Sign language recognition method based on deep learning

Mu He

School of Mathematics and Statistics, Xuzhou University of Technology, 221000, Xuzhou, China

Abstract. Sign language recognition, as an interdisciplinary field involving computer vision, pattern recognition, and natural language processing, holds profound research significance and extensive application value. This technology not only helps people with hearing impairments and those with normal hearing achieve barrier-free communication, but it also enhances their daily living experience while driving the development of sciences such as computer vision and artificial intelligence technologies. The subsequent text offers a thorough examination of the technologies involved in sign language recognition. It starts by detailing the methods for gathering data in sign language recognition, giving particular attention to hand modeling and the techniques used for visual feature extraction. Then, it discusses in detail the two methods of sign language recognition, namely traditional methods and artificial intelligence methods. These two methods have their advantages and disadvantages, providing different ideas for developing sign language recognition technology. Finally, the article proposes a prospect for the future development of sign language recognition technology, hoping that it can play a significant role in more fields and create a more convenient and barrier-free communication environment for people with hearing impairments.

1 Introduction

The roots of sign language recognition technology date back to the 1960s in the United States, where advocacy organizations for the hearing impaired began the quest to investigate technology for sign language communication. In 1972, the American Association of the Deaf completed the translation of the first sign language vocabulary list, which not only marked the initial formation of sign language recognition technology but also heralded its future robust development. In China, sign language recognition technology has gradually developed significantly over a long historical process, moving from complex to simple hand gestures, from a limited to a rich vocabulary, and from regional differentiation to national standardization. With the advancement of technology, the research and application of sign language recognition technology are increasingly expanding, aiming to promote barrier-free communication between deaf people and those with normal hearing.

The deaf-mute community uses sign language as their primary means of communication, making the development of sign language recognition technology of great significance to them. It can help deaf-mute individuals communicate better with others, enhancing their social engagement and quality of life. Through computer vision and artificial intelligence

algorithms, sign language recognition technology can convert sign language gestures into text, which can help the deaf and mute communicate better with others. With the advancement of technologies such as deep learning, the accuracy, and practicality of sign language recognition are continuously improving, providing deaf-mute individuals with more enriched communication tools and social support.

This article first introduces the two common methods of collecting sign language data sets: hand-based modeling data collection and vision-based technology collection. It then discusses sign language recognition methods from the perspectives of traditional methods and artificial intelligence methods. Conventional methods rely mainly on data gloves and visual technology, while artificial intelligence, particularly deep learning techniques, are employed to automatically extract features and facilitate recognition processes. The area of visual sign language recognition has drawn considerable interest within the academic and research communities, covering everything from data collection to recognition techniques and future outlooks. The goal is to highlight the cutting-edge progress and expansive potential in this domain. This paper hope that this paper will act as a valuable asset, providing insight to scholars and experts in the domain, and thereby fostering the ongoing growth and advancement of sign language recognition technologies.

2 Sign language dataset collection method

2.1 Collection method based on hand modeling

Hand modeling recognition is an important research direction in the fields of computer vision and human-computer interaction. It involves capturing and reconstructing the three-dimensional shape and movements of the hand from image or video data. In recent times, the advancement of deep learning has led to considerable achievements in the realm of hand modeling recognition, and a variety of efficient open-source tools and methods have emerged.

Metric-Affine Hand Model (MANO) is a real-time 3D hand model library implemented based on PyTorch[1]. It provides a highly realistic, low-dimensional, and adaptable hand model that fits any human hand shape. The MANO model is capable of mapping hand pose parameters and shape parameters to a 3D hand mesh, supporting end-to-end differentiation, which facilitates its integration into deep learning frameworks for training and optimization. The MANO model is suitable for applications in virtual reality, gesture recognition, human-computer interaction, as well as 3D modeling and animation.

HandRefiner is an open-source project based on deep learning, specifically designed for hand tracking and 3D reconstruction [2]. It uses convolutional neural networks for image processing and, through pre-trained models and optimization algorithms, can capture and reconstruct the 3D structure of the hand in real time with accuracy. HandRefiner supports real-time performance optimization and is applicable in fields such as virtual reality, game development, motion capture, and human-computer interaction.

One-shot Hand Avatar (OHTA) [3] is a new method that can create high-fidelity hand avatars from just a single image. OHTA addresses the inherent difficulties in building virtual avatars under limited data conditions by learning and leveraging data-driven priors. It supports text-to-hand avatar model generation, hand avatar appearance, geometric editing, and latent space editing operations.

The development of these technologies and tools has greatly advanced the field of hand modeling recognition, providing strong support for various applications. As research continues to deepen, future hand modeling recognition technology will become even more precise and practical.

2.2 Collection method based on visual technology

Visual sign language recognition technology leverages computer vision and artificial intelligence algorithms to interpret sign language gestures. This innovative technology serves as a bridge for the deaf and mute community to communicate with individuals who can hear, thereby enriching their social interactions and overall well-being. As deep learning technology progresses, significant interest has been focused on the domain of visual sign language recognition within the research circles. Researchers are tirelessly working to enhance the precision and instantaneous response capabilities of this recognition technology.

Sensory cameras, such as Kinect, are capable of capturing movements and gestures in three-dimensional space. These devices construct a three-dimensional model of the user by emitting infrared light and receiving the reflected signals, thereby enabling the tracking of gestures and body movements. In the field of sign language recognition, sensory cameras can be used to capture the gestures and body language of sign language users, providing input data for recognition algorithms. For instance, Kinect has been utilized in sign language recognition projects to identify sign language gestures through its motion capture functionality and transform them into written or spoken data [4].

The latest research breakthrough has led to the development of a sign language recognition system embedded in smart glasses. This setup employs deep learning techniques to examine and translate live sign language movements, with the goal of bridging the communication gap between the broader population and individuals with hearing and speech impairments using these innovative eyewear devices. For example, Liu has presented a portable sign language detection system that depends on a convolutional neural network. Through the combination of flexible strain sensors and inertial measurement units, the system can detect hand motions and paths, reaching a high level of precision in identifying sign language words and phrases [5].

These research achievements indicate that sign language recognition methods based on visual technology are developing at an unprecedented pace, demonstrating remarkable prospects in terms of improving recognition efficiency, expanding application scenarios, and optimizing user experience. In the future, with the continuous advancement of technology and the deepening of applications, visual sign language recognition technology will undoubtedly bring a more convenient, efficient, and natural communication experience to the deaf and mute community.

3 Sign language recognition method

3.1 Traditional Methods

Traditional sign language recognition methods primarily encompass two major branches: recognition technology based on data gloves and recognition technology based on vision. These two technologies each have their unique strengths and limitations, jointly forming the foundational framework of the sign language recognition field. Data glove-based recognition technology relies on a multi-sensor fusion system, which can accurately capture hand angles, motion trajectories, and timing information. Data gloves typically integrate a variety of high-precision components such as angular sensors, accelerometers, and magnetic sensors, which work together to record every subtle change in hand movement in real-time. Subsequently, this data is fed into advanced classifiers, such as Hidden Markov Models (HMMs) [6], to achieve precise recognition of specific sign language gestures. However, despite the excellent recognition accuracy of this method, its high cost, cumbersome wearing experience, and low portability limit its widespread application in daily life. In contrast, vision-based sign language recognition technology is more in line with societal needs and practical application

scenarios. This technology captures two-dimensional images or videos of sign language through cameras and uses image processing techniques and machine learning algorithms for recognition. The recognition process may include key steps such as background subtraction, contour detection, feature point tracking, and action recognition. Features like the shape, texture, and motion data of images are extracted and utilized to train machine learning algorithms, including Support Vector Machines (SVM) and Neural Networks (NN) [7], for the purpose of automating sign language recognition. Nonetheless, this approach encounters numerous obstacles in its practical use, such as boosting recognition accuracy, upgrading real-time capabilities, and ensuring robustness in intricate settings. Moreover, variables such as fluctuating lighting conditions, obstructions, and intricate scene backgrounds can negatively impact the efficiency of recognition.

The traditional methods of sign language recognition are challenged with the tasks of increasing recognition accuracy, boosting real-time operation, and cutting down on expenses. Data glove technology is limited by the portability and comfort of the equipment, while vision-based technology needs to overcome issues such as changes in lighting, occlusions, and complex backgrounds. Additionally, the diversity of sign languages and individual differences in expression present challenges for recognition algorithms. Researchers need to continuously explore new algorithms and technical approaches to improve the accuracy and stability of recognition while reducing application costs, and rendering sign language recognition technology more attainable for the routine requirements of individuals.

3.2 Artificial Intelligence Methods

Within the domain of sign language detection and interpretation, neural network methods, especially deep learning technologies, play a significant role. Among them, Convolutional Neural Networks (CNNs) are crucial, particularly in dealing with static gesture features in sign language video frames. By cleverly stacking convolutional and pooling layers, CNNs can efficiently learn features of different shapes and sizes, thus accurately capturing high-level semantic information in sign language. This capability allows CNNs to demonstrate high recognition accuracy and robustness when processing complex and variable sign language images. Long Short-Term Memory Networks (LSTMs), as a type of special Recurrent Neural Network (RNN), are adept at handling and learning long-term dependencies in time series data. By introducing LSTM methods into sign language recognition, the system can more accurately understand the continuity and fluidity of gestures, thereby further enhancing the accuracy and real-time performance of recognition.

3.2.1 Convolutional Neural Networks

CNNs is a kind of deep learning architecture that is adept at handling data with a grid format, such as visual images and video frames. It obtains advanced feature representations of the data through the stacking of convolutional layers, activation functions, pooling layers, and fully connected layers.

Convolutional layers utilize kernels that slide over the data to detect local patterns while pooling layers are used to reduce the size of the feature maps. This reduction not only decreases computational complexity but also improves the model's capability to identify patterns invariant to their location in the spatial domain. In sign language recognition, CNN can effectively extract static and dynamic visual features from sign language video frames. These features are crucial for distinguishing different sign language gestures. CNN can automatically learn key shapes, movements, and spatial relationships in sign language, thereby improving the accuracy of recognition. By combining CNN with attention mechanisms, the recognition accuracy of sign language words can be further enhanced [8].

For example, some research combines CNN with a multi-head attention mechanism to construct a lightweight sign language recognition network model. This model achieves effective feature extraction from input sign language video frame images through an improved convolutional neural network and attention mechanism [9].

3.2.2 Recurrent Neural Networks

RNNs are neural networks crafted to manage sequential information, like temporal data or textual data in natural language. The distinctive trait of RNNs lies in their network structure, which includes recurrent connections. These connections enable the network to preserve past information while handling sequential data, allowing RNNs to identify and represent long-term dependencies within time series data.

In the realm of sign language recognition, RNNs are employed to handle the dynamic aspects of information within sign language videos, such as the trajectory and temporal evolution of gestures. With recurrent connections, RNNs can utilize information from previous frames while processing the current sign language frame, thus more accurately recognizing sign language actions. Combining RNNs with attention mechanisms can further enhance the recognition accuracy of sign language words [8]. Attention mechanisms can help RNNs focus on the most relevant parts of the sign language video, ignoring less important background information, thereby improving the model's recognition performance. Additionally, RNNs can be used in conjunction with other network structures such as CNNs, leveraging the latter's advantages in image feature extraction and using RNNs to process time series data, achieving effective sign language recognition.

4 Current limitations and future prospects

Future sign language recognition systems will rely more deeply on diverse data modalities, including visual, auditory, and tactile dimensions, aiming to significantly improve recognition accuracy and system robustness. Deep learning models, with their powerful integration capabilities, can effectively fuse these different modalities of data, capturing more detailed and rich information. As technology continues to evolve, sign language recognition systems will enter a new era of personalization and adaptive learning, customizing learning based on each user's unique gestural habits to provide more accurate and error-free recognition services. Achieving this goal poses higher demands on the model's generalization and adaptive learning capabilities [10].

Future systems will also strive to achieve instant recognition and rapid feedback of sign language, which is of vital importance for applications in education, entertainment, and social interaction, bringing an unprecedented upgrade in user experience. To further enhance recognition accuracy, the construction of a large-scale, rich, and comprehensive sign language dataset is crucial. At the same time, the clever application of transfer learning techniques will ensure that the model maintains its efficiency and accuracy when dealing with different sign language environments and users.

As technology becomes more extensively deployed and utilized, concerns over user data security and privacy have grown more visible, necessitating this issue to be a pressing matter that requires immediate attention. Therefore, future sign language recognition systems must incorporate ethical and privacy considerations from the outset of their design to ensure the healthy development of technology and user trust.

5 Conclusions

This article delves into the technology of sign language recognition, covering data acquisition, recognition methods, and future development trends. It begins by introducing two methods of data acquisition: based on hand modeling and based on visual technology. Hand modeling techniques such as MANO, HandRefiner, and OHTA can capture and reconstruct the three-dimensional structure of the hand, while visual technology uses devices such as depth-sensing cameras and smart glasses to capture sign language gestures. Traditional sign language recognition methods include those based on data gloves and those based on vision. Data gloves can accurately capture hand movements, but they are expensive and have poor portability. Vision-based technology captures images with a camera but is susceptible to environmental factors. Artificial intelligence methods, especially deep learning, play a crucial role in sign language recognition. CNNs excel at extracting static gesture features, while RNNs are adept at processing dynamic information. Combining attention mechanisms can further improve recognition accuracy. Future sign language recognition technology will develop towards the use of diversified data modalities, personalization and adaptive learning, instant recognition, and rapid feedback. Deep learning models will integrate different modalities of data to improve recognition accuracy. Transfer learning technology will ensure that models remain efficient in different sign language environments and for different users. At the same time, system design must take ethical and privacy protection issues into full consideration to ensure the healthy development of technology. In summary, sign language recognition technology is in a stage of rapid development, providing more opportunities for communication for the deaf and driving technological progress in related fields.

References

1. J. Romero, D. Tzionas, M.-J. Black, Embodied hands: Modeling and capturing hands and bodies together, arXiv preprint arXiv:2201.02610 (2022)
2. Z. Zhang, S. Xie, M. Chen, H. Zhu, HandAugment: A simple data augmentation method for depth-based 3D hand pose estimation, arXiv preprint arXiv:2001.00702 (2020)
3. X. Zheng, C. Wen, Z. Su, Z. Xu, Z. Li, Y. Zhao, Z. Xue, OHTA: One-shot Hand Avatar via Data-driven Implicit Priors, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2024), 799-810
4. Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, P. Presti, American sign language recognition with the Kinect, in Proceedings of the 13th International Conference on Multimodal Interfaces, (2011), 279-286
5. Y. Liu, X. Jiang, X. Yu, H. Ye, C. Ma, W. Wang, Y. Hu, A wearable system for sign language recognition enabled by a convolutional neural network, Nano Energy, **116**, 108767 (2023)
6. H. Liu, Construction and Implementation of Hidden Markov Model, J. Shanghai Univ. Electric Power. **37**(5), 467-470 (2021)
7. K. Zhang, J. Hou, An Improved Convolutional Neural Network Image Recognition Method, Sci. Technol. Eng. **20**(01), 252-257 (2020)
8. C. Lu, M. Kozakai, L. Jing, Sign Language Recognition with Multimodal Sensors and Deep Learning Methods, Electronics. **12**(23), 4827 (2023)
9. S.-J. Zhang, Q. Zhang, H. Li, A Survey of Sign Language Recognition Based on Deep Learning, J. Electronics Inf. Technol. **42**(04), 1021-1032 (2020)

10. Wadhawan, P. Kumar, Deep learning-based sign language recognition system for static signs, Neural Comput. Appl. **32**(12), 7957-7968 (2020)