SkeletonNet: Mining Deep Part Features for 3D Action Recognition

Qiuhong Ke, Senjian An, Mohammed Bennamoun, Ferdous Sohel, and Farid Boussaid

***Abstract*—This letter presents SkeletonNet, a deep learning framework for skeleton-based 3D action recognition. Given a skeleton sequence, the spatial structure of the skeleton joints in each frame and the temporal information between multiple frames are two important factors for action recognition. We firstly extract body-part based features from each frame of the skeleton sequence. Compared to the original coordinates of the skeleton joints, the proposed features are translation, rotation and scale invariant. To learn robust temporal information, instead of treating the features of all frames as a time series, we transform the features into images and feed them to the proposed deep learning network which contains two parts, one to extract general features from the input images, while the other to generate a discriminative and compact representation for action recognition. The proposed method is tested on SBU kinect interaction dataset, CMU dataset and the large scale NTU RGB+D dataset and achieves state-of-the-art performance.**

***Index Terms*—convolutional neural networks (CNNs), 3D action recognition, robust features**

I. Introduction

HUMAN action recognition has received increasing attention [1]–[7] due to its wide range of applications such as video surveillance, human-machine interaction and robot control [8]. The 3D representations of human actions provide more comprehensive information than 2D RGB videos [9]–[13]. Recently, many works have investigated skeleton-based 3D action recognition due to the availability of highly-accurate data acquisition devices and real-time skeleton estimation algorithms [14], [15]. A human skeleton can be grouped into five sets of joints corresponding to five body parts, i.e., the trunk, the left arm, the right arm, the left leg and the right leg. Different body parts have their own specific features and importance for various actions. Certain actions may only involve the motion of one limb. For example, the action of waving can be recognised merely from the motion of the hands. Other actions may involve the movements of several or all of the body parts (*e.g.*, picking up an object). In this paper, we propose a body-part based feature learning framework for skeleton-based action recognition.

Given a skeleton sequence, only the 3D coordinates of the skeleton joints are provided in each frame. There are two important factors to recognize the action class from the skeleton sequence. One is to design robust features to describe the spatial structure of the skeleton joints in each frame. Another is to extract temporal information among multiple frames of the sequence [9].

To extract robust spatial structural information, an origin (*e.g.*, the hip joint) and a reference skeleton are usually used to normalize the skeleton data to the same center and scale due to the lack of invariance properties of the absolute positions of the joints [16]. While scale and translation invariance can easily be achieved, rotation invariance is more difficult to handle. In this paper, a vector-based representation, which is scale, translation and rotation invariant, is introduced for the five body parts in each frame of a skeleton sequence. The vectors of each body part are generated from the pairwise relative positions of a selected starting joint to the other joints. For different actions, the cosine distance (CD) between two vectors and the normalized magnitude (NM) of each vector capture the spatial structure of a body part, and its relationships to the other parts. CD and NM are thus used to represent the spatial structural information of the skeleton sequence in each frame.

Previous works model the temporal structure of a skeleton sequence as a time series based on Long Short Term Memory (LSTM) [17], Fourier Temporal Pyramid (FTP) [16], [18], [19], or Hidden Markov Models (HMMs) [20]–[22]. In this paper, a deep learning method, which is based on CNN, is proposed to learn the high-level temporal representations from the low-level features. CNN is used in this paper as it is capable of exploiting more salient and robust features than hand-crafted features. Moreover, it has shown great success for various visual recognition tasks [23], [24]. In addition, the massive image databases such as ImageNet [25] can be leveraged to pre-train CNN models. To learn high-level temporal representations, the CD and NM features of all frames of each body part are concatenated and scaled into gray images with values between 0 to 255. They are further resized into the same dimension and fed to the deep network, which includes two parts. The first part extracts generic CNN features from the CD and NM images, and the second aggregates the extracted features and learns a compact and discriminative representation for action recognition.

The contributions of this paper include: **1)** well-designed vector-based features for each body part of human skeleton sequences, which are translation, scale and rotation invariant,

Qiuhong Ke, Senjian An and Mohammed Bennamoun are with School of Computer Science and Software Engineering, The University of Western Australia, Crawley, Australia.
E-mail: qiuhong.ke@research.uwa.edu.au, senjian.an@uwa.edu.au, mohammed.bennamoun@uwa.edu.au

Ferdous Sohel is with School of Engineering and Information Technology, Murdoch University, Murdoch, Australia.
E-mail: f.sohel@murdoch.edu.au

Farid Boussaid is with School of Electrical, Electronic and Computer Engineering, The University of Western Australia, Crawley, Australia.
E-mail: farid.boussaid@uwa.edu.au

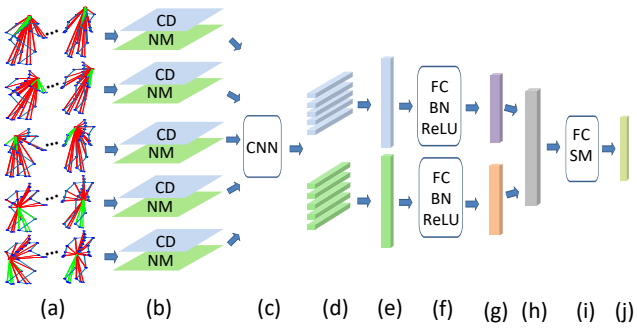

Fig. 1. Overall pipeline of the proposed SkeletonNet. (a) Vector generation of a skeleton sequence for the five body parts. From the top to the bottom are the sequences of the trunk, right arm, left arm, right leg and left leg. (b) Ten feature arrays, including five cosine distance (CD) arrays and five normalized magnitude (NM) arrays calculated from the sequences of the five body parts. They are transformed to a set of images and fed to a deep CNN (c) for feature learning. (d) Outputs of the CNN network, including five high-level CD features and five high-level NM features. (e) Concatenation of the five high-level CD features and the five high-level NM features. They are separately fed to a two-stream network (f) where each contains a fully connected (FC) layer, a batch normalization (BN) layer and a rectified linear unit (ReLU). (g) Outputs of the two steams of networks. (h) Concatenation of the outputs of the two-stream network, which is fed to another network (i) including a FC layer and a Softmax (SM) layer for classification. (j) Classification scores.

2) a deep learning method based on CNN to learn high-level and discriminative representations from the low-level features and **3)** the state-of-the-art performance for skeleton-based action recognition on challenging databases.

II. Approach

This section presents the pipelines (Fig.1) of the proposed SkeletonNet for 3D skeleton-based action recognition. The spatial information is encapsulated in the proposed CD and NM features which capture the spatial structure of a body part, and its relationships to the other parts. CNN is used to process the CD and NM arrays and learn the high-level spatial information. CNN is capable to learn hierarchical features. In contrast, LSTM provides good temporal modelling but has more difficulty to learn high-level features. The novelty of the proposed method is to learn high-level robust and discriminative representations from low-level features. The well designed low-level features are translation, rotation and scale invariant. For a skeleton, if it is rotated or scaled, the cosine distance and the normalized magnitude are still the same. Thus, the features are rotation and scale invariant. Similar to temporal sampling, the temporal invariance with respect to camera speed can be achieved through image resizing in the procedure of generating images from the CD and NM arrays.

A. Robust Features of Body Parts

This paper aims to extract robust features from the skeleton sequences. We propose to compute a set of vectors between two joints to capture the relationship between joint pairs. Subsequently, instead of using the coordinates (x,y,z) to represent each vector, we compute CD and NM to provide scale and rotation invariance. The CD between two vectors can capture the spatial structure of a body part, and its relationships to the other parts, while NM reflects the magnitude variations of the vectors. CD and NM are complementary and their combination is shown to provide superior performance. The proposed CD and NM are translation invariant as they reflect the relative locations of skeleton joints. They are also not affected by the rotation, *e.g.*, when a human skeleton rotates for some degree, the CD and NM between two joints remain to be same.

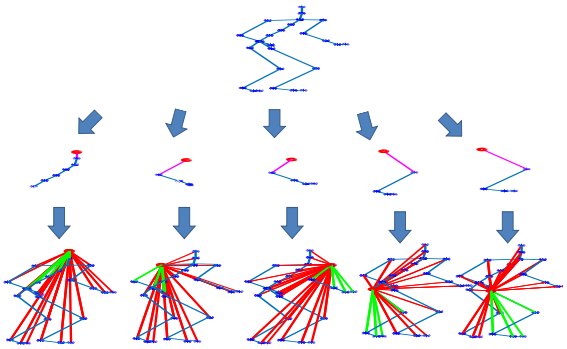

Fig. 2. Vector generation. A human skeleton can be grouped into five parts (from the left to the right are the trunk, the left arm, the right arm, the left leg and the right leg). Five joints (shown in red, i.e., the head, the left shoulder, the right shoulder, the left hip and the right hip) are selected as the starting joints to connect with other joints to generate within-part vectors (shown in green) and between-part vectors (shown in red) for the five parts. In each part, a reference vector (shown in magenta) is also selected to be compared with other vectors to calculate normalized magnitudes.

1) Vector Generation: Given a frame of a skeleton sequence, let the 3D coordinates of the skeleton joints be:

$$\Omega = \{\mathbf{p}_i \in \mathbb{R}^3 : i = 1, \cdots, n\} \tag{1}$$

where n is the number of the skeleton joints, and $\mathbf{p}_i = [x_i, y_i, z_i]$ is the 3D coordinate of the i^{th} joint. All the skeleton joints are separated into five groups corresponding to the trunk, left arm, right arm, left leg and right leg, more precisely:

$$\Omega = \bigcup_{k=1}^{5} \Omega_k \tag{2}$$

where Ω_k is the set of joints in the k^{th} part.

For each body part Ω_k, a joint, namely $\mathbf{p}_0^{(k)}$, is selected as the starting joint. For any other joint, namely $\mathbf{p}$, we define the set of within-part vectors as

$$\mathcal{V}_{\mathrm{w}}^{(k)} \triangleq \left\{ \mathbf{p} - \mathbf{p}_0^{(k)} : \mathbf{p} \in \Omega_k \right\} \tag{3}$$

and the set of between-part vectors as

$$\mathcal{V}_{\mathrm{b}}^{(k)} \triangleq \left\{ \mathbf{p} - \mathbf{p}_0^{(k)} : \mathbf{p} \in \Omega \backslash \Omega_k \right\}. \tag{4}$$

In this paper, the head, left shoulder, right shoulder, left hip and right hip are selected as the starting joints for the trunk, left arm, right arm, left leg and right leg, respectively. This selection is based on the fact that these selected joints are fixed in most actions so that the designed between-part and with-part vectors in Equation 3 and Equation 4 can reflect the motions of the other joints. For the trunk part, the base of the spine seems more fixed than the head joint. However, the base of the spine is close to the left hip and right hip, which might result in information redundancy if it is selected as the starting joint.

2) CD and NM: For any $\mathbf{v} \in \mathcal{V}_{\mathrm{w}}^{(k)}$, and any $\mathbf{u} \in \mathcal{V}_{\mathrm{w}}^{(k)} \cup \mathcal{V}_{\mathrm{b}}^{(k)}, \mathbf{u} \neq \mathbf{v}$, their cosine distance is defined by

$$\frac{\mathbf{v}^T\mathbf{u}}{\|\mathbf{v}\|\|\mathbf{u}\|} \tag{5}$$

All these distances are concatenated as the CD feature for part Ω_k with dimension $(n_k-1)(n-2)$. Here n_k is the number of the human skeleton joints in part Ω_k, and n is the number of the entire human skeleton joints.

For any $\mathbf{u} \in \mathcal{V}_{\mathrm{w}}^{(k)} \cup \mathcal{V}_{\mathrm{b}}^{(k)}$, the NM is defined as

$$\frac{\|\mathbf{u}\|}{\|\mathbf{u}_0^{(k)}\|} \tag{6}$$

where $\mathbf{u}_0^{(k)}$ is the selected reference vector, whose length usually remains fixed in motions, for part Ω_k to normalize other vectors. The five selected reference vectors are the neck, the left upper arm, the right upper arm, the left upper leg and the right upper leg, as shown in magenta in Figure 2. All the normalized magnitudes are concatenated as the NM feature for part Ω_k with dimension $n-1$.

B. High-level Feature Learning

Given a skeleton sequence of t frames, a CD array with dimension $(n_k-1)(n-2)\times t$ and a NM array with dimension $(n-1)\times t$ can be obtained by extracting and aggregating the features of all frames. The CD and NM arrays of each body part are then separately fed to a deep CNN to learn high-level spatial features. Each column of the array represents the spatial structural features of each frame. The temporal information could thus also be learned from all columns of the entire arrays with CNN.

More specifically, the CD and NM arrays of the five parts are firstly scaled into gray images with values between 0 to 255, and further resized to 224×224. Image resizing is similar to temporal sampling, which can handle sequences of different lengths. The advantages of transforming the skeleton features from a sequence into an image are that the sequences of different lengths can be handled with simple image resizing and that the CNN can be used to learn high-level features. For the CD and NM array, the features of the neighbouring joints and the features of the same joint in the neighbouring frames change smoothly. Hence the pixels of the images do not change sharply. The images generated from the ten arrays (two for each body part) are separately fed to a deep network to learn high-level features. The network shares the parameters of the pre-trained VGG-M network [26]. The edges and salient features of the original images are captured after convolution.

The layer of 4096-dimensional (4096D) fc6 is used as the output feature vector. Thus the outputs of the network contains five 4096D feature vectors from each of the CD and NM arrays, respectively, each corresponding to one body part, as shown in Figure 1(d). The five 4096D vectors of the CD or NM arrays are concatenated as two feature vectors, respectively. The two features are then fed in a two-stream network including a fully connected (FC), a batch normalization (BN) [27] and a ReLU [28] layer. Each stream outputs a 512D feature vector, and the two vectors are then concatenated as a 1024D feature vector and sent to another FC layer, followed by a Softmax layer for classification.

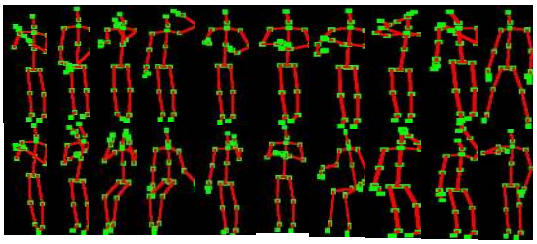

Fig. 3. Sample examples of NTU RGB+D Dataset.

III. EXPERIMENTS

The proposed SkeletonNet is tested on NTU RGB+D dataset [29], SBU kinect interaction dataset [30] and CMU dataset [31]. For all datasets, the learning rate is set to 0.001 and batch size is set to 100. The training is stopped at 25 epochs. The performance of the proposed method on each dataset is compared with previous methods using the same testing protocol. To show the advantages of the proposed robust features and the proposed feature learning method, two ablative analyses are also conducted in the experiments: **1)** To show the contribution of the robust features, the coordinates of the skeletons were used to replace the CD and NM feature arrays, and the remaining learning framework is the same as that of SkeletonNet; **2)** to demonstrate the contribution of the proposed feature learning method, Fourier Temporal Pyramid (FTP) is used to learn temporal information of skeleton sequences based on CD and NM features, followed by SVM for classification. FTP has widely been used to learn temporal information of videos. FTP has been combined with SVM for action recognition [16], [18], [19]. Therefore the combination of FTP and SVM is used as a baseline in comparison to show the advantages of the proposed method.

A. NTU RGB+D Dataset

This dataset contains 56880 sequences of 60 classes of actions. Some examples are shown in Fig.3. These actions are performed by 40 distinct subjects and captured by three cameras. It is a very challenging dataset due to the large intra-class, sequence length and view point variations. The evaluations are performed using the two standard protocols proposed by [29]: **1)** cross-subject evaluation, for which sequences associated to half of the subjects are used for training and the remaining are used for testing; **2)** cross-view evaluation, for which the sequences captured by two cameras are used for training and the rest are used for testing.

The results are shown in Table I. It can be seen that the proposed SkeletonNet performs significantly better than others in both testing protocols. When testing with the cross-subject protocol, the performance is about 75.94%, which is about 6.74% better than the ST-LSTM method [17]. The accuracy is about 81.16% when testing with cross-view protocol. Compared to the ST-LSTM method [17], the improvement is about 3.46%. The good performance of the proposed method

TABLE I
COMPARISONS ON THE NTU RGB+D DATASET.

Methods	Accuracy	
	Cross Subject	Cross View
Lie Group [16]	50.1%	52.8%
Skeletal Quads [32]	38.6%	41.4%
Dynamic Skeletons [33]	60.2%	65.2%
Hierarchical RNN [34]	59.1%	64.0%
Deep RNN [29]	59.3%	64.1%
Deep LSTM [29]	60.7%	67.3%
Part-aware LSTM [29]	62.9%	70.3%
ST-LSTM (Tree) + Trust Gate [17]	69.2%	77.7%
Joints+Network Learning	72.81%	71.61%
Robust Features+FTP+SVM	63.63%	79.09%
SkeletonNet	**75.94%**	**81.16%**

TABLE II
COMPARISONS ON THE SBU KINECT INTERACTION DATASET.

Methods	Accuracy
Raw Skeleton [30]	49.7%
Joint Feature [30]	80.3%
Raw Skeleton [35]	79.4%
Joint Feature [35]	85.9%
Hierarchical RNN [34]	80.35%
Deep LSTM [9]	86.03%
Deep LSTM + Co-occurrence [9]	90.41%
ST-LSTM (Tree) + Trust Gate [17]	93.3%
Joints+Network Learning	85.93%
Robust Features+FTP+SVM	87.95%
SkeletonNet	**93.47%**

is due to the robust CD and NM features, as well as the deep learning method. As shown in Table I, when using joint coordinates instead of the proposed robust features (i.e., Joints+Network Learning), the performance is reduced on both testing protocols. Particularly, when tested on the cross-view protocol, the proposed method is about 10% better than Joints+Network Learning method. This is due to the fact that the proposed CD and NM features are rotation invariant, thus providing invariance against view points and improving performance. From Table I, it can also be seen that the proposed learning method performs better than FTP and SVM (i.e., Robust Features+FTP+SVM). For the NTU dataset of different subjects, the skeleton sequences have variant lengths. The temporal features learned by SkeletonNet are seen to be more robust and powerful than FTP. This is because for the training method of SkeletonNet, the features of all frames in each sequence are transformed into an image. The robust temporal features of the sequences are thus captured by learning translation invariant features from the images with the convolution and pooling operations of SkeletonNet.

B. SBU Kinect Interaction Dataset

The SBU kinect interaction dataset is a two-person interaction dataset collected by the Microsoft Kinect sensor. It contains 283 videos of 8 types of interactions performed by two persons (i.e., approaching, departing, kicking, punching, pushing, hugging, shaking hands and exchanging). The evaluation is done through a 5-fold cross validation, with the same data split as proposed in [30]. For each video, there are two separate human skeletons. For data augmentation, the images generated from the CD and NM arrays are first resized to 250×250, and twenty sub-images with fixed size of 224×224 are then randomly cropped from the original image, with a further random horizontal flipping. For testing, the scores of all the augmented samples are averaged for the final decision for action recognition.

Compared to other methods (Table II), the proposed SkeletonNet achieves the best performance, with an accuracy of 93.47%, which is better than the spatial temporal LSTM (ST-LSTM) [17]. Compared to the Deep LSTM + Co-occurrence method [9], the improvement is about 3.07%. From Table II it can also be seen that the proposed method performs better than the method Joints+Network Learning and the method Robust Features+FTP+SVM. This clearly shows the effectiveness of the proposed robust features and the feature learning method.

From Table I and Table II it can also be seen that for the small SBU kinect interaction dataset, our method and the ST-LSTM method [17] achieve comparable performance. For the large NTU dataset (which is more challenging), our method is more robust and achieves a much better performance than the ST-LSTM method [17].

C. CMU Dataset

This dataset contains 2,235 sequences, which has been categorized into 45 classes [31]. As proposed in [9], the evaluation has been performed with a subset of 664 sequences and the entire dataset. The proposed method achieves an accuracy of 89.46% on the subset, which is slightly better than 88.40% achieved by [9]. For the entire dataset, the performance of the proposed method is 84.83%, which is 3.79% better than [9].

IV. CONCLUSION

In this paper, a novel feature learning framework SkeletonNet has been proposed for skeleton based action recognition. Given a skeleton sequence, a set of vectors are generated with the selected pairs of joints for each body part. Then the spatial structure of each body part and their relationships are modelled using geometric properties of the vectors, including the cosine distances and the normalized magnitudes. The two feature arrays are transformed into gray images, which are then fed to the proposed deep learning architecture for high-level feature learning and action recognition. The proposed feature learning framework is based on image-format inputs, and it is not only suitable for large datasets, but also for small datasets with the help of pre-trained CNN models and image augmentation. It also performs well across multi-type actions (i.e., one-person actions or two-person interactions). Experimental results have demonstrated state-of-the-art performance of the proposed method on three skeleton datasets.

ACKNOWLEDGMENT

This work was partially supported by Australian Research Council grants DP150100294, DP150104251, and DE120102960. This paper used the NTU RGB+D Action Recognition Dataset made available by the ROSE Lab at the Nanyang Technological University, Singapore.

REFERENCES

[1] Y. Zhu, W. Chen, and G. Guo, "Fusing multiple features for depth-based action recognition," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 2, p. 18, 2015.

[2] Y. Zhu and G. Guo, "A study on visible to infrared action recognition," *IEEE Signal Processing Letters*, vol. 20, no. 9, pp. 897–900, 2013.

[3] Y. Song, S. Liu, and J. Tang, "Describing trajectory of surface patch for human action recognition on rgb and depth videos," *IEEE Signal Processing Letters*, vol. 22, no. 4, pp. 426–429, 2015.

[4] B. Hu, J. Yuan, and Y. Wu, "Discriminative action states discovery for online action recognition," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1374–1378, 2016.

[5] J. Song, H. Shen *et al.*, "Beyond frame-level cnn: Saliency-aware 3d cnn with lstm for video action recognition," *IEEE Signal Processing Letters*.

[6] C. Liang, E. Chen, L. Qi, and L. Guan, "Improving action recognition using collaborative representation of local depth map feature," *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1241–1245, 2016.

[7] Y. Ming, G. Wang, and C. Fan, "Uniform local binary pattern based texture-edge feature for 3d human behavior recognition," *PloS one*, vol. 10, no. 5, p. e0124640, 2015.

[8] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: a review," *arXiv preprint arXiv:1601.01006*, 2016.

[9] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-Occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[10] G. Zhang, L. Tian, Y. Liu, J. Liu, X. A. Liu, Y. Liu, and Y. Q. Chen, "Robust real-time human perception with depth camera," in *ECAI 2016: 22nd European Conference on Artificial Intelligence, 29 August-2 September 2016, The Hague, The Netherlands-Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*, vol. 285. IOS Press, 2016, p. 304.

[11] J. Liu, G. Zhang, Y. Liu, L. Tian, and Y. Q. Chen, "An ultra-fast human detection method for color-depth camera," *Journal of Visual Communication and Image Representation*, vol. 31, pp. 177–185, 2015.

[12] J. Liu, Y. Liu, G. Zhang, P. Zhu, and Y. Q. Chen, "Detecting and tracking people in real time with rgb-d camera," *Pattern Recognition Letters*, vol. 53, pp. 16–23, 2015.

[13] J. Liu, Y. Liu, Y. Cui, and Y. Q. Chen, "Real-time human detection and tracking in complex environments using single rgbd camera," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 2013, pp. 3088–3092.

[14] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[15] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Advances in neural information processing systems*, 2014, pp. 1799–1807.

[16] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a Lie group," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 588–595.

[17] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-Temporal LSTM with trust gates for 3D human action recognition," in *European Conference on Computer Vision (ECCV)*, 2016.

[18] J. Luo, W. Wang, and H. Qi, "Group sparsity and geometry constrained dictionary learning for action recognition from depth maps," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 1809–1816.

[19] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1290–1297.

[20] F. Lv and R. Nevatia, "Recognition and segmentation of 3-D human action using HMM and multi-class adaboost," in *European Conference on Computer Vision (ECCV)*, 2006, pp. 359–372.

[21] D. Wu and L. Shao, "Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 724–731.

[22] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 20–27.

[23] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014, pp. 512–519.

[24] Q. Ke, M. Bennamoun, S. An, F. Boussaid, and F. Sohel, "Human interaction prediction using deep temporal features," in *European Conference on Computer Vision*. Springer, 2016, pp. 403–414.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conferences on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[26] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *British Machine Vision Conference (BMVC)*, 2014.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[28] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.

[29] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[30] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 28–35.

[31] CMU, "CMU graphics lab motion capture database. http://mocap.cs.cmu.edu/," 2013.

[32] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *International Conference on Pattern Recognition (ICPR)*, 2014, pp. 4513–4518.

[33] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5344–5352.

[34] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1110–1118.

[35] Y. Ji, G. Ye, and H. Cheng, "Interactive body part contrast mining for human interaction recognition," in *Multimedia and Expo Workshops (ICMEW)*, 2014, pp. 1–6.