contour analysis, corner detection, and other methods, to extract features like color and shape of the gestures from videos and images. However, these methods exhibit poor robustness and generalization capabilities when faced with complex scene changes and noise interference. The application of models such as CNN, RNN, and LSTM has enabled deep learning algorithms to automatically extract features from images. This reduces the reliance on manually designed feature extraction and thereby enhances the accuracy and robustness of recognition [2].

This paper will focus on analyzing the research progress and application effects of deep learning in the field of SL recognition, and will discuss in detail the future research directions, aiming to provide researchers with research ideas and theoretical foundations for SLR technology.

2. The application of deep learning in SLR models.

This section will introduce methods and theories related to deep learning in SLR models, focusing on reviewing methods and related models that have applied deep learning models to SLR over the past 5 years. This includes the principles and applications of various deep learning models such as CNN, RNN, and Transformer in the process of SLR.

2.1. The process of SLR using deep learning.

The process of SLR includes multiple steps, ranging from data collection to model deployment. It is crucial to define the specific goals of SLR, such as recognizing the meaning of individual gestures, handling continuous SL expressions, and satisfying the demands for real-time performance [3].

(1) Data Collection and Labeling

During the steps of data collection, imaging equipment is typically used to capture SL images or video data. To ensure the accuracy and robustness of the SLR system, the dataset should be sufficiently diverse, ranging various SL expressions, hand gestures, and environmental backgrounds. SL data is categorized and annotated through methods such as manual labeling, ensuring the diversity and quality of the dataset.

(2) Data Preprocessing and Augmentation

In the process of SLR, since changes in the background are unavoidable, it is necessary to ensure that the model focuses on learning the key features in the data by preprocessing operations that filter out unclear, incomplete, or mutilated images to ensure the quality of the dataset [4]. To enhance the model's generalization ability, data augmentation is typically performed to increase robustness, reduce overfitting, and lower training costs.

(3) Model Training

After the completion of data preprocessing, the annotated data is fed into the model for training, during which the model may sometimes require optimization. The Adam optimizer is commonly used to accelerate the model's convergence process. The learning rate is typically set to 0.001. Throughout the training process, the model continuously adjusts its parameters to minimize errors and enhance its ability to recognize sign language images.

(4) Model Validation and Testing

After the model training is completed, its need to evaluate its performance. Common evaluation metrics include accuracy, recall, confusion matrix, and F1 score. These metrics can comprehensively measure the recognition effect of the model and ensure its good performance in different scenarios [5]. Once the model achieves satisfactory performance when evaluated on the test set, it can deploy in practical applications, such as developing a program that captures SL images in real-time through a camera, and after system recognition, converts them into text or voice output.

2.2. CNNs, RNNs, and Hybrid CNN-based Hand Feature Recognition

In recent years, CNNs have been widely applied in research of SLR. By emulating the processing mechanisms of the human brain to automatically extract features from images, it reduces the specialized knowledge required for image recognition across different field. In SLR, the shape of the hands and the direction of gestures are different from person to person, existing significant nonlinear issues, which CNNs address by learning the nonlinear relationships between images [6].

2.2.1 Enhancing SLR with CNN and MediaPipe

In the process of SLR, the accuracy of hand tracking may not always reach a high level, leading to insufficient accuracy in SLR. This is extremely important in the practical application of SLR because the expression of SL heavily relies on precise hand movements and positions. even minor errors can lead to semantic inconsistencies. MediaPipe offers a range of pre-trained models and tools that enable developers to quickly build multifunctional applications capable of detecting 21 key points of the hand in real-time dynamic SLR, including the positions of the palm and fingers, for real-time capture and preprocessing of hand movements [7]. In this process, CNNs can be utilized to achieve

dynamic SLR. The research team first performed feature extraction and preprocessing on the ASL (American Sign Language) dataset, converting the keypoint data of the hand into image forms or tensor formats suitable for CNN processing, and then normalizing the extracted features. Min-max Normalization was applied to scale the pixel values to a range of [0,1], optimizing the training speed of the CNN model. The CNN Three convolutional layers, two pooling layers, and a fully connected layer made up the design of the model, using the ReLu function to accelerate training and mitigate the problem of vanishing gradients. The max pooling layer was used to reduce the feature dimension and extract key features, while the fully connected layer mapped the final classification results, and early stopping was employed to prevent model overfitting. In the performance evaluation phase, metrics such as accuracy, recall, precision, and F-measure were used to assess the model's performance. Ultimately, an accuracy rate of 99.12% was achieved on the ASL dataset, demonstrating the accuracy and reliability of the recognition task. In contrast, in another study [8], MediaPipe was used to extract key points of the hand, body, and face, and then combined with Recurrent Neural Network (RNN) models (GRU, LSTM, and Bi-directional LSTM) to address the frame dependency issues in SL movement, with an accuracy rate also exceeding 99% [8].

2.2.2 Deep Learning Model Combining 3D-CNN and 2D-CNN for Dynamic SLR

3D-CNNs (3D-CNNs) are specifically designed to handle data that has both temporal and spatial dimensions. These CNNs utilize 3D convolutional kernels to extract features, making them highly suitable for processing video data. Muneer Al-Hammadi, and colleagues proposed a model that combines 3D-CNN and 2D-CNN to address the issue of insufficient labeled data through transfer learning [9]. In feature extraction, a single 3D-CNN model directly extracts features from video samples, dividing the samples into three parts (beginning, middle, end) and employing three 3D-CNN models to extract features. The study also explored three different fusion techniques—Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), and autoencoders—to process features extracted in parallel from the 3D-CNN structures. In both signer-dependent and signer-independent modes, the MLP technique achieved the highest accuracy rates, as confirmed by the results, while LSTM and autoencoders also showed some effectiveness but were slightly inferior to MLP. Additionally, the researchers used PCA (Principal Component Analysis) and t-SNE (t-Distributed Stochastic Neighbor Embedding) techniques for feature dimension reduction and visualization to better understand the effects of these feature fusion techniques. The results demonstrated that features learned by the parallel 3D-CNN structures were more dispersed in space, indicating better discriminability.

2.2.3 Sign Language Recognition using RNN-Transducer and H2Snet

Recurrent Neural Networks (RNNs) are adept at handling time-series data, thus playing a crucial role in dynamic SLR. Before feeding video data into the model, preprocessing is required, including the extraction of video frames, scaling, normalization, and combining continuous video frames into time-series data. Subsequently, CNNs can be utilized as feature extractors to identify key features, which are then sent into RNNs, such as Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU), to model the temporal dependencies of hand gestures. Finally, model fusion is conducted by inputting features extracted by CNN into RNNs to recognize dynamic gestures and optimizing the model to improve the accuracy and real-time performance of recognition.

Traditional RNN-based SLR methods often formulate the problem between SL video frame sequences and sign language label sequences as a sequence alignment problem. The effective alignment between video sequences and sentence-level labels can be achieved through the use of Connectionist Temporal Classification (CTC). CTC methods are unsuccessful when the output label sequence is longer than the input video sequence and do not take into account the interdependencies between output predictions, leading to reduced recognition accuracy and incoherent semantics. Therefore, Liqing Gao and colleagues proposed an RNN-Transducer-based SLR framework, also known as A visual hierarchy is present in the lexical sequence alignment network (H2SNet) [10]. A visual hierarchy is present in the lexical sequence alignment network (H2SNet) [10]. This framework

consists of a network that transcribes sign language videos using visual hierarchy, capturing the spatial appearance and temporal motion cues. This framework includes a visual hierarchy transcription network that captures the spatial appearance and temporal motion cues of sign language videos, providing rich contextual information for SLR. The RNN-Transducer is responsible for learning the relationship between sequential video features and sentence-level labels. Research results indicate that H2SNet outperforms existing state-of-the-art methods on the Chinese Sign Language (CSL) dataset and is faster in evaluation speed compared to other methods, demonstrating its potential application value in real-time SLR systems.

2.3. Enhancing SLR Accuracy with Transformer Architecture

The Transformer architecture, initially designed for Natural Language Processing (NLP) tasks, revolves around the self-attention mechanism. This mechanism allows the model to consider all elements in a sequence simultaneously when processing sequential data, effectively capturing long-range dependencies. The Transformer architecture typically consists of an encoder and a decoder, each with a similar hierarchical structure. Each layer can process different information in parallel, enabling the model to consider both local features and context of gestures, capturing the dynamics and continuity in sign language videos, which is crucial for SLR.

2.3.1 Transformer-based network model for word-level sign language recognition

Early sign language recognition primarily utilized Hidden Markov Models (HMMs) to process SL videos, while in recent years, the combination of CNNs and Recurrent Neural Networks (RNNs) has provided new possibilities for SLR. However, these methods still have limitations in dealing with complex backgrounds, gesture details, and temporal information. To address the limitations of the aforementioned methods, Yao Du et al. proposed a novel Transformer-based network model for word-level SLR [11]. The model consists of two parts: a spatial encoder and a temporal module. The spatial encoder, with Swin Transformer as the backbone network, is responsible for extracting high-level semantic features from video frames, aiding in capturing global spatial information. The temporal module, composed of multiple future-masked Transformer layers, focuses on understanding the sequential nature of SL movements, which is also the greatest innovation of this study. By masking future frames in the self-attention mechanism, the model can only make judgments based on current and past gesture information, thereby better understanding the immediacy and continuity of sign language. Experimental results show that the model performed excellently on the WLASL and NMFs-CSL datasets, the WLASL dataset is setting a new record for word-level sign language recognition, indicating its significant application the field of sign language recognition has potential.

2.3.2 End-to-End Continuous Sign Language Recognition (CSLR) via Spatial-Temporal Transformer Network (STTN)

Continuous Sign Language Recognition (CSLR) has always been a technical challenge because it requires coordinating low-density video sequences with high-density text sequences. Existing CNN models struggle to balance Visual feature extraction involves the use of spatial and temporal features, which makes it difficult to improve recognition accuracy. Unlike the previously mentioned Transformer-based new network models, Zhenchao Cui and colleagues proposed an end-to-end CSLR network called the Spatial–Temporal Transformer Network (STTN), which is designed based on the architecture of the Transformer model [12]. In STTN, spatial encoding utilizes CNNs and the self-attention and multi-head attention mechanisms from Transformers to effectively extract spatial features from video frames. Temporal encoding is achieved through RNNs and the temporal attention mechanism of Transformers to understand the temporal dependencies in video frame sequences. STTN reduces computational complexity while maintaining recognition performance and supports end-to-end training and recognition, meaning the model can simultaneously address both recognition and translation tasks, thereby increasing efficiency and accuracy. Research results indicate that compared to several existing state-of-the-art methods, STTN demonstrates superior performance in

continuous sign language recognition tasks, further proving its application potential in the field of sign language recognition.

3. Analysis and Discussion

Current sign language recognition systems are limited by the diversity and quality of datasets, lacking large-scale datasets that encompass various sign language expressions, hand gesture variations, and environmental backgrounds. Future research should focus on building databases that include diverse signs, dynamic symbols, and continuous sentences to enhance the generalization and practicality of algorithms.

The existing deep learning models for real-time sign language translation still need to be improved in terms of accuracy and robustness. Future researchers should aim to enhance current deep learning models to more accurately translate sign language for application in real-time translation scenarios. In the practical application process, users play a crucial role, as they can provide valuable feedback to help improve the system and offer guidance on data privacy and ethical issues. User studies are essential to ensure that sign language detection systems are widely accepted and reasonably deployed within an ethical framework.

From a technical perspective, real-time recognition of dynamic sign language sentences in complex environments is a challenge. Future research should explore more complex CNNs structures to better understand the contextual information of sign language. Additionally, researchers might consider deploying sign language translation under contextual constraints, which would facilitate real-time communication between deaf individuals and others. Although real-time recognition of dynamic sign language sentences in complex environments is challenging, it is also an important direction for future research, and it is hoped that future researchers can develop algorithms capable of handling videos of varying lengths.

Considering the diversity of sign language cultures, the development of sign language recognition systems should enable an understanding and respect for cultural differences. This will not only improve the accuracy of the system but also promote cultural exchange among different backgrounds.

Finally, the introduction of multimodal methods greatly enhances the possibilities of sign language recognition systems. By combining visual, auditory, and even tactile information, future sign language recognition systems can more accurately understand and interpret sign language, thereby providing users with a richer communication experience.

The future of SLR systems is promising. We can look forward to achieving a more inclusive and accessible society where individuals who use SL can communicate more freely with the world.

4. Conclusion

The analysis of various deep learning models, including CNN, RNN, LSTM, and Transformer, is presented in this article to introduce the application and progress of deep learning technology in SLR. By automatically extracting features, these models significantly enhance the accuracy and efficiency of sign language recognition and processing time-series data. Research indicates that combining MediaPipe's hand keypoint detection with CNN feature recognition, integrating 3D-CNN with 2D-CNN models, and frameworks based on RNN-Transducer and H2SNet have all effectively enhanced SLR performance. Furthermore, Transformer-based models have shown great potential in handling word-level and continuous SLR.

Through these technologies, the improvement in accuracy and efficiency of SLR has led to a more inclusive and accessible communication environment for the deaf and mute. Although current technologies still face challenges, such as diverse gestures and performance in different environments, there is a wide potential for the use of deep learning in sign language recognition. Future research will continue to optimize these models, looking forward to creating a more friendly and effective communication environment for the deaf and mute community.

References

- [1] Azevedo R V, Coutinho T M, Ferreira J P, Gomes T L, Nascimento E R. Empowering sign language communication: Integrating sentiment and semantics for facial expression synthesis. *Computers & Graphics*, 2024, 124: 104065 von Agris U., Zieren J., Canzler U., Bauer B., Kraiss K.-F. Recent developments in visual sign language recognition. *Universal Access in the Information Society*, 2007, 6(1): 123-141.
- [2] Jiang, X., Sun, J., Zhang, Y., Wang, L., & Han, X. (2024). Review of Chinese Sign Language Recognition Methods and Technologies. *Theoretical and Research*, 2024(3), 47-58.
- [3] Zuo R, Wei F, Mak B. Natural Language-Assisted Sign Language Recognition. *arXiv:2303.12080v1 [cs.CV]*, 2023 Mar 21.
- [4] von Agris, U., Zieren, J., Canzler, U., Bauer, B., & Kraiss, K.-F. (2007). Recent developments in visual sign language recognition. *Transactions of China Electrotechnical Society*, 19(10), 53-58.
- [5] Islam M.Z., Hossain M.S., Islam R.U., Andersson K. Static Hand Gesture Recognition using Convolutional Neural Network with Data Augmentation. *Transactions of China Electrotechnical Society*, 2004, 19(10): 53-58.
- [6] Rainio, O., Teuho, J. & Klén, R. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 2024, 14: 6086.
- [7] Gagandeep Singh, Aditya Raj Verma, Banavath Ramji, Karnim Meghwal, Praveen Kumar Dadheech. Enhancing Sign Language Detection through Mediapipe and Convolutional Neural Networks (CNN). *Transactions of China Electrotechnical Society*, 2024, 19(10): 53-58.
- [8] Cheng Q.Y., Sun C.X., Zhang X.X., et al. Short-Term load forecasting model and method for power system based on complementation of neural network and fuzzy logic. *Transactions of China Electrotechnical Society*, 2004, 19(10): 53-58.
- [9] Al-Hammadi M, Muhammad G, Abdul W, Alsulaiman M, Bencherif M A, Mekhtiche M A. Hand Gesture Recognition for Sign Language Using 3DCNN. *IEEE Access*, 2020, 8: 79491-79509.
- [10] Gao Liqing, Li Haibo, Liu Zhijian, Liu Zekang, Wan Liang, Feng Wei. RNN-Transducer based Chinese Sign Language Recognition. *Neurocomputing*, 2021, 434: 45-54.
- [11] Du Yao, Xie Pan, Wang Mingye, Hu Xiaohui, Zhao Zheng, Liu Jiaqi. Full transformer network with masking future for word-level sign language recognition. *Neurocomputing*, 2022, 500: 115-123.
- [12] Cui Z, Zhang W, Li Z, Wang Z. Spatial-temporal transformer for end-to-end sign language recognition. *Complex Intell Syst*, 2023, 9: 4645-4656.