Advancements in Gesture Recognition: From Traditional Machine Learning to Deep Learning Innovations

Qingyang Wang
School of Artificial Intelligence, Shanghai Normal University Tianhua College, Shanghai, 201815, China

Keywords: Gesture Recognition, Machine Learning, Deep Learning, Transformer.

Abstract: Gesture recognition is essential for creating more efficient human-computer interactions, transforming the way people communicate with and control technology. With improvements in computer performance and the development of image processing technology, researchers have begun to explore how to automatically extract useful information to achieve effective gesture recognition. This paper focuses on the advancements in gesture recognition, highlighting the progression from conventional machine learning to state-of-the-art deep learning approaches. Traditional machine learning is limited by its feature dependency and offers limited accuracy but has low computational complexity and strong interpretability. Convolutional Neural Network (CNN)-based methods are characterized by automatic feature extraction, high recognition accuracy, and adaptability to complex environments, but they come with high computational demands and data dependence. Transformer-based methods excel in capturing global information and have high recognition accuracy potential but are affected by extremely high computational complexity and a vast model optimization space. In summary, each of the three gesture recognition methods has its own benefits and disadvantages, and in real-world applications, the best approach should be chosen depending on specific needs and scenarios.

1 INTRODUCTION

Gesture recognition makes interactions more natural and intuitive. Direct control of electronic devices via gestures can smoothen communication between humans and machines (Khan, 2012). It also has the ability to increase effectiveness and accuracy, particularly in fields demanding precise control, such as medical surgery or industrial manufacturing (Oudah, 2020). Here, gesture recognition facilitates more accurate operations, reduces errors, and enhances work efficiency. Furthermore, gesture recognition empowers special groups to utilize electronic devices more effectively. For instance, individuals with poor eyesight may struggle to discern screen buttons clearly, but gesture recognition allows them to control devices with greater ease. In summary, the significance of gesture recognition lies in its contribution to a more convenient and efficient lifestyle, as well as fostering a more diverse and inclusive world.

Machine learning enhances gesture recognition through high accuracy and real-time performance. This is achieved by algorithms that learn from extensive data, ensuring precision in control-intensive applications like surgery and manufacturing (Mahesh, 2020). It also provides real-time responses, thanks to advancements in machine learning frameworks and hardware. The technology is adaptable, accommodating various environmental conditions and recognizing a wide range of gestures, from static to dynamic, through continuous learning. This adaptability enhances the user experience by facilitating natural and intuitive interactions, reducing reliance on external devices, and improving engagement.

Furthermore, machine learning automates the feature extraction from gesture data, simplifying system development and enhancing generalization (Liakos, 2018). It also processes large datasets to uncover patterns, thereby improving the accuracy of gesture recognition. With a broad range of applications, from educational tools that enhance classroom dynamics to medical applications that assist in surgery and rehabilitation, gesture recognition is a transformative technology. It also enriches entertainment, offering immersive gaming experiences and extending to smart homes and industrial controls, where it simplifies device operation and robotic line management. These

Wang, Q.
Advancements in Gesture Recognition: From Traditional Machine Learning to Deep Learning Innovations.
DOI: 10.5220/0013332100004558

In *Proceedings of the 1st International Conference on Modern Logistics and Supply Chain Management (MLSCM 2024)*, pages 372-376
ISBN: 978-989-758-738-2

capabilities position gesture recognition as a pivotal technology in advancing human-computer interaction.

2 RECOGNITION OF GESTURES FOUND IN TRADITIONAL MACHINERY LEARNING

The conventional gesture recognition with machine learning primarily relies on manually designed feature extractors to identify the characteristics of gestures, and then utilizes these characteristics to train machine learning models to achieve gesture classification and recognition.

When dealing with gesture recognition tasks, the method typically follows these steps in the experimental process: (1) Data collection: It is necessary to gather a large amount of gesture information, including videos, images, or depth data, which will be used to train machine learning models. (2) Preprocessing: The collected gesture data is preprocessed to enhance the accuracy of subsequent feature extraction and recognition. Preprocessing steps may include denoising, image enhancement, normalization, etc. (3 Feature extraction: This is an essential phase in traditional machine learning-based gesture recognition. At this step, a manually designed feature extractor is needed to identify key features from the preprocessed gesture data. (4) Feature selection: Since many features are extracted, to enhance the effectiveness and precision of the model, feature selection is usually required. This involves choosing the most useful subset of features for classification tasks. (5) Model training: A machine learning model can be trained using the selected feature set. Commonly used models in traditional machine learning for gesture recognition include Support Vector Machine (SVM), Hidden Markov Model (HMM), and Random Forest, among others. (6) Model assessment and refinement: To increase the trained model's accuracy and robustness, it is evaluated using a test set, and necessary adjustments and optimizations are made based on the evaluation's findings. (7) Deployment and application: The optimized model is deployed to practical application scenarios, such as human-computer interaction systems, smart home control, virtual reality experiences, etc. (Jordan, 2015).

Typical features for classification include the Histogram of Oriented Gradients (HOG), a feature descriptor for object detection. It creates features by computing and statistically analyzing the local region gradient direction histograms in an image. Additionally, there is the Scale-Invariant Feature Transform (SIFT), an algorithm for extracting local features from images. SIFT demonstrates stability against perspective changes, affine transformations, noise, and rotation while being invariant to scaling, brightness, and rotation (Khalid, 2014).

Traditional machine learning gesture recognition algorithms heavily rely on manually designed feature extractors, which require researchers to have extensive domain knowledge and experience. However, these manually designed feature extractors might not adequately capture the intricate nuances of gestures. The design of feature extractors and classifiers is often tailored for specific scenarios and tasks, which can result in weak generalization to new scenarios and tasks. Additionally, the performance of these algorithms is constrained by the standard and quantity of the training data; insufficient data volume or poor data quality can compromise recognition accuracy. Despite these limitations, traditional machine learning offers certain advantages. For instance, its gesture recognition algorithms, which are based on manual feature extractors and classifiers, have a relatively intuitive decision-making process that is easy to understand and explain. Compared to deep learning, these algorithms typically demand fewer computing resources, making them suitable for resource-constrained situations. In certain cases, with carefully designed feature extractors and classifiers, traditional machine learning algorithms can achieve high recognition accuracy and robustness, justifying their continued relevance in gesture recognition research.

3 RECOGNITION OF GESTURES DUE TO CONVOLUTIONAL NEURAL NETWORK

This line of work refers to applying convolutional neural networks (CNNs) to process input gesture images or video streams, automatically extracting image features, learning gesture patterns, and classifying them to achieve accurate recognition of gesture actions. This process usually comprises actions such as gathering and preparing data, building and training network models, and evaluating and applying models (Alzubaidi, 2021).

In representative previous research, a dynamic gesture recognition method is presented that leverages a 2D CNN and feature fusion to achieve high efficiency and accuracy. It aims to address the high complexity and low efficiency of traditional 3D

CNN-based dynamic gesture recognition methods by proposing a more efficient two-dimensional CNN approach utilizing feature fusion for improved precision and reduced computational demands. It employs original frames and keyframes for optical flow to capture both temporal and spatial characteristics, which are then fused and recognized by the 2D CNN. A fractional-order Horn and Schunck method extracts high-quality optical flow, and an improved clustering algorithm identifies keyframes, reducing data redundancy. The suggested dynamic gesture identification technique achieved high accuracy rates of 98.6% on the Cambridge dataset and 97.6% on the Northwestern University dataset, outperforming alternative techniques. The model, with only 0.44 million parameters, significantly reduced computational complexity and training time compared to conventional 3D CNN models. Ablation studies confirmed the efficiency of fractional-order optical flow and keyframe retrieval, enhancing recognition accuracy by over 10%. The approach demonstrates efficient gesture recognition with minimal parameters and fast computation time (Yu, 2022).

Another representative work proposes enhancing human-computer interaction by developing a highly accurate, hardware-free static hand gesture recognition system using CNNs. The method involves preprocessing with skin segmentation and data augmentation to enhance model accuracy. The CNN architecture consists of seven layers, including max-pooling and convolutional layers, followed by a fully connected layer. Dropout is applied to prevent overfitting. The model is trained using the cross-entropy loss function and Adam optimizer. The study's results demonstrated outstanding performance of the proposed CNN model in recognizing static hand gestures, achieving testing accuracies of 96.5% on the NUS II dataset and 96.57% on the Marcel dataset. The accuracy of the model was greatly increased by the incorporation of skin segmentation and data augmentation, reducing misclassification rates. The experiments confirmed the effectiveness of the CNN approach in gesture recognition tasks, even with complex backgrounds (Eid, 2023).

In summary, the accuracy of CNN-based gesture recognition systems may reach over 90%, or even approach or exceed 99%, in some relatively simple application scenarios, such as recognizing a limited number of gesture types with little variation. However, in more complex and diverse application scenarios, such as recognizing a large number of gesture types with subtle differences and presented under various lighting conditions and angles, the accuracy may be relatively lower, but it can still outperform traditional machine learning or image processing methods. The field of gesture recognition based on convolutional neural networks is currently experiencing rapid development and continuous innovation. As a result of ongoing technological advancements and the expansion of application scenarios, gesture recognition technology will become increasingly essential in various industries.

4 GESTURE RECOGNITION BASED ON TRANSFORMER

The Transformer model possesses significant advantages due to its self-attention mechanism, which allows for dynamic weighting of input data, enabling it to focus on relevant features. This results in efficient processing and understanding of sequences, making it particularly adept at handling long-range dependencies. In gesture recognition, these capabilities translate to robust interpretation of gesture sequences, facilitating accurate identification even in complex environments. The model's capacity to capture minute details and temporal dynamics within gestures leads to enhanced recognition accuracy and real-time performance, making it a potent tool for gesture-based human-computer interaction (Ahmed, 2023).

Motivated by the goal of developing an efficient and accurate hand gesture recognition framework, a representative work utilizes high-density surface Electromyography (EMG) signals and deep learning, aiming to enhance prosthetic hand control and human-machine interactions. The research presents a Vision Transformer network-based Compact Transformer-based Hand Gesture Recognition (CTHGR) framework for hand gesture classification using high-density surface EMG signals. The framework employs a method of attention for feature extraction and leverages both spatial and temporal features without requiring transfer learning. It incorporates a hybrid model that fuses macroscopic EMG data with microscopic neural drive information extracted via Blind Source Separation, enhancing gesture recognition accuracy. The method is evaluated using various window sizes and electrode channels, demonstrating improved performance over conventional deep learning and machine learning models. The study's results show that the proposed CTHGR framework achieves high accuracy in identifying hand movements using HD-sEMG signals, with average accuracies ranging from 86.23% to 91.98% across different electrode channels and window widths. The framework outperforms 3D CNN models and traditional machine learning,

showing significant improvements in accuracy, precision, and recall. Instantaneous recognition and a hybrid model that incorporates information on both macroscopic and microscopic neural drives further enhance the framework's performance, validating its effectiveness for real-time applications (Montazerin, 2023).

An additional representative study seeks to bridge the gap between feature data and gesture recognition needs by developing the Long Short-Term Transformer Feature Fusion Network (LST-EMG-Net), a deep learning model that fuses long and short-term surface electromyography (sEMG) features for accurate gesture recognition in various applications. It employs an extended brief encoder to extract multi-scale features from sEMG windows, as well as a cross-attention feature module for efficient feature fusion. The model dynamically adjusts channel weights using sEMG channel attention and incorporates signal augmentation to expand the training dataset. LST-EMG-Net demonstrates improved accuracy and stability in recognizing gestures across different datasets. The Ninapro DB2E2, DB5E3 partial gesture, and CapgMyo DB-c datasets yielded high accuracy rates of 81.47%, 88.24%, and 98.95% for the LST-EMG-Net, respectively. The model outperformed other networks in accuracy and stability, effectively recognizing various gesture types. The results validate LST-EMG-Net's capability for accurate and stable sEMG-based gesture recognition, demonstrating its potential for applications in rehabilitation and human-computer interaction (Zhang, 2023).

In summary, gesture recognition based on the Transformer model encompasses multiple aspects such as model application, data processing, sequence modeling, classification recognition, and application scenarios. As deep learning technology advances further, gesture recognition based on the Transformer is currently experiencing rapid development, and technological innovation continues to promote the expansion of its application scenarios. In the future, as technology continues to advance and mature, gesture recognition technology based on the Transformer is expected to play an increasingly important role in a variety of fields.

5 DISCUSSIONS

Comparing these methods, it is evident that each has its own specialty. Traditional machine learning is suitable for simple, explainable tasks with limited data. CNNs are ideal for environments where high accuracy and adaptability are paramount. Transformers, with their advanced feature handling capabilities, are poised to excel in complex, real-time gesture recognition scenarios.

Looking ahead, the future of gesture recognition lies in the continued development of more efficient algorithms that can balance accuracy with computational efficiency. There is a need for models that can generalize well across diverse datasets and scenarios without compromising performance. Additionally, combining data from multiple modalities and the development of more robust preprocessing techniques will likely enhance the accuracy and reliability of gesture recognition systems.

In conclusion, while each method has its strengths and weaknesses, the field is moving towards more sophisticated deep learning models that can better interpret and respond to human gestures. As research progresses, it is expected that we will see more innovative solutions that address the existing constraints and extend the boundaries of human-computer interaction.

6 CONCLUSIONS

Utilizing conventional machine learning, convolutional neural networks, and Transformer-based methods for gesture recognition, although they differ in specific implementations, they all aim to achieve accurate recognition of gestures by analyzing and understanding gesture information in images or videos. These methods share a consistent core purpose: to extract effective features from gestures and classify or recognize them based on these features.

Utilizing conventional machine learning, convolutional neural networks, and Transformer-based methods for gesture recognition, although they differ in specific implementations, they all aim to achieve accurate recognition of gestures by analyzing and understanding gesture information in images or videos. These methods share a consistent core purpose: to extract effective features from gestures and classify or recognize them based on these features.

However, all three types of research share a common drawback: insufficient robustness. This means that all gesture recognition methods face challenges from complex and changing environmental factors, such as lighting changes, occlusion, and background interference, which may lead to a decline in recognition performance.

Regarding the future prospects of gesture recognition technology. (1) Model lightweighting:

Developing lighter model structures to reduce computational complexity and parameter count can improve real-time performance and applicability. (2) Cross-modal fusion: Combining multimodal information such as images, speech, and text can lead to more natural and intelligent gesture recognition and interaction. (3) Robustness enhancement: Introducing techniques such as adversarial training and data augmentation can increase the resilience of gesture recognition techniques in a range of challenging situations.

With the continuous progress and innovation in fields such as deep learning and computer vision, it is believed that gesture recognition technology will achieve more widespread applications and breakthroughs in the future.

REFERENCES

Ahmed, S., Nielsen, I. E., Tripathi, A., Siddiqui, S., Ramachandran, R. P., & Rasool, G. 2023. Transformers in time-series analysis: A tutorial. *Circuits, Systems, and Signal Processing*, *42*(12), 7433-7466.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... & Farhan, L. 2021. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data*, *8*, 1-74.

Eid, A., & Schwenker, F. 2023. Visual Static Hand Gesture Recognition Using Convolutional Neural Network. *Algorithms*, *16*(8), 361.

Jordan, M. I., & Mitchell, T. M. 2015. Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Khalid, S., Khalil, T., & Nasreen, S. 2014. A survey of feature selection and feature extraction techniques in machine learning. In *science and information conference*. 372-378.

Khan, R. Z., & Ibraheem, N. A. 2012. Hand gesture recognition: a literature review. *International journal of artificial Intelligence & Applications*, *3*(4), 161.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. 2018. Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674.

Mahesh, B. 2020. Machine learning algorithms-a review. *International Journal of Science and Research*, *9*(1), 381-386.

Montazerin, M., Rahimian, E., Naderkhani, F., Atashzar, S. F., Yanushkevich, S., & Mohammadi, A. 2023. Transformer-based hand gesture recognition from instantaneous to fused neural decomposition of high-density EMG signals. *Scientific reports*, *13*(1), 11000.

Oudah, M., Al-Naji, A., & Chahl, J. 2020. Hand gesture recognition based on computer vision: a review of techniques. *journal of Imaging*, *6*(8), 73.

Yu, J., Qin, M., & Zhou, S. 2022. Dynamic gesture recognition based on 2D convolutional neural network and feature fusion. *Scientific Reports*, *12*(1), 4345.

Zhang, W., Zhao, T., Zhang, J., & Wang, Y. 2023. LST-EMG-Net: Long short-term transformer feature fusion network for sEMG gesture recognition. *Frontiers in Neurorobotics*, *17*, 1127338.