Rehabilitation recognition skeleton data depth learning based on RNN

Qingzhi Zhang*, *Panfeng* Wu, *Xiaohui* Du, *Hualiang* Sun, and *Lijia* Yu

Shandong Institute of Space Electronics Technology, China

Abstract. With the extensive application of deep learning in the field of human rehabilitation, skeleton based rehabilitation recognition is becoming more and more concerned with large-scale bone data sets. The key factor of this task is the two intra frame representations of the combined co-and the inter-frame. In this paper, an inter frame representation method based on RNN is proposed. Pointtion of each joint is joint-coded they are assembled into semantic both spatial and temporal domains.we introduce a global spatial aggregation which is able to learn superior joint co features over local aggregation.

1 Introduction

Human rehabilitation recognition is the basic scenario of AI application. The description of human actions to joints is the best description of human rehabilitation. Firstly, compared with other optical flow and RGB, the data amount is very small, and the target is clear. Secondly, the joint data has the best robustness to the interference of the background noise. These two characteristics make skeleton based models have the inherent advantages of lightweight and high recognition rate.

In this paper, we focus on the problem of Workflow for Rehabilitation Recognition Skeleton (Figure 1).

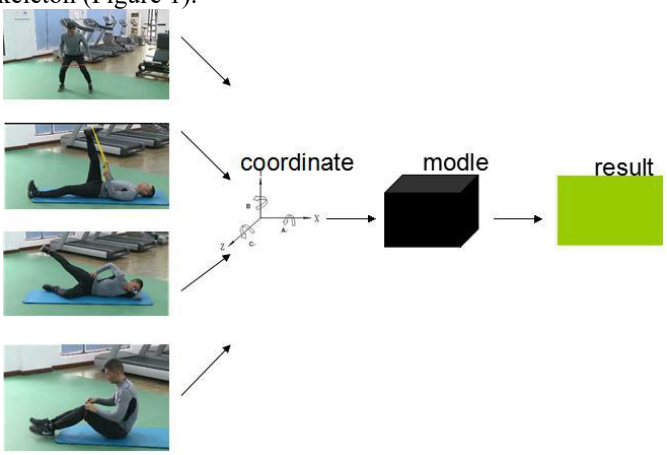

Fig. 1. Workflow for Rehabilitation Recognition Skeleton.

* Corresponding author: Qzzhang28@gmail.com

2 Related Work

The RNN network is designed to address the entire process simulation of exercise rehabilitation. The RNN network is a natural choice that has been used on a large scale to complete deep learning from the skeleton sequence. With the widespread use of deep learning, more and more literature uses RNN to learn skeleton features and complete large-scale citations in various scenarios.

3 Methods

3.1 Co-occurrence Feature Learning with RNN

RNN is one of the most powerful and successful neural network models, which has been widely used in image classification, object detection, video classification and so on. Compared with RNN and other sequence structures, RNN can utilize historical information, take action sequence into account, and use time and space to encode asynchronously. Using the convolution characteristics of RNN, the human body's rehabilitation action is decomposed into two steps .that is, the action characteristics of the cross space domain, including the motion position and the range of motion, and the motion feature fitting in the Ag and the whole motion process. Finally, the results obtained from soft max can give the accuracy of the rehabilitation result to the posture and range of movement. It indicates that T is a D1 D2 D3 3D tensor flow. In the convolution process, the array values of motion posture and range of motion can be transposed to meet different rehabilitation needs. All tensors of dimension DI can be aggregated in whole process. If the key actions are specified in the action process, the two preceding and subsequent actions of the key actions can be used as key convolution frames.

3.2 Explicit Skeleton Motion

In addition to the co-occurrence of movement, a group of multiple time movement of joints is the key frame of rehabilitation action. Therefore, in human rehabilitation, skeleton action is taken as the key frame and the convolution network of RNN is introduced.

We formulate it as Kt ={D1t; D2t; : : : ; DNt}, where N is the number of joint and D = (x; y; z) is a 3D joint coordinate. The skeleton motion is defined as the temporal difference of each joint between two consecutive frames:

$$Zt = Kt+1-Kt$$
$$=\{D1t+1-D1t; D2t+1\_D2t; :::; DNt+1-DNt\}:$$

The original skeleton coordinate K and skeleton motion D are sent into convolution respectively by spatial coordinates and time coordinates. In the subsequent calculation, the two are combined to merge.

	Temporal difference		Point-level feature learning
conv1	1x1x64	1x1x64	
conv2	3x1x32	3x1x32	
	(0,2,1)	(0,2,1)	
conv3	3x3x32, /2	3x3x32, /2	
conv4	3x3x64, /2	3x3x64, /2	
			Co-occurrence feature learning
	Concat		
conv5	3x3x128, /2		
conv6	3x3x256, /2		
	Flatten		Convolution
fc7	256		Transpose
			FC
fc8	class		

Fig. 2. Overview of the proposed hierarchical co-occurrence.

3.3 Hierarchical Co-occurrence Network

In this section, the detailed Rehabilitation recognition base on RNN will be described. The network architecture is shown in Figure 3. The joint tensor sequence X can be represented by Z G W tensor, Z represents the number of frames, G represents the number of joints in the skeleton, W represents coordinates. Skeletal movement is involved in network feedback in the same way as X. They are as the two input traffic. The two networks interact in the same system and their parameters. Their characteristics are data fusion after unified convolution. After a given state and motion requirement of a given motion skeleton, hierarchical learning is performed. In the first calculation, the point level features are computed by convolution with 1 (1) and 1 (2). The size of the body is fixed, so the independent three-dimensional coordinate system of each joint is fixed, and the feature of point level is added to this joint coordinate system. After that, we use RNN to transform the 3D coordinates of the joints into the moving channels. In the second calculation, all the global features will be classified by two fully linked layers of recovery action and rehabilitation time.

In the three converged calculations, the best accuracy occurs at the most operational time. The results of the test are as shown in Table 1.

Table 1. Performance of different fusion methods for multi-person feature on the NTU RGB+D dataset in the cross-subject setting.

Method		Accuracy (%)
Early fusion		84.7
Late fusion	Mean	88.3
	Concat	88.4
	Max	88.9

4 Experiments

The NTU RGB+D data set is so far the largest skeleton-based human action recognition dataset. It contains 56880 skeleton sequences, which are annotated as one of 60 action classes.

There are two recommended evaluation protocols, i.e. Cross Subject (CS) and Cross-View (CV).In the training process, according to the international standard method, the random subsequence is extracted, and the extraction rate is uniformly distributed in [0.5, 1]. In the calculation process, the ratio of skeletal sequence is 0.9. The time of the sample is different. We normalize the sample and interpolate the sample with insufficient time. The total input training model of the algorithm is 100k, each sequence is 1024, the initial learning ability is 0.0001, the exponential decay is 0.999 per 10K step.

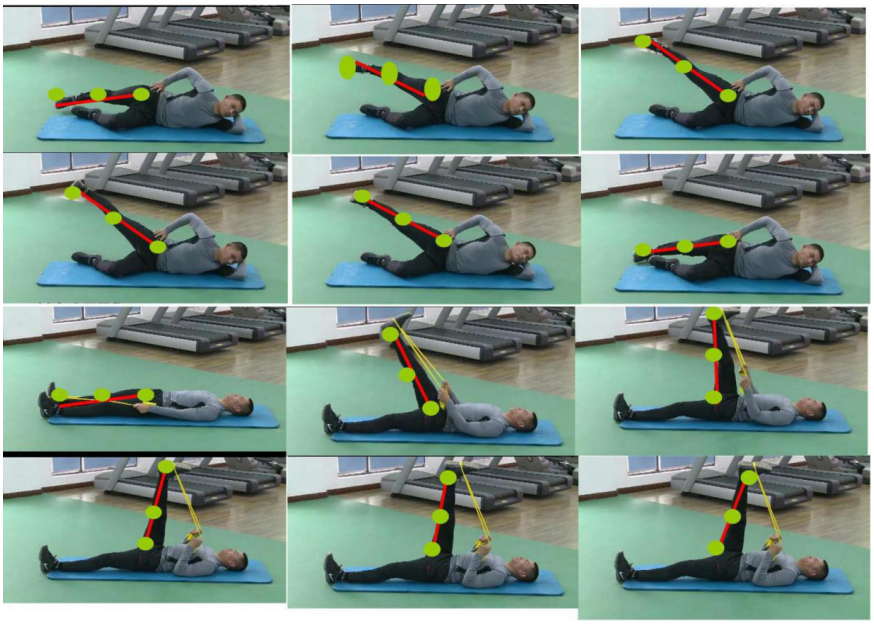

Fig. 3. Exemplary action recognition results.

5 Conclusions

We present a deep learning framework for skeleton rehabilitation based on RNN for human rehabilitation. By using the convolution network of RNN, a three-dimensional coordinate

based on the body skeleton is established between the confusions and the time of rehabilitation, and then the characteristics of the current skeleton and the time of motion are fused. The experimental results show that it can obviously improve the recognition of rehabilitation actions and improve the accuracy of evaluation of rehabilitation actions.

References

1. [Du et al., 2017] Chao Li, Qiaoyong Zhong, Di Xie, Shiliang Pu. Co-occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation, pages 579–583, 2016.
2. [Girshick et al., 2014] Ross B Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierar-chies for accurate object detection and semantic segmentation. CVPR, pages 580–587, 2014.
3. [He et al., 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recog-nition. In CVPR, pages 770–778, 2016.
4. [Hussein et al., 2013] Mohamed E. Hussein, Marwan Torki, Mohammad A. Gowayyed, and Motaz El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In IJCAI, pages 639–44, 2013.
5. [Ji et al., 2014] Yanli Ji, Guo Ye, and Hong Cheng. Inter-active body part contrast mining for human interaction recognition. In ICMEW, pages 1–6, 2014.
6. [Jin and Choi, 2012] Sou Young Jin and Ho Jin Choi. Essen-tial body-joint and atomic action detection for human ac-tivity recognition using longest common subsequence algorithm. In ICCV, pages 148–159, 2012.
7. [Ke et al., 2017] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A New Representation of Skeleton Sequences for 3D Action Recognition. In CVPR, July 2017.
8. [Kingma and Ba, 2015] Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. ICLR, 2015.
9. [Li et al., 2016] Yanghao Li, Cuiling Lan, Junliang Xing, Wenjun Zeng, Chunfeng Yuan, and Jiaying Liu. On-line human action detection using joint classification-regression recurrent neural networks. ECCV, pages 203– 220, 2016.
10. [Li et al., 2017a] Bo Li, Huahui Chen, Yucheng Chen, Yuchao Dai, and Mingyi He. Skeleton boxes: Solving skeleton based action detection with a single deep convolutional neural network. In ICMEW, pages 613–616, July 2017.
11. [Li et al., 2017b] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with con-volutional neural networks. In ICMEW, pages 597–600, July 2017.
12. [Liu et al., 2016] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In ECCV, pages 816–833, 2016.
13. [Liu et al., 2017] Chunhui Liu, Yueyu Hu, Yanghao Li, Si-jie Song, and Jiaying Liu. PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding. ACM Multimedia workshop, 2017.