Human Action Recognition Methods

Subjects: Computer Science, Artificial Intelligence

Contributor: Changxuan Yang , Feng Mei , Tuo Zang , Jianfeng Tu , Nan Jiang , Lingfeng Liu

In the field of artificial intelligence, human action recognition is an important part of research in this area, making human interaction with the external environment possible. While human communication can be conveyed with words, facial expressions, written text, etc., the relationship between computers and sensors to understand human intentions and behaviour is now a popular area of research. As a result, more and more researchers are devoting their time and experience to the study of human action recognition.

action recognition ARMA attention mechanism

1. Introduction

In recent years, research on human action recognition has developed by leaps and bounds and is now used in various fields, such as video surveillance, intelligent medical care, human–machine collaboration, and intelligent human–machine interfaces [1][2][3][4]. This also means that there are increasingly higher requirements for human action recognition algorithms in terms of performance, which is a classic and challenging topic in computer vision research. To date, many methods based on hand-crafted feature representations have been widely used for action recognition due to their advantages, such as simplicity and robustness [5][6][7]. However, due to the limitations of human cognitive abilities, the method is often database-oriented and difficult to apply to real-life scenarios.

With the development of deep learning techniques, deep learning algorithms have more advantages in the field of human motion recognition than traditional methods [8]. Currently, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are frequently used in the field of human motion recognition. The 3D CNN [9] is a typical algorithm studied in human action recognition tasks. In that work, 3D convolutions are employed to extract features from the spatial and temporal dimensions of video data. This works well for capturing spatial information and has a better performance in image recognition at the moment, but temporal information is inevitably lost when sequences are encoded into images, and temporal motion plays a key role in human action recognition. This problem can be mitigated with RNNs, in particular long short-term memory (LSTM), which has been shown to effectively model long-term cues of motion sequences [10]. The gate unit in LSTM can choose whether to update specific information or not while ensuring long-term memory of valid data and forgetting or discarding useless information, thereby maximizing the utilization of data information.

2. Traditional Machine Learning and Hand-Crafted Feature-Based Action Recognition

In traditional recognition methods based on machine learning and manual features, hand-crafted feature extractors and action classifiers based on traditional machine learning algorithms are often used [11]. Action classifiers are used to recognise and classify human movement actions based on the characteristics of that action. For example, Cho et al. [12] used joint distance features for feature extraction. The category of each pose is labelled by an artificial neural network (ANN). Finally, discrete Hidden Markov Models (HMMs) are applied to classify and recognise action sequences. Meanwhile, in order to effectively improve the recognition performance of the system, some researchers have adopted a key-frame-based approach to reduce the processing time of the system [13][14]. A recognition system for human action sequences was developed using traditional machine learning algorithms combined with key-frame selection. In past research, action recognition methods based on traditional machine learning and manual features were combined with great success. However, for the construction and extraction of features [15], they need to rely on human cognition. Moreover, based on human expertise, only superficial features can be learned, making it difficult to cope with the needs of real environments.

3. Deep Learning-Based Action Recognition

In recent years, a number of new methods have been developed, especially regarding the application of deep learning methods in action recognition [16]. The main representative works can be summarized as discussion methods based on convolutional neural networks and discussion methods based on LSTM.

Traditional CNN models are currently limited to processing 2D inputs and are not suitable for the feature capture of 3D skeleton data. To shift CNNs from images to temporal motion sequences, Tran et al. [17] extended traditional CNNs to 3D CNNs, which are more suitable for spatio-temporal feature learning. Related experiments have shown that this scheme outperforms traditional 2D CNNs in terms of analytical functionality. Another common strategy is to employ two-stream CNNs to deal with the problem of capturing motion information between consecutive frames. Zhu et al. [18] proposed a CNN architecture based on a two-stream approach that implicitly captures motion information between adjacent frames and uses an end-to-end CNN approach to learn optical streams. Task-specific motion representations can be obtained while avoiding expensive computation and storage. Since then, many improved models have been proposed, and the two-stream CNN has made significant contributions to the development of motor action recognition [19]. It can even be referenced to realistic and complex real-world environments; for example, Hu et al. [20] introduced a video triple model to obtain additional timestamp information, thus extending behaviour recognition to workflow recognition. Moreover, with extensive simulation experiments, it was shown that the algorithm is robust and efficient in the recognition of real environments.

However, these algorithms have been shown to be only effective for short-term temporal feature learning and are not applicable to long-term temporal feature encoding. With the development of RNNs, LSTM networks suitable for long-term motion sequences have been developed. They have been gradually applied to human action recognition, demonstrating their ability to effectively alleviate the recognition problem of long-term motion sequences [21][22]. Wang et al. [23] introduced long short-term memory (LSTM) to model the high-level temporal features generated by a kinetically pretrained 3D CNN model, with satisfactory results in the recognition and classification of long-term motion sequences. However, the traditional frame-skipping pattern of LSTM [24] also limits performance in action

recognition. The problem of data redundancy accompanies the task of the recognition of long-term motion action data.

4. Action Recognition Based on Joint-Aware and Attention Mechanisms

In recent years, many researchers have turned their attention to joint-aware and attention mechanisms and have achieved good recognition performance in long-term temporal reasoning tasks. Regarding joint-aware-based recognition methods, Oikonomou et al. [25] argue that each action in real life can be effectively perceived by observing only a specific set of joints and associate a specific joint with each action to point out the joint that contributes the most; Shah et al. [26] separately extracted the motion features of each joint using a motion encoder and then performed collective reasoning and selected the most discriminative joint for the recognition task. Regarding recognition methods based on attention mechanisms, Dai et al. [24] proposed an LSTM network based on end-to-end two-stream attention, which can selectively focus on the effective features of the original input image and give different levels of attention to the output of each depth feature map to effectively improve the recognition performance of the model by adopting a visual attention mechanism to address the problem that features of different frames have different learning roles; Li et al. [27] proposed a spatio-temporal attention (STA) network to learn discriminative feature representations of actions by representing useful information at the frame level and channel level, which can be inserted into state-of-the-art 3D CNN architectures for video action detection and recognition with better recognition performance; in the article [28], the authors proposed a bi-directional long short-term memory (BiLSTM)-based attention mechanism. The attention mechanism is used to improve performance and extract additional high-level selective action-related patterns and cues, thereby obtaining a high-performance recognition model.

References

1. Antonik, P.; Marsal, N.; Brunner, D.; Rontani, D. Human action recognition with a large-scale brain-inspired photonic computer. Nat. Mach. Intell. 2019, 1, 530–537.
2. Kwon, Y.; Kang, K.; Bae, C. Unsupervised learning for human activity recognition using smartphone sensors. Expert Syst. Appl. 2014, 41, 6067–6074.
3. Wang, P.; Liu, H.; Wang, L.; Gao, R.X. Deep learning-based human motion recognition for predictive context-aware human-robot collaboration. CIRP Ann. 2018, 67, 17–20.
4. Barnachon, M.; Bouakaz, S.; Boufama, B.; Guillou, E. Ongoing human action recognition with motion capture. Pattern Recognit. 2014, 47, 238–247.
5. Xia, L.; Chen, C.C.; Aggarwal, J.K. View invariant human action recognition using histograms of 3d joints. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision

and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 20–27.

6. Mozafari, K.; Moghadam Charkari, N.; Shayegh Boroujeni, H.; Behrouzifar, M. A novel fuzzy hmm approach for human action recognition in video. In Proceedings of the Knowledge Technology Week, Kajang, Malaysia, 18–22 July 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 184–193.
7. Li, X.; Zhang, Y.; Liao, D. Mining key skeleton poses with latent svm for action recognition. Appl. Comput. Intell. Soft Comput. 2017, 2017, 5861435.
8. Kansizoglou, I.; Bampis, L.; Gasteratos, A. Deep feature space: A geometrical perspective. IEEE Trans. Pattern Anal. Mach. Intell. 2021, 44, 6823–6838.
9. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2012, 35, 221–231.
10. Tang, P.; Wang, H.; Kwong, S. Deep sequential fusion LSTM network for image description. Neurocomputing 2018, 312, 154–164.
11. Ramasamy Ramamurthy, S.; Roy, N. Recent trends in machine learning for human activity recognition—A survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2018, 8, e1254.
12. Wang, Y.; Sun, S.; Ding, X. A self-adaptive weighted affinity propagation clustering for key frames extraction on human action recognition. J. Vis. Commun. Image Represent. 2015, 33, 193–202.
13. Gharahbagh, A.A.; Hajihashemi, V.; Ferreira, M.C.; Machado, J.J.; Tavares, J.M.R. Best Frame Selection to Enhance Training Step Efficiency in Video-Based Human Action Recognition. Appl. Sci. 2022, 12, 1830.
14. Cho, T.Z.W.; Win, M.T.; Win, A. Human Action Recognition System based on Skeleton Data. In Proceedings of the 2018 IEEE International Conference on Agents (ICA), Salt Lake City, UT, USA, 18–22 June 2018; pp. 93–98.
15. Wang, J.; Chen, Y.; Hao, S.; Peng, X.; Hu, L. Deep learning for sensor-based activity recognition: A survey. Pattern Recognit. Lett. 2019, 119, 3–11.
16. Zhang, H.B.; Zhang, Y.X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.X.; Chen, D.S. A comprehensive survey of vision-based human action recognition methods. Sensors 2019, 19, 1005.
17. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 4489–4497.
18. Zhu, Y.; Lan, Z.; Newsam, S.; Hauptmann, A. Hidden two-stream convolutional networks for action recognition. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 363–378.

19. Sarabu, A.; Santra, A.K. Distinct two-stream convolutional networks for human action recognition in videos using segment-based temporal modeling. Data 2020, 5, 104.
20. Hu, H.; Cheng, K.; Li, Z.; Chen, J.; Hu, H. Workflow recognition with structured two-stream convolutional networks. Pattern Recognit. Lett. 2020, 130, 267–274.
21. Meng, B.; Liu, X.; Wang, X. Human action recognition based on quaternion spatial-temporal convolutional neural network and LSTM in RGB videos. Multimed. Tools Appl. 2018, 77, 26901–26918.
22. Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
23. Wang, X.; Miao, Z.; Zhang, R.; Hao, S. I3d-lstm: A new model for human action recognition. In Proceedings of the IOP Conference Series: Materials Science and Engineering, Kazimierz Dolny, Poland, 21–23 November 2019; IOP Publishing: Bristol, UK, 2019; Volume 569, p. 032035.
24. Dai, C.; Liu, X.; Lai, J. Human action recognition using two-stream attention based LSTM networks. Appl. Soft Comput. 2020, 86, 105820.
25. Oikonomou, K.M.; Kansizoglou, I.; Manaveli, P.; Grekidis, A.; Menychtas, D.; Aggelousis, N.; Sirakoulis, G.C.; Gasteratos, A. Joint-Aware Action Recognition for Ambient Assisted Living. In Proceedings of the 2022 IEEE International Conference on Imaging Systems and Techniques (IST), Kaohsiung, Taiwan, 21–23 June 2022; pp. 1–6.
26. Shah, A.; Mishra, S.; Bansal, A.; Chen, J.C.; Chellappa, R.; Shrivastava, A. Pose and joint-aware action recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 3850–3860.
27. Li, J.; Liu, X.; Zhang, W.; Zhang, M.; Song, J.; Sebe, N. Spatio-temporal attention networks for action recognition and detection. IEEE Trans. Multimed. 2020, 22, 2990–3001.
28. Muhammad, K.; Ullah, A.; Imran, A.S.; Sajjad, M.; Kiran, M.S.; Sannino, G.; de Albuquerque, V.H.C. Human action recognition using attention based LSTM network with dilated CNN features. Future Gener. Comput. Syst. 2021, 125, 820–830.