Skeleton-based human activity analysis using deep neural networks with adaptive representation transformation

1st Jiahui Yu
School of Computing
University of Portsmouth
Portsmouth, UK
jiahui.yu@port.ac.uk

2nd Hongwer Gao
School of Automation and Electrical Engineering
Shenyang Ligong University
Shenyang, China
ghw1978@sohu.com

3rd Qing Gao
School of Science and Engineering
The Chinese University of Hong Kong
Shenzhen, China
gaoqing@cuhk.edu.cn

4th Dalin Zhou
School of Computing
University of Portsmouth
Portsmouth, UK
dalin.zhou@port.ac.uk

5th Zhaojie Ju
School of Computing
University of Portsmouth
Portsmouth, UK
zhaojie.ju@port.ac.uk

***Abstract*—Compared with RGB-D-based human action analysis, skeleton-based works reach higher robustness and better performance, which are widely applied in the real world. However, the diversity of action observation perspectives hinders the improvement of recognition accuracy. Most of the existing works solve this problem by increasing the amount of training data, which brings a huge computational cost and cannot improve the robustness of the models. In this paper, an adaptive model is proposed to obtain high-performance representations to improve human action recognition accuracy. First, a skeleton representation transformation method is proposed to transform the skeleton model to the best perspective, in which all parameters can be adaptively learned. This is more robust and cost-effective than hand-crafted features. Next, a re-designed backbone is proposed to train the model with a small computational cost based on the 3D-CNN. In the training process, a data enhancement method is also introduced to improve the robustness of the model. Finally, extensive experimental evaluations are conducted on two benchmarks, including the NTU RGB-D dataset and the SBU interaction dataset. The results show that the proposed model can effectively and adaptively obtain high-performance skeleton representation and its performance is better than other state-of-the-art methods.**

***Index Terms*—human action analysis, skeleton representation, CNN,deep learning**

I. Introduction

Human activity analysis is a key technology in the research area of human robot/computer interaction, which is widely applied in many applications, such as video analysis, surveillance systems, medical services, and entertainment games [1]–[3]. Motivated by the development of neural network, in recent years, existing human action analysis methods commonly are categorized in two forms, including CNN-based models (convolutional neural network) and RNN-based models (recurrent neural network).

Corresponding author: Zhaojie Ju

Normally, human action analysis methods can be classified based on the types of input data, including RGB-based works, RGB-D-based works, skeleton-based works, and mixed data-based works. Considering the robustness of the model in the real-world, a skeleton-based work is proposed in this paper. Because it is difficult to obtain high-performance feature representation, skeleton-based research is a challenge. Typical works, such as [4], [5], [6], and [7], extract skeleton representation by introducing hand-crafted features based on human prior experience, and then these representations are inputted CNN/RNN-based neural works for training. However, most of these methods cannot achieve satisfactory results in the case of changes in viewing perspectives. For example, observing the "sit down" action from the front and observing from the top can give different recognition results. This is because high-level feature representation cannot be obtained due to the effect of the perspective variations.

In this paper, a new skeleton representation transformation method is proposed reduce the effect of perspective variations based on deep learning. This study is partially motivated by the recent success of the idea of the skeleton transfer module and three-dimensional CNN (3D-CNN), such as in [8], [9] , [10], and [11]. Based on the above, the proposed model can obtain high-performance representations from skeleton-based input data. Additionally, spatiotemporal features can be learned effectively to analyze human action and interaction.

The contributions proposed in this work are summarized as follows:

1) An adaptive skeleton representation transformation method is proposed to re-observe skeleton models. The skeletal model in all frames of a sequence can be

automatically learned to find the best transformation perspective. This method can be achieved based on various neural networks to obtain high-performance skeleton representations.
2) A re-designed 3D-CNN-based backbone is proposed to train the input data and obtain the final classification results. 3D convolution can effectively model spatial and temporal information at the same time.
3) An effective training scheme is proposed. Random enhancement of input data in the training process improves the difficulty of training, thereby reducing the data requirements of the model and improving its performance.
4) A series of evaluations are conducted on two benchmark datasets, in which evaluation indicators for multi-view data analysis are used. The results show the correctness of the model improvement ideas and the superiority of recognition performance.

The remainder of this paper is organized as follows: Section II briefly reviews related work about human action recognition. Section III describes the proposed method in detail, including the skeleton representation transformation method, the re-designed backbone, and a proposed training scheme. Section IV discusses the experiment results. Section V concludes this work and gives future research directions.

II. RELATED WORK

A. Deep learning-based human action analysis

CNN-based works. With the rapid developments, neural networks are widely applied in term of human action analysis. The main idea is to train and extract key features. CNNs have strong graphics classification capabilities. Normally, the image and skeleton coordinates are mapped into a matrix representation, where the pixel points and coordinate information correspond to the rows and columns of the matrix, such as [12], [13], [8], [9], and [14]. Next, the 3D-CNN-based models are proposed to model spatiotemporal features by introducing a high-dimension convolution module, such as, [10], [11], and [15].

RNN-based works. To model long-term and short-term temporal features, RNN-based models are proposed, which are effective for continuous sequence processing. A more memorable neural network is proposed for video analysis, that is, long short-term memory (LSTM). Normally, the attention mechanism is introduced into RNNs, which allows the model to selectively remember or forget certain features. Typical works are shown in [16], [17], [11], and [18].

B. Challenges

Perspective variations. To imitate the way humans observe the object in the dataset collection, the same action is usually collected from multiple perspectives. This brings challenges due to the diversity of perspectives. To solve this issue, in recent works, data enhancement is utilized, that is, increasing the amount of data allows the model to see multiple manifestations of the same action. However, the above methods cannot improve the robustness of the model and increase a lot of computational costs, which do not work well in the real-world.

High-performance representation. To alleviate the impact of changes in perspective, skeleton model transformation methods are proposed. The main idea is to transform the random input skeleton joints according to a hand-crafted feature algorithm. However, this can only consider the transformation of several common perspectives in the design, so the robustness is not good. Next, hand-crafted features are time-consuming and laborious.

III. METHOD

A. Overall

To alleviate the effects of skeleton model diversity, in this paper, an adaptive deep model is proposed to learning various skeleton models and obtain high performance representations. The proposed transformation method proves effective in multitype backbones, including RNN-based networks and CNN-based networks. To achieve the best performance, the 3D-CNN-based network is designed as the backbone in this work. The technology details about the skeleton representation transformation are introduced, and then the typical backbone is described and a new training scheme is proposed. The overall structure of the proposed method is shown in Fig. 1.

B. Skeleton representation transformation

Because the perspective of data collection is different, there are frames based on different coordinate systems in the same sequence. First, the coordinate systems of all the input frames are transformed to that of the first frame of the input sequence. Next, the proposed method can imitate the habit of human observing action and transform all skeleton models from different perspectives into the same angles. The main idea is to re-assign the same coordinate system to the skeleton model of each frame with the best observation perspective. These transformed skeleton data have high-performance representations that can be learned better in the classification network. The schematic illustration of the skeleton representation transformation process is shown in Fig. 2, which is described as follows: 1) 1) The set of each joint is denoted as $j_{t,i} = [x_{t,i}, y_{t,i}, z_{t,i}]$, where t is the number of frame and i is the number of the skeleton joints. Hence, the set of all joints is denoted as $J_t = \{j_{t,1}, j_{t,1}, ..., j_{t,i}\}$. 2) The skeleton model is transformed into the best observation coordinate system by rotating and translating the x, y, z axis in the original coordinate system. Hence, rotation angles and translation distance are key parameters, denoted as $K_t = \{a_t, b_t, c_t, d_t\}$. Where a_t , b_t , and c_t are the rotation angle of the t-th joint point around the x, y, and z axes; d_t is the translation distance. Since the human skeleton structure is rigid, it shares the same parameters during transformation. 3) R_t denotes the rotation process and T_t is the transformation process, as shown in (1) and (2). Where $r_{t,a}$, $r_{t,b}$, and $r_{t,c}$ are the rotation matrices. 4)The t-th transformed joint can be described in (3) and the set of transformed joints is shown in (4). 5) After deep model training, the model can automatically learn k_t and assign the

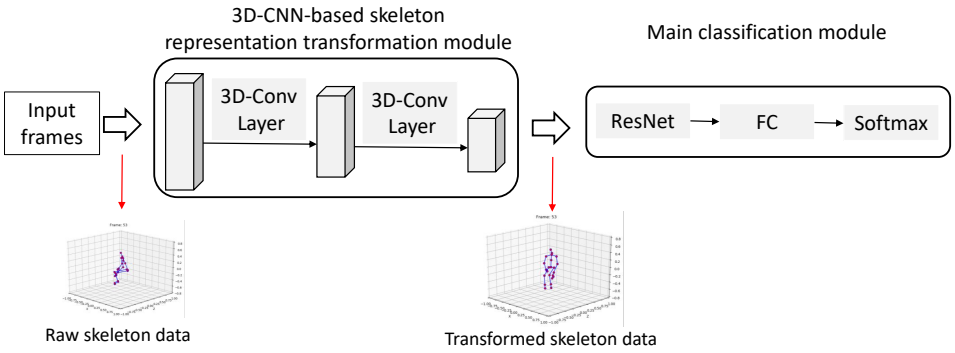

Fig. 1. Structure of the proposed method

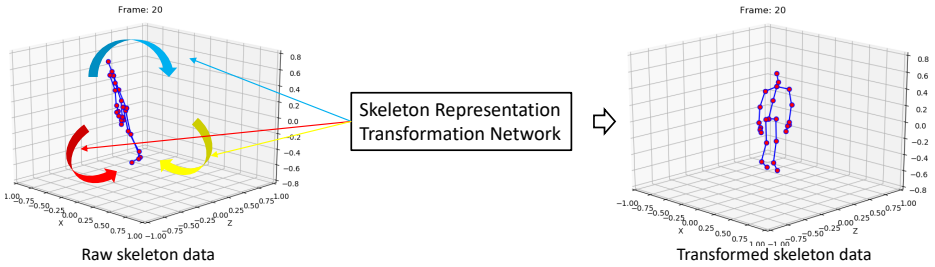

Fig. 2. Schematic illustration of the proposed skeleton representation transformation method. The red, blue and yellow arrows indicate the transformation operation along the coordinate axes.

best parameters to the skeleton joints in each frame to obtain the best skeleton representation features. Both the designed backbone and the proposed training process are shown in the following subsections.

$$R_t = r_{t,a} r_{t,b} r_{t,c} \tag{1}$$

$$T_t = j_{t,i} - d_t \tag{2}$$

$$j_{t,i}{}' = R_t T_t \tag{3}$$

$$J_t{}' = \{j_{t,1}{}', j_{t,2}{}', ..., j_{t,i}{}'\} \tag{4}$$

In short, the proposed skeleton representation transformation method has three merits for human action recognition.

1) The proposed method no longer needs extra instance-related training.
2) Representation pre-processing and transformation algorithms designed based on human prior knowledge are no longer required because the proposed model can complete automatic transformation.
3) Higher performance representations can be obtained in this work because a single transformation algorithm (hand-designed) is replaced by multiple transformation methods (deep model-learned).

C. Backbone

The proposed skeleton representation transformation method is achieved based on a re-designed 3D-CNN-based network. The main purpose is to automatically learn key parameters (the set of k_t) to obtain the high-performance skeleton representations, and then the RNN can finish the classification process. Specifically, following the works in [9], [19], the input skeleton vector is mapped to an image matrix. Next, convolution layers are utilized to learn key parameters, including a_t , b_t , c_t, and d_t . Based on the works [10], the backbone is re-designed, and the learning process is described in (5). Where $f[]$ denotes the greatest integer function, b denotes the number of pixel values, and n denotes the number of joints in the input data. Compared with the above works, the re-designed deep model proves more effective for spatiotemporal feature modeling. This is because the image map in this work has four dimensions, in which the temporal and spatial features promote each other. Additionally, although the overall idea of perspective transformation is similar to the methods proposed in [9] and [8], this paper introduces a re-designed 3D-CNN network on this basis to process temporal information. After that, the ResNet [20] is utilized as the classification network.

$$j_{t,i}{}' = R_{t,i} f[b \times \frac{j_{t,i} - n_{\min}}{n_{\max} - n_{\min}}] \tag{5}$$

D. Training

Because the backbone is designed based on CNNs, the proposed model is trained by an end-to-end scheme. To reduce the

impact of the amount of data on the deep model, additionally, a new training method is proposed. Specifically, the input 3D skeleton model is randomly rotated around the x, y, and z coordinate axes, resulting in more new data. There will be a large number of challenging perspectives in these data. The training method proves effective for the small-size datasets. For the hyper-parameters, we follow the settings given in [21], and the SoftMax function is also used to obtain classification probabilities.

IV. EXPERIMENTS

A. Datasets and experimental settings

To give a comprehensive evaluation, two benchmark datasets are utilized to test the proposed method, including the NTU RGB-D dataset (NTU) and the SBU Kinect Interaction dataset (SBU). The NTU consists of 56880 video samples and contains 60 action classes, and then all samples are collected from various perspectives, which is a challenge for human action analysis [22]. Next, the SBU consists of 8 interactive actions and includes 282 sequences, which is mainly used for human interaction analysis [23].

The model is built under the Tensorflow framework, the Ubuntu system is utilized, the NVDIA GTX2060 GPU is used for training, and the batch size is set to 32. Both evaluation indexes are utilized on the NTU, including the cross-subject (CS) and the cross-view (CV). The main purpose is to reduce the impact of changes in observational perspective on action analysis, so the CV is used as the standard evaluation. The cross-subject evaluation is used on the SBU.

B. Ablation study

TABLE I
COMPARISON OF HAND-CRAFTED FEATURE METHODS

Method	SBU (%)	NTU (CV, %)
G-based	87.7	84.2
BSW-based	58.6	57.6
Strucutured-based	87.4	85.9
Proposed	**89.6**	**86.5**

TABLE II
COMPARISON OF DIFFERENT BACKBONE

Method	SBU (%)	NTU (CV, %)
CNN-based	82.7	81.4
RNN-based	84.3	82.8
3D-CNN-based	88.5	85.1
Proposed	**89.6**	**86.5**

To evaluate the effectiveness of proposed different components, a group of ablation study is conducted on the NTU and the SBU. As shown in Table 1, first, we compare the proposed skeleton representation transformation method with other popular hand-crafted feature methods. These methods are widely introduced in many works, such as [24], [25], and [26]. It can be seen that the adaptive representation learning method can improve the recognition performance in two benchmark datasets, which proves that the skeleton models collected under multiple perspectives can be transformed. Next, we introduce the proposed method into various popular backbones and discuss the results, and the results are shown in Table 2. It can be seen that the re-designed backbone achieves state-of-the-art recognition rates on both datasets, which proves that the high-dimensional feature processing network is effective for the learning of spatiotemporal features.

As shown in Fig. 3, the demonstration results of the skeleton transformation on the NTU are given. The second line is the transformed results, and the first line is the input of the model. It can be seen that the proposed method can imitate the habit of human vision to transform the skeleton model from different perspectives to the best perspective.

C. Comparison with other state-of-the-art works

NTU dataset. As shown in Table 3, the performance comparison between the proposed method and other state-of-the-art methods on the NTU dataset. For a fair comparison, we follow the settings in [22]. This dataset contains 60 action classes, including daily actions, medical conditions, and interactions, which is a challenge for human activity recognition. Thanks to the effectiveness of the skeleton representation transformation, the proposed model achieves the higher accuracy than other works, which is as high as 86.5%. The results also prove that the proposed method can adaptively learn a variety of complex perspective transformation methods. Note that nearly 80 collection perspectives are included in the NTU dataset.

TABLE III
RESULTS ON THE NTU DATASET

Method	NTU (CV, %)
3D-representing-based [27]	52.8
HBRNN-L-based [28]	64.0
P&C FW-AEC [29]	76.1
STA-LSTM [30]	81.2
GCA-LSTM [31]	82.8
URNN-2L-T-based [32]	83.2
TSA [33]	84.7
CNN-MTLN-based [26]	84.8
Proposed	**86.5**

TABLE IV
RESULTS ON THE SBU DATASET

Method	SBU (%)
Velocity features [23]	48.4
Plane features [23]	73.8
Joint features [23]	80.3
CFDM-based [24]	89.4
CHARM-based [34]	83.9
HBRNN-based [28]	80.4
Proposed	**89.6**

SBU dataset. Table 4 shows the comparison results of various state-of-the-art works. Although there are not too many complex data collected with various perspectives in this dataset, interactive action has always been a difficulty in behavior analysis. The SBU dataset consists of eight interactions,

Fig. 3. Demonstration results of skeleton representation transformation method.

including "Approach", "Leave", "Kick", "Punch", "Push", "Hug", "Shake hand", "Exchange item". 89.6% recognition accuracy is achieved, which is 41.2%, 15.8%, and 9.3% higher than hand-crafted feature-based methods proposed in [23].

V. Conclusion

In this work, an adaptive model is proposed to alleviate the effect of perspective variations to reach better robustness. The contributions of the paper are summarized as follows: 1) A skeleton representation transformation method is proposed to obtain high-performance representations by adaptive learning key parameters; 2) A 3D-CNN-based backbone is designed to model spatiotemporal features and output final classification results 3) A series of experiments are conducted on two benchmark datasets and the experimental results show that the proposed model has superior performance to other state-of-the-art models.

Group activity analysis, that is, more than two participants are included, will be studied in future work and the network structure will be further simplified.

Acknowledgment

The authors would like to acknowledge the support from the AiBle project co-financed by the European Regional Development Fund, National Key R&D Program of China (Grant No. 2018YFB1304600), CAS Interdisciplinary Innovation Team (Grant No. JCTD-2018-11), LiaoNing Province Higher Education Innovative Talents Program Support Project (Grant No. LR2019058), and National Natural Science Foundation of China (grant No. 52075530, 62006204, and 51575412).

References

[1] B. Liu, H. Cai, Z. Ju, and H. Liu, "Rgb-d sensing based human action and interaction analysis: A survey," *Pattern Recognition*, vol. 94, pp. 1–12, 2019.

[2] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4405–4425, 2017.

[3] S. A. Abu-Bakar, "Advances in human action recognition: an updated survey," *IET Image Processing*, vol. 13, no. 13, pp. 2381–2394, 2019.

[4] B. Liu, H. Cai, X. Ji, and H. Liu, "Human-human interaction recognition based on spatial and motion trend feature," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 4547–4551.

[5] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 9–14.

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

[7] B. Liu, H. Yu, X. Zhou, D. Tang, and H. Liu, "Combining 3d joints moving trend and geometry property for human action recognition," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2016, pp. 000 332–000 337.

[8] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2017, pp. 597–600.

[9] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1963–1978, 2019.

[10] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[11] J. Yu, H. Gao, W. Yang, Y. Jiang, W. Chin, N. Kubota, and Z. Ju, "A discriminative deep model with feature fusion and temporal attention for human action recognition," *IEEE Access*, vol. 8, pp. 43 243–43 255, 2020.

[12] Z. Tu, W. Xie, J. Dauwels, B. Li, and J. Yuan, "Semantic cues enhanced multimodality multistream cnn for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1423–1437, 2018.

[13] Z. Tu, W. Xie, Q. Qin, R. Poppe, R. C. Veltkamp, B. Li, and J. Yuan, "Multi-stream cnn: Learning representations based on human-related regions for action recognition," *Pattern Recognition*, vol. 79, pp. 32–43, 2018.

[14] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional lstm with cnn features," *IEEE access*, vol. 6, pp. 1155–1166, 2017.

[15] J. Li, X. Liu, W. Zhang, M. Zhang, J. Song, and N. Sebe, "Spatio-temporal attention networks for action recognition and detection," *IEEE Transactions on Multimedia*, vol. 22, no. 11, pp. 2990–3001, 2020.

[16] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1586–1599, 2017.

[17] M. Majd and R. Safabakhsh, "Correlational convolutional lstm for human action recognition," *Neurocomputing*, vol. 396, pp. 224–229, 2020.

[18] Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei, "Eidetic 3d lstm: A model for video prediction and beyond," in *International conference on learning representations*, 2018.

[19] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2015, pp. 579–583.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[22] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[23] 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 *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012, pp. 28–35.

[24] Y. Ji, H. Cheng, Y. Zheng, and H. Li, "Learning contrastive feature distribution model for interaction recognition," *Journal of Visual Communication and Image Representation*, vol. 33, pp. 340–349, 2015.

[25] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "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*, 2013, pp. 2752–2759.

[26] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3288–3297.

[27] R. Vemulapalli, F. Arrate, and R. Chellappa, "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*, 2014, pp. 588–595.

[28] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[29] K. Su, X. Liu, and E. Shlizerman, "Predict & cluster: Unsupervised skeleton based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9631–9640.

[30] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[31] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1647–1656.

[32] W. Li, L. Wen, M.-C. Chang, S. Nam Lim, and S. Lyu, "Adaptive rnn tree for large-scale human action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1444–1452.

[33] C. Caetano, J. Sena, F. Brémond, J. A. Dos Santos, and W. R. Schwartz, "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)*. IEEE, 2019, pp. 1–8.

[34] W. Li, L. Wen, M. C. Chuah, and S. Lyu, "Category-blind human action recognition: A practical recognition system," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4444–4452.