Decoupled Representation Network for Skeleton-Based Hand Gesture Recognition

Zhaochao Zhong[1], Yangke Li[2(✉)], and Jifang Yang[3]

[1] Xuzhou Xinzhi Science and Technology Co., Ltd., Xuzhou, China
[2] Xi'an Jiaotong University, Xi'an, China
liyangke@stu.xjtu.edu.cn
[3] Nanjing LES Information Technology Co., Ltd., Nanjing, China

Abstract. Skeleton-based dynamic hand gesture recognition plays an increasing role in the human-computer interaction field. It is well known that different skeleton representations will have a greater impact on the recognition results, but most methods only use the original skeleton data as input, which hinders the improvement of accuracy to a certain extent. In this paper, we propose a novel decoupled representation network (DR-Net) for skeleton-based dynamic hand gesture recognition, which consists of temporal perception branch and spatial perception branch. For the former, it uses the temporal representation encoder to extract short-term motion features and long-term motion features, which can effectively reflect contextual information of skeleton sequences. Besides, we also design the temporal fusion module (TFM) to capture multi-scale temporal features. For the latter, we use the spatial representation encoder to extract spatial low-frequency features and spatial high-frequency features. Besides, we also design the spatial fusion module (SFM) to enhance important spatial features. Experimental results and ablation studies on two benchmark datasets demonstrate that our proposed DR-Net is competitive with the state-of-the-art methods.

Keywords: Dynamic hand gesture recognition · Skeleton representation · Feature fusion

1 Introduction

In recent years, the human-computer interaction field has owned increasing diverse interaction methods, including speech recognition [16], hand gesture recognition [3] and touch recognition [20]. Hand gesture, as the second mainstream human communication method, provides solutions for interactive environments that use non-touch interfaces. Currently, it has been applied in many application fields, such as virtual reality systems [6], somatosensory games [22], sign language communication [25], and so on. Meanwhile, the increasing demand for intuitive interaction promotes research on hand gesture recognition.

E. Pimenidis et al. (Eds.): ICANN 2022, LNCS 13530, pp. 469–480, 2022.
https://doi.org/10.1007/978-3-031-15931-2_39

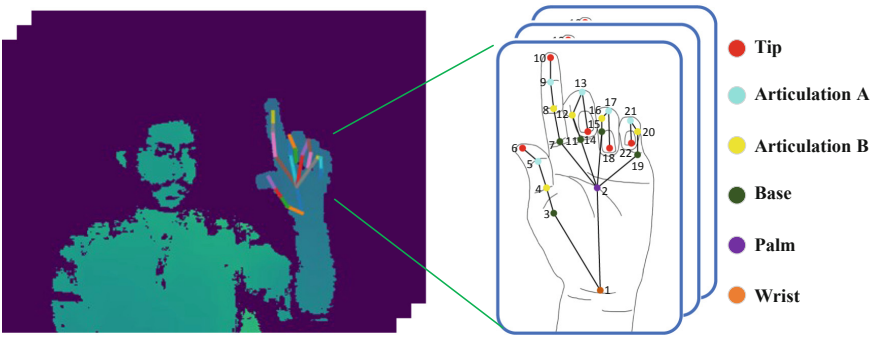

Fig. 1. Captured depth images and hand skeleton images. Each hand skeleton consists of 22 joints, including: one joint for the center of the palm, one joint for the position of the wrist and four joints for each finger.

At present, the related research on hand gesture recognition can be divided into two categories: static hand gesture recognition and dynamic hand gesture recognition. The former is mainly to recognize hand gestures from a single image, while the latter has more extensive application value, which is mainly to understand the information conveyed by hand gesture sequences. In recent years, low-cost depth sensors can capture the hand pose with reasonably good quality and provide the precise 3D hand skeleton. As shown in Fig. 1, we present the depth images and hand skeleton images captured by Intel RealSense. Compared to original RGB images, hand skeleton data can provide more intuitive information, and it is more robust to varying lighting conditions and occlusions. Therefore, skeleton-based dynamic hand gesture recognition has gradually become a current research hotspot.

The hand is an object with complex topology and has no fixed variation period, which makes skeleton-based dynamic hand gesture recognition still a challenging topic. Original skeleton data can not effectively reflect temporal motion features and spatial structure features. Meanwhile, most methods do not make full use of multi-scale features to provide the discriminative basis for hand gesture recognition. To solve these problems, we propose a novel DR-Net to realize dynamic hand gesture recognition. On the one hand, it uses the temporal representation encoder to obtain short-term motion features and long-term motion features, which have lower intra-class variance and higher inter-class variance. Meanwhile, it also introduces the TFM to perceive multi-scale temporal features, which can effectively reduce time dependency. On the other hand, it uses the spatial representation encoder to obtain low-frequency spatial features and high-frequency spatial features, which reduces the impact of location-viewpoint variation on the recognition result. Besides, it uses the SFM to enhance important spatial features. The DR-Net uses the cross-entropy loss function as the loss term, which effectively improves the recognition results.

In summary, our main contributions are summarized as follows:

- We propose a temporal representation encoder and a spatial representation encoder to enrich original skeleton data, which makes DR-Net use short-term motion features, long-term motion features, low-frequency spatial features, and high-frequency spatial features as input sources.
- We design an efficient feature fusion module for DR-Net in the temporal and spatial domains, respectively. Specifically, our proposed TFM can effectively capture multi-scale temporal features, while SFM can effectively enhance important spatial features.
- We conduct comprehensive experiments on two public benchmark datasets to verify the effectiveness of our method. Related experimental results demonstrate that DR-Net is competitive with the state-of-the-art methods.

2 Related Works

2.1 Skeleton Representations

Different skeleton representations can have an important impact on hand gesture recognition and human action recognition. Li *et al.* [17] proposed to use the Lie group to model the skeleton representation, which can effectively describe the three-dimensional geometric relationship between joints. Jiang *et al.* [15] proposed a spatial-temporal skeleton transformation descriptor, which describes the relative transformations of skeletons, including the rotation and translation during movement. Wei *et al.* [30] proposed a novel high-order joint relative motion feature to describe the instantaneous status of the skeleton joint, which consists of the relative position, velocity, and acceleration. Caetano *et al.* [2] proposed to encode the temporal dynamics by explicitly computing the magnitude and orientation values of the skeleton joints. Liu *et al.* [19] proposed to use 3D hand posture evolution volume and 2D hand movement map to represent hand posture variations and hand movements, respectively.

2.2 Deep Neural Networks

In recent years, deep neural networks have been widely used in dynamic hand gesture recognition and achieved satisfactory results. Nguyen *et al.* [23] presented a new neural network for hand gesture recognition that learns a discriminative SPD matrix encoding the first-order and second-order statistics. Chen *et al.* [4] proposed a novel motion feature augmented network for hand gesture recognition. Guo *et al.* [10] proposed a novel spatial-based GCNs called normalized edge convolutional networks for hand gesture recognition. Nunez *et al.* [24] proposed a deep learning approach based on a combination of a convolutional neural network and a long short-term memory network for hand gesture recognition. Chen *et al.* [5] proposed a dynamic graph-based spatial-temporal attention network for skeleton-based hand gesture recognition. Hou *et al.* [11] proposed an end-to-end spatial-temporal attention residual temporal convolutional network for hand gesture recognition. Weng *et al.* [31] proposed a deformable pose traversal convolution network for dynamic hand gesture recognition.

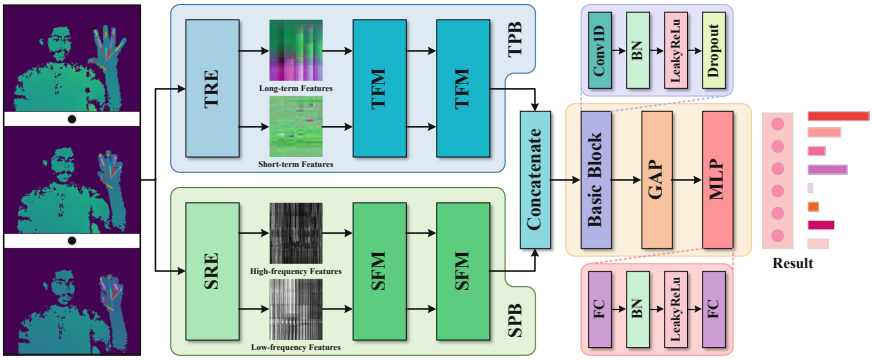

Fig. 2. The overall architecture of our proposed DR-Net. The temporal perception branch (TPB) consists of the temporal representation encoder (TRE) and the temporal fusion module (TFM). The spatial perception branch (SPB) consists of the spatial representation encoder (SRE) and the spatial fusion module (SFM).

3 Our Approach

3.1 Overview

As shown in Fig. 2, our proposed DR-Net mainly contains temporal perception branch and spatial perception branch. For the former, it uses the TRE to extract long-term motion features and short-term motion features. Besides, it uses the TFM to effectively capture and fuse multi-scale temporal features. For the latter, it uses the SRE to extract high-frequency spatial features and low-frequency spatial features. Besides, we propose the SFM to effectively enhance and fuse important spatial features. To balance the model size and recognition accuracy, the DR-Net adopts two continuous TFMs and SFMs.

3.2 Temporal Perception Branch

As we all know, dynamic hand gesture recognition not only needs to obtain the spatial information between the joints in the frame, but also needs to extract the temporal information of each joint between the frames. To solve the above problems, we propose the temporal representation encoder to process original skeleton data. In this paper, we assume the total frame number is T and the number of joints included in each frame is J. For the j-th skeleton joint of the t-th frame, in the 3D Cartesian coordinate system, it can be expressed as $S_j^t = (x_j^t, y_j^t, z_j^t)$. The set of all skeleton joints in the t-th frame of the k-th hand gesture can be expressed as: $G_k^t = \{S_1^t, S_2^t, S_3^t, \cdots, S_J^t\}$. Our proposed temporal representation encoder designs two different temporal skeleton representations as input sources. As shown in Fig. 3, short-term motion features refer to the difference between adjacent frames, while long-term motion features mean computing the difference between all other frames and the first skeleton frame.

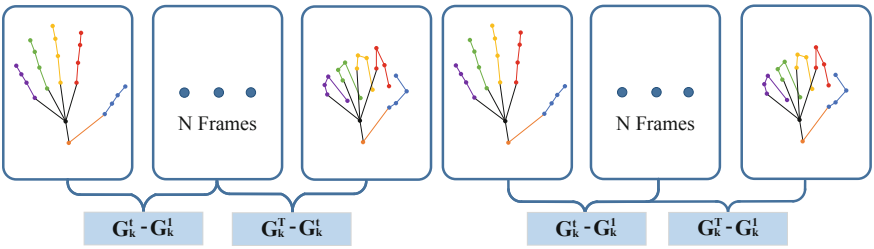

Fig. 3. Illustration of the temporal representation encoder. The left figure shows the acquisition method of short-term motion features, and the right figure shows the acquisition method of long-term motion features.

To effectively fuse short-term motion features and long-term motion features, we design the temporal fusion module, which can help DR-Net obtain multi-scale motion features. As shown in Fig. 4, the TFM has two input sources and two output sources. For two consecutive TFMs, the output of the former will be the input of the latter, and the latter will use the concatenation operation to fuse the output result. Besides, the TFM processes motion features by using different convolution kernels with different scales of receptive fields, which can tolerate a variety of temporal extents in a complex hand gesture. Specifically, we fuse short-term motion features and long-term motion features, and send them into three different branches. Their convolution kernel sizes are all 3, and the dilation rates are 1, 3, and 5, respectively. We do not use the addition method to aggregate the results of multi-scale perception, but adopt the concatenation method, which can avoid the loss of information. Finally, we use the average pooling operation to process the aggregated motion features.

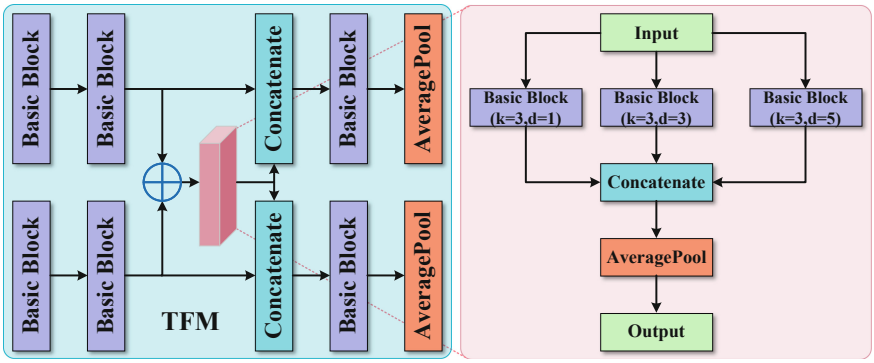

Fig. 4. The left figure shows the overall architecture of the TFM. The right figure shows the internal details of the TFM.

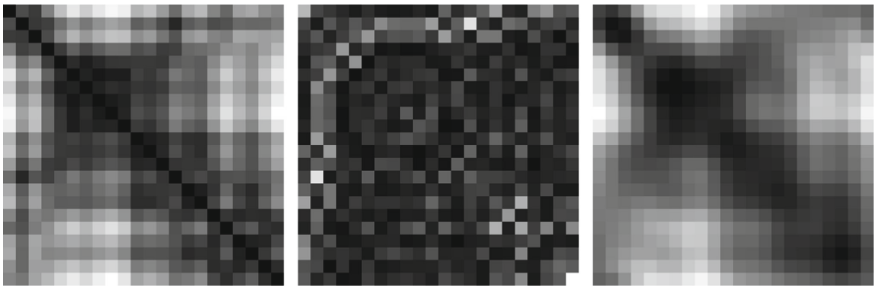

Fig. 5. Images generated by the spatial representation encoder. The left figure is the original spatial map. The middle figure is the high-frequency spatial map. The right figure is the low-frequency spatial map.

3.3 Spatial Perception Branch

To enhance the generalization ability of the model, we often need to perform extra operations such as translation, flipping, and rotation. However, the Cartesian coordinate feature is very sensitive to these data enhancement operations. Meanwhile, geometric features can fully reflect the spatial relationship of the skeleton joints. Therefore, we design an effective spatial representation encoder to extract geometric features. As shown in Fig. 5, the SRE can generate three different skeleton maps. For each skeleton frame, the original spatial map is generated by computing the Euclidean distance between any two joints, its specific calculation formula is as follows:

$$D_{i,j}^{t} = \sqrt{\left(x_i^t - x_j^t\right)^2 + \left(y_i^t - y_j^t\right)^2 + \left(z_i^t - z_j^t\right)^2} \tag{1}$$

where i and j represent the serial ID of the skeleton joints respectively. x, y, and z represent the data of the skeleton on different coordinate axes. Next, we use the fast Fourier transform to transform the spatial domain image into the frequency domain image, and use a circular filter to filter out the low-frequency information or high-frequency information. Finally, we use inverse fast Fourier transform to generate the high-frequency spatial map and low-frequency spatial map. To reduce redundant information, we only choose the upper triangular of each map to represent the geometric features of each hand skeleton.

To effectively fuse low-frequency spatial features and high-frequency spatial features, we design a spatial fusion module with the attention mechanism, which is similar to the temporal fusion module. Specifically, the SFM also has two input sources and two output sources. In the upper and lower branches, we use 1D convolutions with different dilation rates to perceive multi-scale spatial features, and use the concatenation operation to fuse them. In addition, we subtract the low-frequency features from the high-frequency features to obtain the significant difference features. Meanwhile, we take the difference feature as the input of the middle layer, and use the basic block and full connection layer to obtain the attention weight, so as to obtain the weighted features by multiplication.

4 Experiments

In this section, we evaluate our method on two public datasets: FPHA dataset and SHREC'17 Track dataset. Extensive ablation studies and comparative results show the effectiveness of our model.

4.1 Datasets

SHREC'17 Track Dataset. The SHREC'17 Track Dataset [27] is a public dynamic hand gesture dataset, which contains 2800 sequences. 28 participants perform each gesture between 1 and 10 times in two ways: using one finger and the whole hand. The depth images and hand skeletons are captured at 30 frames per second, with a resolution of 640×480. The length of hand gestures ranges from 20 to 50 frames. Each skeleton frame provides the coordinates of 22 hand joints in the 3D world space, which forms a full hand skeleton.

FPHA Dataset. The FPHA dataset [9] is a challenging 3D hand pose dataset, which provides first-person dynamic hand gestures interacting with 3D objects. The dataset contains 1,175 action videos corresponding to 45 different action categories and performed by 6 actors in 3 different scenarios. It provides the 3D coordinates of 21 hand joints except for the palm joint. We used the 1:1 setting with 600 sequences for training and 575 sequences for testing.

4.2 Implementation Details

We perform all our experiments on an NVIDIA GeForce GTX 2080Ti with Keras using the TensorFlow backend. The learning rate is initially set to be 0.001. If the loss remains unchanged after 25 iterations, we set it to 0.5 times the current learning rate. The minimum learning rate is set to be $1e^{-8}$. The batch size is set to be 64 and the network train 400 epochs. We employ the Adam algorithm with default parameters to optimize the network. Besides, we use median filtering operations to preprocess the original skeleton data and use linear interpolation to adjust the skeleton sequence with different lengths to 32 frames. To avoid over-fitting, we set the dropout rate to 0.5.

4.3 Ablation Study

Different Network Branches. To examine the influence of different network branches on hand gesture recognition accuracy, we conduct related ablation experiments according to different input sources. As shown in Table 1, the performance of the temporal perception branch is significantly better than that of the spatial perception branch. In addition, compared with long-term motion features, short-term motion features can provide a more discriminative recognition basis for the network. Meanwhile, we can get better recognition results by fusing SPB and TPB, which can achieve 96.31% and 93.21% recognition accuracy on 14 hand gestures and 28 hand gestures, respectively.

Table 1. Recognition accuracy (%) of our method for different network branches on the SHREC'17 Track dataset.

Network branches	14 Gestures(%)	28 Gestures(%)
TPB(short-term)	95.71	91.19
TPB(long-term)	95.48	89.88
SPB(low-frequency)	72.14	67.38
SPB(high-frequency)	71.31	65.36
TPB + SPB	**96.31**	**93.21**

Table 2. Recognition accuracy (%) of our method for different joint distances on the SHREC'17 Track dataset.

Joint distances	14 Gestures(%)	28 Gestures(%)
Correlation distance	95.48	89.52
Cosine distance	95.95	91.43
Cityblock distance	96.07	92.86
Euclidean distance	**96.31**	**93.21**

Different Joint Distances. To investigate the influence of different joint distances on recognition accuracy, we design four related ablation experiments. As shown in Table 2, using Euclidean distance as the metric can obtain the best recognition performance, which can reach 96.31% and 93.21% on 14 hand gestures and 28 hand gestures, respectively. Besides, we find that the recognition result of using Correlation distance as the metric is the worst. This is mainly because it reduces the inter-class variance to a certain extent. Besides, we find that Cityblock distance also can obtain satisfactory results, which reflects the spatial position relationship of skeleton joints.

4.4 Comparison with the State-of-the-Art

In this section, we compare our method with the state-of-the-art methods on the SHREC'17 Track dataset [27] and FPHA dataset [9]. For the former, we use 1960 sequences for training and 840 sequences for testing. As shown in Table 3, our proposed DR-Net achieves 96.31% and 93.21% recognition accuracy for 14 hand gestures and 28 hand gestures, respectively. The DR-Net adopts temporal skeleton representation and spatial skeleton representation as input sources, which effectively improves recognition results. For the latter, we quote related results of compared methods from this paper [23] to demonstrate the effectiveness of our proposed DR-Net. As shown in Table 4, our approach outperforms the state-of-the-art methods. Besides, ST-TS-HGR-NET [23] uses SVM for hand gesture recognition, which is not suitable for larger datasets. HMM+HPEV [19] uses deep neural networks to recognize hand gestures, which performs poorly on smaller datasets. The related experimental results demonstrate that our method is suitable for datasets of various sizes.

Table 3. Comparison of recognition accuracy (%) with the state-of-the-art methods on the SHREC'17 Track dataset.

Method	14 Gestures(%)	28 Gestures(%)
Key-Frame CNN [27]	82.90	71.90
SoCJ+Dir+Rot [26]	86.90	84.20
3D PAT [1]	90.50	80.50
Two-stream 3DCNN [28]	83.45	77.43
SEM-MEM+WAL [18]	90.83	85.95
Res-TCN [11]	91.10	87.30
STA-Res-TCN [11]	93.60	90.70
MFA-Net [4]	91.31	86.55
DG-STA [5]	94.40	90.70
DD-Net [33]	94.60	91.90
DeepGRU [21]	94.50	91.40
ST-GCN [32]	92.70	87.70
ST-TS-HGR-NET [23]	94.29	89.40
HMM+HPEV [19]	94.88	92.26
NC-CNN [10]	94.80	92.90
Ours	**96.31**	**93.21**

Table 4. Comparison of recognition accuracy (%) with the state-of-the-art methods on the FPHA dataset (**C**, **D**, **S** represent color images, depth images and skeleton data).

Method	C	D	S	45 Gestures(%)
JOULE-color [12]	✓	✗	✗	66.78
JOULE-depth [12]	✗	✓	✗	60.17
JOULE-pose [12]	✗	✗	✓	74.60
JOULE-all [12]	✓	✓	✓	78.78
1-layer LSTM [36]	✗	✗	✓	78.73
2-layer LSTM [36]	✗	✗	✓	80.14
Moving Pose [34]	✗	✗	✓	56.34
Lie Group [29]	✗	✗	✓	82.69
HBRNN [7]	✗	✗	✓	77.40
Gram Matrix [35]	✗	✗	✓	85.39
TF [8]	✗	✗	✓	80.69
Riemannian-Net [13]	✗	✗	✓	84.35
Grassmann-Net [14]	✗	✗	✓	77.57
ST-TS-HGR-NET [23]	✗	✗	✓	93.22
HMM+HPEV [19]	✗	✗	✓	90.96
NC-CNN [10]	✗	✗	✓	87.60
Ours	✗	✗	✓	**94.09**

5 Conclusion

In this paper, we propose a novel DR-Net for skeleton-based dynamic hand gesture recognition, which decouples the original skeleton data into different skeleton representations as input. For the temporal perception branch, we use short-term motion features and long-term motion features as the temporal skeleton representation. Meanwhile, we design the TFM to capture multi-scale temporal features. For the spatial perception branch, we use spatial low-frequency features and spatial high-frequency features as the spatial skeleton representation. Besides, we also propose the SFM to enhance important spatial features. On the two public benchmark datasets, related experimental results demonstrate that our proposed DR-Net is competitive with the state-of-the-art methods. At present, our method cannot adaptively learn spatial geometric relationships, which leads to unsatisfactory performance for the spatial perception branch. In the future, we intend to use GCN to automatically learn the spatial geometric relationship between joints.

References

1. Boulahia, S.Y., Anquetil, E., Multon, F., Kulpa, R.: Dynamic hand gesture recognition based on 3d pattern assembled trajectories. In: 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1–6. IEEE
2. Caetano, C., Sena, J., Brémond, F., Dos Santos, J.A., Schwartz, W.R.: Skelemotion: a new representation of skeleton joint sequences based on motion information for 3d action recognition. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–8. IEEE (2019)
3. Cao, W.: Application of the support vector machine algorithm based gesture recognition in human-computer interaction. Informatica **43**(1), 123–127 (2019)
4. Chen, X., Wang, G., Guo, H., Zhang, C., Wang, H., Zhang, L.: Mfa-net: motion feature augmented network for dynamic hand gesture recognition from skeletal data. Sensors **19**(2), 239 (2019)
5. Chen, Y., Zhao, L., Peng, X., Yuan, J., Metaxas, D.N.: Construct dynamic graphs for hand gesture recognition via spatial-temporal attention. In: 30th British Machine Vision Conference, pp. 103–116 (2019)
6. Côté, S., Beaulieu, O.: VR road and construction site safety conceptual modeling based on hand gestures. Front. Robot. AI **6**, 15 (2019)
7. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1110–1118 (2015)
8. Garcia-Hernando, G., Kim, T.K.: Transition forests: learning discriminative temporal transitions for action recognition and detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 432–440 (2017)
9. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.K.: First-person hand action benchmark with RGB-d videos and 3d hand pose annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 409–419 (2018)

10. Guo, F., He, Z., Zhang, S., Zhao, X., Fang, J., Tan, J.: Normalized edge convolutional networks for skeleton-based hand gesture recognition. Pattern Recogn. **118**, 108044 (2021)
11. Hou, J., Wang, G., Chen, X., Xue, J.-H., Zhu, R., Yang, H.: Spatial-temporal attention res-tcn for skeleton-based dynamic hand gesture recognition. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11134, pp. 273–286. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11024-6_18
12. Hu, J.F., Zheng, W.S., Lai, J., Zhang, J.: Jointly learning heterogeneous features for RGB-d activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5344–5352 (2015)
13. Huang, Z., Gool, L.V.: A riemannian network for SPD matrix learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 2036–2042 (2017)
14. Huang, Z., Wu, J., Van Gool, L.: Building deep networks on grassmann manifolds. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), pp. 3279–3286. AAAI Press (2018)
15. Jiang, X., Xu, K., Sun, T.: Action recognition scheme based on skeleton representation with ds-LSTM network. IEEE Trans. Circ. Syst. Video Technol. **30**(7), 2129–2140 (2019)
16. Lee, M., Lee, J., Chang, J.H.: Ensemble of jointly trained deep neural network-based acoustic models for reverberant speech recognition. Digital Sig. Process. **85**, 1–9 (2019)
17. Li, Y., Guo, T., Liu, X., Xia, R.: Skeleton-based action recognition with lie group and deep neural networks. In: 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), pp. 26–30. IEEE (2019)
18. Liu, H., Tu, J., Liu, M., Ding, R.: Learning explicit shape and motion evolution maps for skeleton-based human action recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1333–1337. IEEE (2018)
19. Liu, J., Liu, Y., Wang, Y., Prinet, V., Xiang, S., Pan, C.: Decoupled representation learning for skeleton-based gesture recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5751–5760 (2020)
20. Liu, J., Liu, N., Wang, P., Wang, M., Guo, S.: Array-less touch position identification based on a flexible capacitive tactile sensor for human-robot interactions. In: 2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM), pp. 458–462. IEEE (2019)
21. Maghoumi, M., LaViola, J.J.: DeepGRU: deep gesture recognition utility. In: Bebis, G., et al. (eds.) ISVC 2019. LNCS, vol. 11844, pp. 16–31. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33720-9_2
22. Nasri, N., Orts-Escolano, S., Cazorla, M.: An semg-controlled 3d game for rehabilitation therapies: real-time time hand gesture recognition using deep learning techniques. Sensors **20**(22), 6451 (2020)
23. Nguyen, X.S., Brun, L., Lézoray, O., Bougleux, S.: A neural network based on SPD manifold learning for skeleton-based hand gesture recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12028–12037. IEEE (2019)
24. Nunez, J.C., Cabido, R., Pantrigo, J.J., Montemayor, A.S., Velez, J.F.: Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recogn. **76**, 80–94 (2018)
25. Rastgoo, R., Kiani, K., Escalera, S.: Sign language recognition: a deep survey. Expert Syst. Appl. **164**, 113794 (2021)

26. de Smedt, Q.: Dynamic hand gesture recognition-From traditional handcrafted to recent deep learning approaches. Ph.D. thesis, Université de Lille 1, Sciences et Technologies; CRIStAL UMR 9189 (2017)
27. de Smedt, Q., Wannous, H., Vandeborre, J.P., Guerry, J., Le Saux, B., Filliat, D.: Shrec 2017 track: 3d hand gesture recognition using a depth and skeletal dataset. In: 3DOR-10th Eurographics Workshop on 3D Object Retrieval, pp. 1–6 (2017)
28. Tu, J., Liu, M., Liu, H.: Skeleton-based human action recognition using spatial temporal 3d convolutional neural networks. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE Computer Society (2018)
29. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 588–595 (2014)
30. Wei, S., Song, Y., Zhang, Y.: Human skeleton tree recurrent neural network with joint relative motion feature for skeleton based action recognition. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 91–95. IEEE (2017)
31. Weng, J., Liu, M., Jiang, X., Yuan, J.: Deformable pose traversal convolution for 3d action and gesture recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 136–152 (2018)
32. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
33. Yang, F., Wu, Y., Sakti, S., Nakamura, S.: Make skeleton-based action recognition model smaller, faster and better. In: Proceedings of the ACM multimedia Asia, pp. 1–6 (2019)
34. Zanfir, M., Leordeanu, M., Sminchisescu, C.: The moving pose: an efficient 3d kinematics descriptor for low-latency action recognition and detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2752–2759 (2013)
35. Zhang, X., Wang, Y., Gou, M., Sznaier, M., Camps, O.: Efficient temporal sequence comparison and classification using Gram matrix embeddings on a Riemannian manifold. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4498–4507 (2016)
36. Zhu, W., et al.: Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp. 3697–3703 (2016)