Journal of Frontiers of Society, Science and Technology
Clausius Scientific Press, Canada

DOI: 10.23977/jfsst.2021.010204
Volume 1, Number 2, 2021

Action Recognition Based on Kinect Deep Learning

Wenlu Yang[1], Ying Peng[2], Hong Xie[3]

1. College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

2.College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

3.College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

Keywords: Action recognition, Kinect, Deep learning

Abstract: With the development of science and technology, human motion recognition technology is widely used in rehabilitation medicine, artificial intelligence, somatosensory games and many other fields. More and more scholars have carried out in-depth research on how to improve the accuracy of human motion recognition.This paper systematically reviews the methods of human motion recognition based on Kinect, summarizes the accuracy of these methods, compares the advantages and disadvantages, summarizes the performance of each method, and provides reasonable suggestions for researchers with different needs.

1. Introduction

Human action recognition is an important part of computer vision [1], involving image processing, machine learning, artificial intelligence and other disciplines ; It can be divided into two categories : traditional method and deep learning method. The traditional method is divided into three categories: template matching, spatio-temporal interest points and trajectory. The deep learning method is divided into three categories: three-dimensional convolution neural network, time convolution network and two-stream convolution neural network [2].

With the rapid development of modern technology, the birth of Kinect has pushed action recognition to a new peak. Kinect[3] is a non-contact somatosensory device launched by Microsoft, which is mainly composed of infrared transmitters, RGB cameras and infrared receivers. Compared with traditional somatosensory devices, Kinect is cheaper and smaller[4], and is widely used in human action recognition. In recent years, the research methods of human action recognition based on Kinect are increasing. This paper classifies and summarizes these methods to provide reference for subsequent researchers.

2. Common Data Sets

The recognition method based on Kinect depth learning mainly uses the RGB camera of Kinect to transform the distance between the human body and the camera into the depth of the image color to represent the depth information. Combined with the network model, the accuracy is verified on the action recognition data set. The commonly used data set is shown in Table 1.

Table 1 Commonly Used Data Sets

Data set	Time	category	Number of samples	Action
KTH	2004	6	599	walking, jogging, running, boxing, waving, applause
Weizmann	2005	9	93	bend, jump, run, etc
Hollywood	2008	8	663	people ' s expressions, postures, and clothes in the film
Hollywood 2	2009	12	3669	people ' s expressions, postures, and clothes in the film
UCF 101	2012	101	13320	climbing, playing, haircut, etc.
HMDB 51	2011	51	6849	body, interactive movements
NTU RGB+D	2016	60	56880	RGB video, depth map sequence,
Sports 1M	2014	487	1133158	Various sports activities
Kinetics	2017	400	266000	Interaction between people and things

3. Action Recognition Method Based on Deep Learning

Deep learning is an unsupervised learning method, which can effectively increase the human action in the video image from two-dimensional space to three-dimensional space[5]. According to the different structure of deep learning network, it can be divided into three directions : action recognition based on two-stream convolution network, action recognition based on three-dimensional convolution network and action recognition based on long-short-term memory network.

3.1 Action Recognition Based on Long-Short Memory Network

Long short term memory network(LSTM) has good performance in temporal data correlation and dynamic modeling. But the original LSTM is difficult to grasp the dynamic of the whole sequence data. To this end, Reference [8] proposed a fusion model, which extracted the skeleton features of each step of LSTM from different time periods, and made full use of the skeleton data of multi-stream LSTM for action recognition. The feasibility was verified on NTU RGB + D dataset and SBU interactive dataset.[9] proposed an action recognition algorithm combining skeleton geometric features with LSTM network, and verified its average recognition accuracy on SBU interactive dataset and UT Kinect dataset. The network model is shown in Fig.1 :

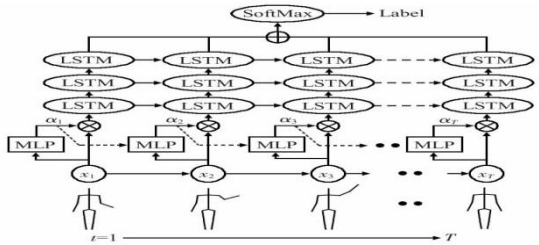

Fig.1 Lstm Network Model Structure[9]

3.2 Action Recognition Based on 3d Convolution Network

Convolutional neural network(CNN) is a deep model, but it is limited to processing 2D models. In order to break this limitation, a new 3D CNN is proposed for action recognition for the first time in literature[10]. The model can extract time and space dimensions at the same time, and verify its average recognition accuracy on KTH dataset. In order to effectively extract visual and temporal information from videos, Reference [11]proposed an improved deep network model, which used Softmax classifier to identify and classify, and verified its average recognition accuracy on KTH dataset. The network model is shown in Fig. 2.

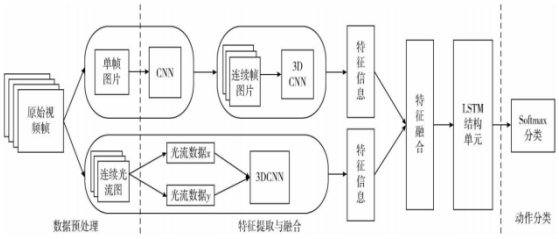

Fig.2 Structure Diagram of 3d Cnn Network Model[11]

3.3 Action Recognition Based on Dual-Flow Convolution Network

The dual-flow convolutional neural network uses optical flow diagram and RGB diagram as time and spatial information, which improves the integrity of the network. Reference[12]proposed an improved FasterR-CNN algorithm using K-means algorithm to cluster objects in the image. Reference[13]obtained the RGB image and optical flow image of the video by sparse sampling, and sent them to the VGG-16 network to extract the spatio-temporal characteristics of the video. The mid-level spatio-temporal fusion feature was sent to the C3D CNN to identify the category of the action, and its accuracy was verified in the HMDB51 and UCF101 data sets, as shown in Fig. 3 :

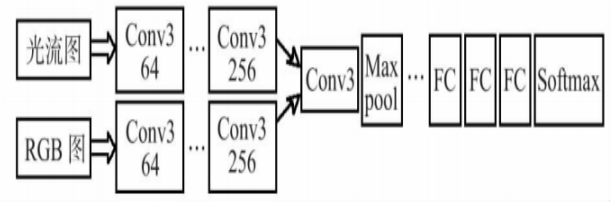

Fig.3 Space-Time Dual-Flow Network[13]

3.4 Method Summary

At present, the types of public data sets are constantly improved, and the action recognition method based on Kinect deep learning is also gradually developed. Good progress has been made in obtaining long-term time information, preprocessing input images, and real-time recognition of human actions. Table 2 summarizes the research methods with high recognition accuracy in recent years.

Table 2 Comparison of Recognition Rate of Each Method

method data set	NTU RGB+D	SBU	UT KINECT	KTH	UCF101	HMDB51
Multi-Stream LSTM[8]	83.81%	92.50%				
WU et al.[9]		99.75%	98.79%			
3D CNN[10]				90.20%		
SHEN et al.[11]				93.89%		
DS-FasterRCNN[12]					92.80%	
C3D CNN[13]					94.80%	69.50%
Jeyanthi et al.[14]					92.00%	91.00%
LTC-CNN[15]					92.70%	67.20%
ST-LSTM[16]		93.30%	95.00%			
GCA-LSTM[17]		94.10%	98.50%			
C2-LSTM[18]					92.80%	61.39%

4. Summary and Outlook

This paper mainly introduces the research of action recognition based on Kinect deep learning. In general, the birth of Kinect somatosensory equipment has greatly improved the accuracy of human action recognition, but there are still many difficulties to be solved in the future :

(1)Action recognition in complex environment With the development of human-computer interaction technology, action recognition in the early simple environment has been unable to meet people ' s needs. The environment of human action recognition is transforming from simple to complex. However, the accuracy of action recognition is affected by illumination, noise and shading in complex scenes, which brings new challenges to the research of human action recognition.

(2) Human differences At present, the application of human action recognition is more and more widely, which means that human action recognition is not limited to early single action recognition. However, due to the obvious differences in people ' s age, height, weight and other aspects, making the same action, different people ' s performance will be very different, which adds difficulty to multi-person action recognition.

(3) Algorithm optimization In order to improve the accuracy and robustness of action recognition, researchers have proposed many algorithms for action recognition, and achieved good progress, but still did not achieve accurate recognition. Therefore, in order to improve the accuracy of action recognition, it is still necessary to continuously optimize the algorithm in the future, reduce the complexity of the algorithm and improve the computational efficiency.

(4) Data set labeling At present, the number of behavior categories, samples and videos of public datasets is increasing, and the scene is becoming more and more complex, which ensures the robustness and stability of the algorithm to a certain extent. However, a large number of public datasets lack good labeled datasets, which brings greater difficulty to improve the accuracy of human action recognition.

In summary, researchers have proposed many methods to improve the accuracy of human action recognition, but the complex action recognition in complex environments has not achieved ideal results. As an important subject in the field of computer vision, human action recognition still faces many difficulties. The real-time performance and spatial difference of action recognition are still the difficulties for future research.

5. Acknowledgment

National Natural Science Foundation of China (No. 61550110252).

References

[1] Pichao Wang,Wanqing Li,Chuankun Li,et al. Action recognition based on joint trajectory maps with convolutional neural networks[J]. Knowledge-Based Systems,2018.

[2] Yang Zhengyuan,Li Yuncheng,Yang Jianchao,et al. Action recognition with visual attention on skeleton images[C]/ /Proc of the 24th Interna- tional Conference on Pattern Recognition Piscataway,NJ: IEEE Press,2018: 3309-3314.

[3] Wasenmüller O,Stricker D. Comparison of kinect v1 and v2 depth ima- ges in terms of accuracy and precision[C]/ /Proc of Asian Conference on Computer Vision.Berlin: Springer,2016: 34-45.

[4] Anonymous. New All-in-One Mount for Xbox 360 Kinect Sensor,[J]. Multimedia Publisher,2011,22(7).

[5] Tsai Jen Kai,Hsu Chen Chien,Wang Wei Yen,Huang Shao Kang. Deep Learning-Based Real-Time Multiple-Person Action Recognition System.[J].Sensors (Basel, Switzerland),2020,20(17).

[6] Nesrine Grati,Achraf Ben-Hamadou,Mohamed Hammami. Learning local representations for scalable RGB-D face recognition[J]. Expert Systems With Applications,2020,150.

[7] Saranya Rajan,Poongodi Chenniappan,Somasundaram Devaraj, et al. Novel deep learning model for facial expression recognition based on maximum boosted CNN and LSTM[J]. IET Image Processing,2020,14(7).

[8] Lei Wang;Xu Zhao;Yuncai Liu. Skeleton Feature Fusion Based on Multi-Stream LSTM for Action Recognition[J].

IEEE ACCESS, 2018, Vol.6(0):50788-50800.

[9] Wu Qian, Wu Fei, Luo Lizhi. Motion recognition algorithm based on geometric features combined with LSTM network [J]. Sensors and microsystems, 2020,39 (10) : 111-114.

[10] Ji Shuiwang,Yang Ming,Yu Kai. 3D convolutional neural networks for human action recognition.[J]. IEEE transactions on pattern analysis and machine intelligence,2013,35(1).

[11] Shen Xiting, Yu Sheng, Dong Yao, etc. Human action recognition method based on deep learning [J]. Computer engineering and design, 2020,41 (04) : 1153-1157.

[12] Liu Yabin,Yu Jun,Hu Zhiyi. Improved Faster R-CNN Algorithm for Sea Object Detection Under Complex Sea Conditions[J]. International Journal of Advanced Network, Monitoring and Controls,2020,5(2).

[13] Wang Qian, Sun Xiankun, Fan Dongyan. Spatiotemporal feature fusion based on deep learning for human action recognition [J]. Sensors and microsystems, 2020,39 (10) : 35-38.

[14] A. Jeyanthi Suresh;J. Visumathi. Inception ResNet deep transfer learning model for human action recognition using LSTM[J]. Materials Today: Proceedings, 2020, (0)

[15] Varol Gul,Laptev Ivan,Schmid,et al.Long-Term Temporal Convolutions for Action Recognition.[J].IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, Vol.40(6):1510-1517.

[16] Liu Jun,Shahroudy, Amir, et al. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition[J]. COMPUTER VISION - ECCV 2016, PT III, 2016, Vol.9907(0):816-833.

[17] LIU J,WANG G,HU P,et al.Global context-aware attention LSTM networks for 3D action recognition[C]//2017 IEEE Con- ference on Computer Vision and Pattern Recognition (CVPR),IEEE Computer Society,2017.

[18] Mahshid Majd, Reza Safabakhsh. Correlational Convolutional LSTM for human action recognition[J]. Neurocomputing, 2020, 396.