WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n1.04
Journal homepage: https://wepub.org/index.php/IJCSIT/index

A Review on Deep Learning Based Sign Language Recognition Translation

Junkai Huang *

Department of Faculty of Innovation Engineering, Macao university of Science and Technology, Macao, 999078, China

*Corresponding Author: junkai12356@iCloud.com

ABSTRACT

Hand posture is often referred to as gesture. It refers to the specific movements and postures that occur when a person utilizes his or her arms. It is one of the earliest and still widely used communication tools. In general, gestures are both dynamic and static. In the process of long-term social practice, gestures are given a variety of specific meanings, has a rich expressive power, the range of motion is extensive when the fingers, wrists, elbows, shoulders, and other joints of the hand are considered. The capacity for flexibility is considerable. Gestures have become a primary means of expression for humans, enabling the conveyance of emotions. In this context, body language occupies a pivotal role. Sign language is defined as the use of hand gestures with or without the use of facial expressions to convey meaning. The combination of hand gestures and facial expressions can be used to imitate images or to represent syllables, which together constitute a certain meaning or word. Sign language is a means of communication for individuals who are deaf or hard of hearing. It is an essential tool for communication, enabling them to express themselves and interact with others in a way that is accessible and understandable. This paper summarizes the deep learning based gesture recognition techniques in recent six years. In recent years, gesture recognition has been in the hot spot, and every year the articles on gesture recognition are constantly updated, so it is evident that a review of the research on gesture recognition is very necessary. This paper also compares and analyzes the existing gesture recognition methods and analyzes their advantages and disadvantages, and puts forward an outlook on the future development.

KEYWORDS

Hand Gesture Recognition; Sign Languages; Deep Learning Techniques; Performance Metrics

1. INTRODUCTION

Sign language is the basic means of communication between deaf people, and it is a very important human body language and a well-developed natural language. Hearing impaired people need to understand national and international events, national laws and regulations, as well as learn professional knowledge and make contact with the society in their social survival and development, so Proficiency in sign language is a prerequisite for those seeking employment in the field of deafness education. In the daily life of human beings, the use of sign language is also ubiquitous, such as traffic command, video call interactive functions and intelligent home appliance control. Consequently, the advancement of sign language recognition is of significant consequence to the enhancement of humanity's quality of life and the advancement of human scientia and technology. The research on gesture recognition has never stopped, but due to the diversity of sign languages and the need to pay attention to the detailed features of the hand, the sign language recognition technology needs to be constantly upgraded to improve its accuracy.

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

Research for gesture recognition is challenging. Gesture recognition is mainly categorized into static gesture recognition and dynamic gesture recognition, in this paper we focus on discussing a review of methods based on deep learning for dynamic gesture recognition. Traditionally, dynamic gesture recognition systems use several methods to extract hand-crafted features based on sequences modeling approaches such as Hidden Markov Models (HMM). However, since 2006, when Hinton et al proposed a deep learning approach, new opportunities have opened up for sign language recognition [1]. Convolutional neural networks (CNNs) have become a prevalent tool for visual feature learning in the field of computer vision. Nevertheless, 3D CNN techniques have been employed for video modelling purposes and are regarded as an extension of the standard CNN techniques associated with spatial-temporal filters. A particularly noteworthy attribute of 3D CNN is its capacity to directly generate a hierarchy of spatiotemporal data representations. However, it requires a large number of parameters compared to 2D CNN, which is considered to be the main drawback. In addition, 3D CNN contains a supplementary kernel dimension, which increases the complexity of network training [2].

In addition to this, there are other challenges for gesture recognition techniques, for example, the differentiation of transitional actions between two consecutive actions is difficult because from time to time, there is no obvious pause between two actions, and the system may not be able to differentiate between the two gestures, and therefore it may lead to incorrect or under-recognition of the action, which creates difficulty in the development of consecutive gesture Language Detection [3].

The remainder of this paper is organised as follows. The second part reviews the previous methods for deep learning-based gesture recognition in recent years, and then in the section three, we conduct an evaluation and comparative analysis of the methods in the second part, pointing out their strengths and weaknesses and looking forward to the future, and finally in the fourth part, we summarize the whole paper.

2. LITERATURE REVIEW ON HAND GESTURE RECOGNITION

2.1. Existing Works Based On Deep Learning Methods

In 2019, Sruthy Skaria et al. employed a miniature radar sensor to idenitify and analyse the Doppler signatures of precisely 14 different hand gestures [4]. A deep convolutional neural network (DCNN) is trained to categorise the gestures captured, with the input comprising the in-phase and cross-phase components of the beat signals from two receiving antennas of a continuous-wave Doppler radar. The two beat signals are then mapped into the DCNN, which comprises three input channels, as two 2-D spectrograms and one 3-D arrival angle matrix. The confusion between different gestures is very low.

In 2019, Liao et al., a novel multimodal approach to the recognition of dynamic sign languages was presented [5]. The method, which employs a deep 3-dimensional residual convolutional network (ConvNet) and a bidirectional long short-term memory network (LSTM network) to achieve its goal, has been designated the BLSTM-3D residual network (B3D ResNet). First, the hand was located within the video frames. The segmented hand features are then fed into LSTM with the original RGB data for long time dynamic feature modelling and final classification output. The experiment is carried out on two large datasets of isolated Chinese sign language words, the final output of the classification results achieves accurate recognition.

In 2019, Xiao et al also proposed a similar approach for dynamic gesture recognition [6].The system employs a 2-layer LSTM and a CHMM to integrate data from both hand and skeleton sequences.This method has good recognition results under different lighting and variable skin colors. It is also tested on two CSL datasets and the experimental results show that the method has better results.

In 2020, Zhang et al. proposed the end-to-end sign language recognition model based on spatiotemporal attention, which was formed based on Residual 3D Convolutional Networks

(Res3DCNN) and Time Attention Networks [7]. The Spatial Attention Networks automatically find the more important information of each frame, and the Time Attention Networks assign different weights to different frames, paying more attention to the important regions of the gesture.The recognition accuracy can be as high as 95.3% by using the Time Attention Networks to assign different weights to different frames, paying more attention to the important regions of the gesture.

In 2020, Aly and Aly devised a groundbreaking framework for signer-independent sign language recognition, utilising a combination of deep learning architectures comprising hand semantic segmentation and deep recurrent neural networks (RNNs) [8]. The DeepLabv3+, a recently developed semantic segmentation methodology, is trained using a set of pixel-labelled hand images with the objective of extracting hand regions from each video frame. The sequence of extracted feature vectors is then recognised using a deep Bi-directional Long Short-Term Memory (BiLSTM) recurrent neural network.

In 2020, Ameur et al. present a dynamic gesture recognition method using non-contact hand movements on the Leap Motion device [9]. First, sequential time series data collected from Leap Motion are analysed for recognition using a Long Short-Term Memory (LSTM) recurrent neural network. By combining the above models with other components, propose a final predictive network called HBU-LSTM. The proposed network markedly enhances the model performance by accounting for the spatial and temporal dependencies of the Leap Motion data with the network layers during the forward and backward passes.

In 2021, Kumar proposed a model using a series of 3D convolutional operations to convolve multiple frame blocks in one flow, followed by 3D maxpooling operations [10]. The obtained results are then flattened and fed into a MLP, and finally a soft maximum layer is applied to obtain the probabilities for each category to display the corresponding gesture representation. The trained framework has ensured the natural language outcomes corresponding to the signs ISL-associated.

In 2021, Microsoft Research proposed the Swin Transformer model. This model has the ability to focus on the global information of the transformer model, and also draws on the design concept of convolutional neural network, adopting downsampling and TSM technology, and comprehensively transforming the original transformer model for better understanding and processing.

In 2022, Prakash et al. put forth a methodology for fusing the outputs of two fine-tuned convolutional neural networks (CNNs), namely AlexNet and VGG-16 [11]. This technique entails an entire process of fine-tuning the deep CNNs from start to finish, with the training gesture samples of the target dataset serving as the input. Subsequently, a technique for combining the output scores of the fine-tuned deep CNNs is applied at the score level.

In 2022, Bhaumik et al present a portable end-to-end CNN-based network called ExtriDeNet [12]. ExtriDeNet demonstrates superior performance when confronted with challenging circumstances, such as variations in lighting and the complexities inherent in complex backgrounds. This is attributed to its straightforward structural design and the incorporation of multiscale features. ExtriDeNet is composed of two principal blocks: The IFFB and IFAB blocks are used in conjunction to fuse the outputs of two distinct receptive fields in order to capture the detailed spatial information present in hand postures. Only the dominant features are retained, while any unnecessary information present in the feature responses generated by the two differently sized filters and the IFFB output map are discarded. The multiscale filters facilitate the learning of features across a range of scales, enhancing ExtriDeNet's ability to extract subtle, localised features that are present consistently throughout the input image. In addition, the number of parameters to be trained is reduced, which reduces the complexity of the hand gesture recognition.

In 2022, H. Wu et al. have developed a new architecture, Convolutions to Vision Transformations (CvT), which improves Vision Transformations (ViTs) performance and efficiency by incorporating convolution into ViTs [13]. The motivation for the CvT architecture is also the introduction of locality

into the vision transformer architecture, with the expectation that the introduction of locality will result in a better trade-off between performance and efficiency. In particular, two improvements are proposed: a transformer hierarchy that is embedded in new convolutional markers, and a convolutional transformer module that makes use of convolutional projections.

In 2023, Altaf and et al. propose a deep learning-based approach for the static gesture recognition of numbers and alphabetic characters in the ISL dataset [14]. The method, which makes use of DenseNet201 in conjunction with transfer learning techniques, has the potential to achieve higher recognition accuracy. The DenseNet architecture is distinguished by a series of key advantages: vanishing gradient resolution, deep supervision, as well as regularisation techniques that minimise the likelihood of overfitting due to the inherent limitations of small sample sizes in training sets.

In 2024, Huang et al present a novel framework for sign language recognition that integrates the R(2+1)D and LSTM networks [15]. The initial stage of the process entails the extraction of spatiotemporal features from sign language sequences, which is achieved through the utilisation of the R (2+1)D network. To capture long-term features and eliminate redundant information, these features are then fed into the LSTM network. They evaluate the proposed network on two different sign language datasets, which achieved 96.21% accuracy on the CSL dataset and 99.69% accuracy on the LSA64 dataset.

In 2024, Kankariya et al proposed a method using Convolutional Neural Network (CNN) and Inception v3 algorithms [16]. These algorithms have shown favourable results in the field of image-based recognition tasks. The Inception network is a variant of CNNs, introducing a novel architecture to address computational efficiency and network depth. It aims to improve the performance and accuracy of models used to recognise gestures in signing. The Inception v3 algorithm achieved an accuracy of 98.94%.

3. FINDINGS OF THE REVIEW

Table 1.1. Comparision of existing systems

Author	Techniques	Advantages	Disadvantages	Year
Kankariya et al	CNN, Inception v3 algorithms	With a multi-layered network structure and a large amount of training data, this approach is able to learn subtle gesture features for high accuracy recognition.Inception v3 is a modular network structure that allows modules to be added or removed to adjust the depth and width of the network as needed. This allows the method to be adapted and optimised for specific application scenarios and computational resources.	High consumption of computational resources.Models have limited generalisation ability and may misrecognise new gestures when faced with large differences from the training data.High challenge for dynamic gesture recognition.	2024
Huang et al	Integrates the R(2+1)D and LSTM networks	Powerful feature extraction; Long-term dependency capture; Efficient, R(2+1)D networks reduce the computation by decomposing the 3D convolution, while LSTM networks prevent the gradient vanishing and gradient explosion problems by their gating mechanism.	The number of parameters increases; the training time is relatively long; and the system is more costly and difficult to maintain.	2024
Altaf et al	CNN, DenseNet201, Transfer Learning	Enhanced feature propagation and reuse helps to improve gradient flow and sparsity learning, which increases the accuracy of sign language recognition; reduces the number of parameters; and improves training efficiency.	If the differences between the source and target domains are very large, then transfer learning may not work well; weight selection and similarity measures rely on experience.	2023
H. Wu et al	CvT, ViT, CNN	Capable of capturing both local details and the overall structure of the gesture; handles inputs of arbitrary size;	Sensitive to the size of the embedding vectors, CvT is quite sensitive to the size of the embedding vectors for each element; possible overfitting.	2022

Table 1.2. Comparision of existing systems

Author	Techniques	Advantages	Disadvantages	Year
Bhaumik et al	CNN, ExtriDeNet, IFFB and IFAB	Strong feature extraction capability; high real-time performance; ExtriDeNet may enhance the robustness of the model to background noise, light variations, and other disturbing factors by introducing an attention mechanism or employing other techniques.	ExtriDeNet is not a widely recognised standard model, so it may lack substantial research support and experimental validation. This makes it difficult to evaluate its performance in gesture recognition.	2022
Prakash et al	CNN, AlexNet and VGG-16	High recognition rate in image classification problems; lightweight network structure for real-time gesture recognition; good scalability.	Compute resource requirements are relatively high. This may limit applications on some resource-constrained devices; the performance of fine-tuned convolutional neural networks is highly dependent on the amount and quality of training data.	2022
Microsoft Research	Swin Transformer model, downsampling and TSM	Swin Transformer employs a hierarchical structure; a shift window; a multi-head self-attention mechanism that allows the model to learn different feature representations in different subspaces; and residual connectivity to speed up the model training process.	The computational cost is relatively high; insufficient or low-quality training data may lead to degradation of model performance; and high complexity, including a large number of parameters and computations.	2021
Kumar	3D convolutional operations, MLP	Dynamic gesture recognition capability; efficient processing of video or continuous image frame data; wide range of applications.	More computational resources are required; the performance of the model is highly dependent on the quality and quantity of training data. Further optimisation is required in application scenarios with high real-time requirements.	2021
Ameur et al	Leap Motion device, LSTM, HBU-LSTM	High accuracy, LSTM networks have advantages in handling sequential data and long-term dependencies. Ease of use and scalability, the system can also be adapted to different application scenarios by adding new gesture actions. Improve the efficiency and intelligence of human-computer interaction.	Dependence on hardware; Requirement on environment; Training of LSTM networks requires a large amount of labelled data and computational resources. At the same time. fine tuning and parameter adjustment of the model may be required to obtain better recognition results.	2020
Aly and Aly	RNN, DeepLabv3+, BiLSTM	BiLSTM is able to handle both forward and backward time series information; long term dependency modelling, LSTM networks are designed to address long term in RNNs are able to better deal with long term dependencies that may exist in gestures; adaptable; stable performance.	High computational complexity; high data requirements, if there is imbalance or noise in the data, it will affect the performance of the model; high complexity of the model; high hardware requirements, the cost of the system and the difficulty of maintenance.	2020
Zhang et al	spatiotemporal attention, Res3DCNN and Time Attention Networks	Enhancement of spatio-temporal feature learning capability; more focus on frames that are important to the recognition result, reducing the interference of redundant information; and reduction of computational cost.	Model complexity increases, model training becomes more difficult, requiring more computational resources and time; data requirements increase; hyperparameter tuning is difficult, requiring more experimentation and debugging.	2020
Xiao et al	LSTM and a hidden Markov model (CHMM)	It can effectively capture long-term dependencies in sequence data; it can flexibly describe the state transfer and observation generation process in sequence data; LSTM and HMM can complement each other. Performance Enhancement.	High computational complexity slows down training and inference; numerous parameters may lead to overfitting problems; increased workload for data collection and labelling.	2019
Liao et al	BLSTM-3D residual network (B3D ResNet)	Powerful spatio-temporal feature learning capability, good at processing 3D data and capturing complex spatio-temporal features in gestures; efficiently dealing with long-term dependencies in gesture sequences; mitigating the problem of gradient vanishing in deep neural networks, allowing the network to be trained deeper.	High computational complexity; high data requirements; high model complexity; reduced model interpretability, which may be relatively low due to the complexity of BLSTM-3D ResNet. This makes it difficult to understand and analyse the behaviour and performance of model.	2019
Sruthy Skaria et al	miniature radar sensor, DCNN	Strong feature extraction ability; DCNN has strong generalisation ability and can adapt to different types of gesture recognition tasks;Strong ability to handle complex scenarios	DCNN contains a large number of parameters and computations, and has high requirements for computational resources; it requires a large amount of labelled data for training. In some extreme cases, light changes and background interference may have a large impact on the recognition results.	2019

4. CONCLUSION

Sign language recognition and translation research based on deep learning has become a popular and challenging field with the rapid development of science and technology. In this paper, the advantages and limitations of different recognition methods are analysed, and the research progress of sign language recognition technology under deep learning framework in recent years is reviewed. The accuracy of sign language recognition and translation is crucial to promote communication between the hearing impaired and society, as the main mode of communication for the hearing impaired.

Through the review of the literature, we can see that the development of sign language recognition technology has seen a shift from the traditional methods of pattern recognition to deep learning techniques. However, deep learning still faces some challenges in the field of sign language recognition, despite the significant progress it has made. Data collection and annotation are essential in sign language recognition, yet the lack of large, high-quality sign language data sets limits the training and performance improvement of deep learning models. To address the above challenges, in terms of improving data collection and annotation, future research in sign language recognition can build larger and higher quality sign language datasets.

To sum up, deep learning sign language recognition and translation studies have great theoretical value and practical importance. With the continuous advancement of technology and deep learning research, the sign language recognition system will become more accurate in understanding and translating sign language in the future, thus providing a more convenient and efficient way for people with hearing disabilities to communicate.

REFERENCES

[1] HINTON G E, OSINDERO S, and TEH Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18(7): 1527–1554. doi: 10.1162/neco.2006.18.7.1527.

[2] Snehal Abhijeet Gaikwad, Dhananjay Upasani, Virendra Shete, “Review and Trends on Hand Gesture Recognition of Sign Language based on Deep Learning Approaches” IEEE Xplore Part Number: CFP23BC3-ART; ISBN: 978-1-6654-5630-2.

[3] H. Cooper, B. Holt, and R. Bowden, “Sign language recognition,” in Visual Analysis of Humans. London, U.K.: Springer, 2011, pp. 539–562.

[4] S. Skaria, A. Al-Hourani, M. Lech and R. J. Evans, "Hand-Gesture Recognition Using Two-Antenna Doppler Radar With Deep Convolutional Neural Networks," in IEEE Sensors Journal, vol. 19, no. 8, pp. 3041-3048, 15 April15, 2019, doi: 10.1109/JSEN.2019.2892073.

[5] LIAO Yanqiu, XIONG Pengwen, MIN Weidong, et al.Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks [J]. IEEE Access, 2019, 7: 38044–38054. doi: 10.1109/ACCESS.2019.2904749.

[6] Xiao Q, Qin M, Guo P, et al. Multimodal fusion based on LSTM and a couple conditional hidden Markov model for Chinese sign language recognition [J]. IEEE Access, 2019, 7: 112258-112268.

[7] Luo Yuan, Li Dan, Zhang Yi. Chinese sign language recognition based on spatio-temporal attention network [J]. Semiconductor Photonics, 2020, 41 (03) 414-419. DOI: 10. 16818 /j.issn1001-5868.2020.03.021.

[8] S. Aly and W. Aly, "DeepArSLR: A Novel Signer-Independent Deep Learning Framework for Isolated Arabic Sign Language Gestures Recognition," in IEEE Access, vol. 8, pp. 83199-83212, 2020, doi: 10.1109/ACCESS.2020.2990699.

[9] Safa Ameur, Anouar Ben Khalifa, Med Salim Bouhlel, A novel hybrid bidirectional unidirectional LSTM network for dynamic hand gesture recognition with Leap Motion, Entertainment Computing, Volume 35, 2020, 100373, ISSN 1875- 95 21, https://doi.org/10.1016/j.entcom.2020.100373.

[10] Dushyant Kumar Singh, 3D-CNN based Dynamic Gesture Recognition for Indian Sign Language Modeling, Procedia Computer Science, Volume 189, 2021, Pages 76-83, ISSN 1877- 050 9, https://doi.org/10.1016/j.procs.2021.05.071.

[11] Sahoo, Jaya Prakash, Allam Jaya Prakash, Paweł Pławiak, and Saunak Samantray. 2022. "Real-Time Hand Gesture Recognition Using Fine-Tuned Convolutional Neural Network" Sensors 22, no. 3: 706. https://doi.org/10.3390/s22030706

[12] Bhaumik, G., Verma, M., Govil, M.C. et al. ExtriDeNet: an intensive feature extrication deep network for hand gesture recognition. Vis Comput 38, 3853–3866 (2022). https://doi.org/10.1007/s00371-021-02225-z

[13] H. Wu et al., "CvT: Introducing Convolutions to Vision Transformers," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 22-31, doi: 10.1109/ICCV48922.2021.00009.

[14] Y. Altaf, A. Wahid and M. M. Kirmani, "Deep Learning Approach for Sign Language Recognition Using DenseNet201 with Transfer Learning," 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2023, pp. 1-6, doi: 10.1109/SCEECS57921.2023.10063044.

[15] J. Huang, J. Chaijaruwanich and V. Chouvatut, "Video-based Sign Language Recognition with R (2+1)D and LSTM Networks," 2024 16th International Conference on Knowledge and Smart Technology (KST), Krabi, Thailand, 2024, pp. 214-219, doi: 10.1109/KST61284.2024.10499646.

[16] S. Kankariya, K. Thakre, U. Solanki, S. Mali and A. Chunawale, "Sign Language Gestures Recognition using CNN and Inception v3," 2024 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2024, pp. 1-6, doi: 10.1109/ESCI59607.2024.10497401.