ISSN 2959-6157

A Review of Deep Learning Based Video Action Recognition Techniques

Mingyuan Zhu

Hyde College, Ocean University of China & The University of Adelaide, Qingdao, Shandong, 266100, China
Email: mingyuan29@gmail.com

Abstract:

Video action recognition is an important research direction in the field of computer vision and pattern recognition, with extensive applications in intelligent video surveillance, human-computer interaction, and sports analysis. The development of data storage and computing hardware over the past decade has driven a shift from traditional feature extraction and machine learning algorithms to deep learning-based approaches. This paper reviews the current state of development, problems, and future research directions of video action recognition techniques. Traditional methods are gradually being replaced by deep learning methods such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long-short-term memory networks (LSTMs). These methods automatically extract features and handle time-dependency, significantly improving the accuracy and robustness of action recognition. In particular, models based on the attention mechanism further enhance action recognition performance by dynamically adjusting the focus of attention, a current hot spot in research. Despite many advances, video action recognition still faces several challenges, including high computational resource requirements, complex model training, dataset bias issues, and variations in real-world application scenarios such as viewpoint changes, lighting changes, and occlusion. Future research can explore multi-modal fusion, lightweight models, self-supervised learning, and cross-domain transfer learning to improve the accuracy, robustness, and generalization of action recognition. The review provided aims to offer researchers a comprehensive perspective on the current state of development and future research directions of video action recognition technology.

Keywords: Video Action Recognition, Deep Learning, CNN, RNN, Attention Mechanisms, Multi modal Fusion, Self-Supervised Learning, Cross-Domain Migration Learning

1. Introduction

Video action recognition technology is one of the important research directions in the field of computer vision and pattern recognition, and it has a wide range of applications in intelligent video surveillance, human-computer interaction, and sports analysis [1]. Only 10 years ago, the field mainly relied on manual feature extraction and general machine learning algorithms due to limitations in data storage capacity and computing hardware [2]. With the development of computing chips and storage units, deep learning-based video action recognition methods have gradually become a research hotspot, making significant progress in several aspects [3][4]. Video content analysis is a key technology for realizing intelligent video applications, with action recognition playing a crucial role [4]. In intelligent video surveillance, recognizing human actions in the video enables functions like abnormal behavior detection and intrusion detection, thereby improving security levels [5]. In human-computer interaction, recognizing user actions allows for more natural and efficient interactions [5]. In sports analysis, automating the recognition of athletes' actions assists referees in judging penalties and improves the accuracy and efficiency of game analysis [1]. Video action recognition technology has evolved from traditional methods to deep learning methods. Traditional methods mainly relied on hand-designed features, such as optical flow features and trajectory features, which could recognize simple actions to some extent. However, with the increase in video data volume and complexity, the limitations of traditional methods have become apparent [3]. In recent years, the introduction of deep learning, especially CNNs and RNNs, has provided new approaches for video action recognition. Deep learning methods have dramatically improved the accuracy and robustness of action recognition by automatically learning features from video data [3][5]. The purpose of this review is to provide an overview of the development status and application of deep learning-based video action recognition technology. Traditional video action recognition methods, including model-based and feature-based methods, are introduced first. The application of deep learning in video action recognition is then discussed in detail, covering models such as CNNs, RNNs, and LSTMs. Commonly used action recognition datasets and performance evaluation indexes are also introduced. Finally, current challenges and problems in action recognition are analyzed, and future research directions are discussed.

2. Literature review

2.1 Traditional Video Action Recognition Methods

Traditional video action recognition techniques are mainly divided into model-based methods and feature-based methods. Model-based methods simulate human motion for action recognition. Early research primarily used models like Hidden Markov Model (HMM) and Dynamic Time Warping (DTW) to achieve action recognition by tracking and matching the key points of the human body [6][7]. These methods perform well for simple actions but struggle with complex actions and occlusion problems [7][8]. Feature-based methods rely on features extracted from the video, such as optical flow, trajectory, and shape [7][8]. For instance, Improved Dense Trajectories (IDT) features perform well in complex scenes and actions [7]. However, these methods often require extensive preprocessing and feature engineering, have high computational complexity, and are less robust to environmental changes and occlusions [8].

2.2 Deep Learning Based Action Recognition Methods

With the rise of deep learning, especially the application of CNN and RNN, video action recognition techniques have made significant progress [9]. CNNs excel in image feature extraction and are suitable for processing spatial information in video frames. Simonyan and Zisserman proposed a dual-stream CNN model that achieves efficient recognition of video actions by fusing RGB images and optical flow images [11]. Additionally, the 3D CNN model can capture both spatial and temporal information in the video, improving the accuracy of action recognition [12]. RNN and LSTM are effective at processing sequential data and are suitable for capturing temporal dependencies in videos. The unsupervised video representation learning method based on LSTM proposed by Srivastava et al. efficiently handles long video sequences, significantly improving action recognition performance [13]. LSTM, through its gating mechanism, can memorize and forget important information in time series, making it suitable for complex action recognition tasks [14]. In addition to CNN and RNN, other deep learning models such as Transformer are also gradually being applied to video action recognition. Transformer captures long-distance dependencies in videos through its self-attention mechanism, enhancing the accuracy of action recognition [15].

2.3 Action Recognition Dataset and Evaluation Metrics

Commonly used action recognition datasets include KTH, UCF101, and HMDB-51 [6][8]. The KTH dataset contains six basic actions, such as walking and running, performed by 25 individuals in four different scenarios, totaling 2,391 videos. This dataset features simple scenes and is mainly used to initially verify the effectiveness of algorithms [6]. The UCF101 dataset includes 101 movement categories and 13,320 videos, covering a wide range of fields such as sports and daily life, making it one of the standard datasets widely used today [3]. The HMDB-51 dataset comprises 51 movement categories and 5,100 videos from various sources and complex scenes, serving as a standard dataset to test algorithm robustness [6][8]. The complexity of this dataset is a good test for the robustness and generalization ability of algorithms [3]. In terms of evaluation metrics, action recognition ability is assessed using Accuracy and Mean Average Precision (mAP) [8]. These metrics measure the classification ability and robustness of the models and reflect their practical application potential.

2.4 Challenges and Issues in Action Recognition

Although deep learning methods have made significant progress in video action recognition, they still face several challenges and problems. One major challenge is viewing angle variation; the performance of the same action can vary greatly under different camera angles, leading to a decrease in model recognition accuracy [6]. Occlusion and lighting variations also pose problems; in practical applications, human actions may be partially occluded or performed under different lighting conditions, affecting feature extraction and recognition [8]. The dataset bias problem is another issue; the samples in the training dataset are often similar, leading to insufficient generalization ability of the model in new environments [9]. Additionally, deep learning models usually require significant computational resources, and processing long video sequences in real-time remains a challenge [6].

2.5 Directions for future research

Future research can explore the following aspects: multimodal fusion, which improves the accuracy and robustness of action recognition by combining multiple modal information such as RGB images, depth images, and infrared images [9]. Research on lightweight models aims to reduce computational complexity and achieve real-time action recognition [3]. Self-supervised learning leverages unlabeled data for pre-training to enhance model performance in small sample environments [9][15]. Cross-domain migration learning, where a model trained on one dataset is transferred to another different but related dataset, can improve model performance and generalization ability in new environments [6][9][15].

3. Deep learning based action recognition method

3.1 CNN

CNNs excel in image feature extraction and are particularly suitable for processing spatial information in video frames. The dual-stream CNN model proposed by Simonyan and Zisserman achieves efficient recognition of video actions by fusing RGB images and optical flow images [11]. Additionally, the 3D CNN model captures both spatial and temporal information in videos, improving the accuracy of action recognition [12]. The advantages of CNNs include efficient feature extraction and strong robustness, making them suitable for a variety of complex scenes. However, CNNs are computationally resource-intensive, challenging to process in real-time, and difficult to capture temporal dependencies in videos when used alone.

3.2 RNN and LSTM

RNN and LSTM excel at processing sequential data and are suitable for capturing temporal dependencies in videos. The unsupervised video representation learning method based on LSTM proposed by Srivastava et al. effectively processes long video sequences, significantly improving action recognition performance [13]. LSTM, through its gating mechanism, can memorize and forget important information in time series, making it suitable for handling complex action recognition tasks [14]. The advantage of RNN and LSTM lies in their ability to capture temporal dependencies effectively, making them suitable for long time-series data. However, training these models is complex, prone to vanishing or exploding gradient problems, and requires high computational resources, posing challenges for real-time applications.

3.3 Models based on attention mechanisms

In recent years, attention mechanisms have been widely used in video action recognition. The attention mechanism enhances the accuracy and robustness of models by automatically focusing on important regions and time segments in the video. The time-domain attention mechanism captures important action changes in a video sequence by assigning different attention weights to various time steps. The Temporal Segment Networks (TSN) model proposed by Wang et al. improves action recognition by segmenting the video into several parts and applying the attention mechanism to each segment, capturing the temporal information of the action, thus improving accuracy [15]. Additionally, the spatial-domain attention mechanism captures important features in each frame by assigning different attention weights to different regions of the image. The Action Transformer model proposed by Girdhar et al. achieves efficient video action recognition by combining spatial-domain and temporal-domain attention mechanisms [16]. The multimodal attention mechanism integrates information from different modalities, such as RGB images, optical flow images, and depth images, to improve action recognition accuracy. The CoViAR model proposed by Liu et al. significantly enhances performance by fusing features of RGB images and optical flow images and applying the attention mechanism on top of them [17]. The advantages of attention-based models include the ability to dynamically adjust the model's focus, improving recognition accuracy and better handling of complex scenes and long-duration videos. However, the implementation and training processes of these models are complex and require substantial data and computational resources.

In the future, models based on the attention mechanism may have better development prospects. The attention mechanism can dynamically adjust the focus of the model, enhancing its accuracy and robustness in recognition tasks. Models that combine multimodal information with attention mechanisms are expected to further improve the performance and application scope of video action recognition. Additionally, research into lightweight models and self-supervised learning will provide new development directions for video action recognition, reducing the demand for computational resources and enabling real-time applications.

4. Conclusion

This paper reviews the current development status, challenges, and future research directions of video action recognition techniques based on deep learning. Significant progress has been made in video action recognition technology in recent years, with traditional methods gradually being replaced by deep learning methods such as CNN, RNN, and LSTM. These methods can automatically extract features and handle temporal dependencies, significantly improving the accuracy and robustness of action recognition [1][2][3][11][12][13][14]. Models based on attention mechanisms further enhance action recognition performance by dynamically adjusting the focus of attention, making them a current research hotspot [15][16][17]. However, despite many advances, video action recognition still faces several challenges. First, high computational resources are required, and real-time processing of long-duration videos remains difficult [11][13]. Second, model training is complex and prone to gradient vanishing or explosion problems [14]. Additionally, the dataset bias issue leads to insufficient generalization ability of the model in new environments [16][17]. Finally, variations in real application scenarios, such as different camera angles, lighting changes, and occlusion, continue to challenge the robustness of the model [6][8][9].

Future research can explore the following aspects: multimodal fusion, which enhances the accuracy and robustness of action recognition by combining multimodal information such as RGB images, optical flow images, and depth images [15][17]; the development of lightweight models, which reduces computational complexity and enables real-time action recognition [3][15]; the application of self-supervised learning and cross-domain migratory learning, which utilizes unlabeled data for pre-training to improve model performance in small sample environments and enhance the generalization ability of the model [9][16]; and models that combine attention mechanisms with multimodal information to further improve the performance and application scope of video action recognition [15][16][17].

Through the review in this paper, we hope to provide researchers with a comprehensive perspective on the current state of development and future research directions of video motion recognition technology.

References

[1] Aggarwal, J., & Ryoo, M. (2011). “Human activity analysis: a review.” ACM Computing Surveys.

[2] angelov, p., & habib, z. (2017). “A Comprehensive Review on Handcrafted and Learning-Based Action Representation Approaches for Human Activity Recognition.” Applied Sciences.

[3] Bi Chunyan. A Review of Deep Learning-Based Human Action Recognition in Video. Tsinghua University Press, 2020.

[4] Cao Jinqi, Jiang Xi nghao, Sun Pongfeng. Video human action recognition algorithm based on training graph CNN features. Computer Engineering, 2017, 43(11): 234-238.

[5] girdhar, r., carreira, j., doersch, c., & z isserman, a. (2019). A Video Representation Learning Framework Using Convolutions and Transformers. cvpr.

[6] Hoch reiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. neural Computation, 9(8), 1735-1780.

[7] Huang, Q. (2021). “A review of video-based human action recognition algorithms.” Journal of Computing.

[8] Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D Convolutional Neural Networks for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1), 221-231.

[9] Kulbacki, M., Klempous, R., & Wojciechowski, K. (2023). “Intelligent Video Analytics for Human Action Recognition: the State of Knowledge.” Sensors.

[10] Liu, Z., Luo, Y., Wang, L., Wang, Y., Tai, Y., & Qiao, Y. (2018). COVIAR: Compression Video Action Recognition. cvpr.

[11] Luo, Huilan. A Review of Advances in Deep Learning-Based Human Action Recognition in Video. Peking University Press, 2021.

[12] Qi, Yanwei. Deep learning in video action recognition. Electronics and Software Engineering, 2018.

[13] Zhang Shujun, Lan Shanzhen, Bu Qi, Wang Yang. A Brief Introduction to Deep Learning Based Action Recognition Methods. School of Information and Communication Engineering, Communication University of China, 2020.

[14] simonyan, k., & z isserman, a. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. nIPS.

[15] srivastava, N., Mansimov, E., & Salakhutdinov, R. (2015). Unsupervised Learning of Video Representations using LSTMs. icml.

[16] vaswani, A., S hazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł ., & Polosukhin, I. (2017). Attention is All You Need. nIPS.

[17] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal Segment Networks: towards Good Practices for Deep Action Recognition. eccv.

[18] Zhou, Yanqing. A review of vision-based human action recognition. Journal of Shandong Light Industry Institute, 2012, 26(1): 85-92.