Skeleton Image Representation for 3D Action Recognition based on Tree Structure and Reference Joints

Carlos Caetano
Smart Sense Laboratory
Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
carlos.caetano@dcc.ufmg.br

François Brémond
INRIA
Sophia Antipolis
Valbonne, France
francois.bremond@inria.fr

William Robson Schwartz
Smart Sense Laboratory
Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
william@dcc.ufmg.br

***Abstract*—In the last years, the computer vision research community has studied on how to model temporal dynamics in videos to employ 3D human action recognition. To that end, two main baseline approaches have been researched: (i) Recurrent Neural Networks (RNNs) with Long-Short Term Memory (LSTM); and (ii) skeleton image representations used as input to a Convolutional Neural Network (CNN). Although RNN approaches present excellent results, such methods lack the ability to efficiently learn the spatial relations between the skeleton joints. On the other hand, the representations used to feed CNN approaches present the advantage of having the natural ability of learning structural information from 2D arrays (i.e., they learn spatial relations from the skeleton joints). To further improve such representations, we introduce the Tree Structure Reference Joints Image (TSRJI), a novel skeleton image representation to be used as input to CNNs. The proposed representation has the advantage of combining the use of reference joints and a tree structure skeleton. While the former incorporates different spatial relationships between the joints, the latter preserves important spatial relations by traversing a skeleton tree with a depth-first order algorithm. Experimental results demonstrate the effectiveness of the proposed representation for 3D action recognition on two datasets achieving state-of-the-art results on the recent NTU RGB+D 120 dataset.**

I. Introduction

Human action recognition plays an important role in various applications such as surveillance systems, health care systems and robot and human-computer interaction. Significant progress on the action recognition task has been achieved due to the design of discriminative representations based on appearance information by using RGB frames. However, due to the development of cost-effective RGB-D sensors (e.g., Kinect), it became possible to employ different types of data such as depth information as well as human skeleton joints to perform 3D action recognition. Compared to RGB or depth information, skeleton based methods have demonstrated impressive results by modeling temporal dynamics in videos. These approaches have the advantage of being computationally efficient due to smaller data size and being robust to illumination changes, background noise and invariance to camera views [1].

On the last decade, many works for 3D action recognition model temporal dynamics in videos by employing Dynamic Time Warping (DTW), Fourier Temporal Pyramid (FTP) or Hidden Markov Model (HMM) in conjunction with skeleton handcrafted feature descriptors [2]–[7]. Nowadays, large efforts have been directed to the employment of deep neural networks to model skeleton data by using two main approaches: (i) Recurrent Neural Networks (RNNs) with Long-Short Term Memory (LSTM) [8]–[11]; and (ii) skeleton image representations used as input to a Convolutional Neural Network (CNN) [12]–[20]. Although the former approach present excellent results in 3D action recognition task due to their power of modeling temporal sequences, such structures lack the ability to efficiently learn the spatial relations between the skeleton joints [18]. On the other hand, the latter takes the advantage of having the natural ability of learning structural information from 2D arrays and is able to learn spatial relations from the skeleton joints.

As the forerunner of skeleton image representations, Du et al. [12] take advantage of the spatial relations by employing a hierarchical structure. The authors represent each skeleton sequence as 2D arrays, in which the temporal dynamics of the sequence is encoded as variations in columns and the spatial structure of each frame is represented as rows. Finally, the representation is fed to a CNN to perform action prediction. Such type of representations is very compact since it encodes the entire video sequence in a single image.

In this paper, we introduce a novel skeleton image representation, named *Tree Structure Reference Joints Image (TSRJI)*, to be used as input for CNNs. We improve the representation of skeleton joints for 3D action recognition encoding temporal dynamics by combining the use of reference joints [15] and a tree structure skeleton [18]. The method takes advantage of a structural organization of joints that preserves spatial relations of more relevant joint pairs and also by incorporating different spatial relationships between the joints. To perform action classification, we train a small CNN architecture with only three convolutional layers and two fully-connected layers.

Since the network is shallow and takes as input a compact representation for each video, it is extremely fast to train.

Our hypothesis is based on the assumption that the rearrangement of the structural organization of joints to be used as inputs helps on guiding the network to extract certain information, possibly complementary, that would not be extracted by using other modalities, such as RGB or depth information. Aligned with our hypothesis, other works mention that although the temporal evolution patterns can be learned implicitly with CNN using RGB data, an explicit modeling is preferable [19].

According to the experimental results, our proposed skeleton image representation can handle skeleton based 3D action recognition very well being able to recognize actions accurately on two well-known large scale datasets (NTU RGB+D 60 [9] and NTU RGB+D 120 [21]). We achieve the state-of-the-art performance on the large scale NTU RGB+D 120 [21] dataset. Moreover, we show that our approach can be combined with a temporal structural joint representation [19] to obtain state-of-the-art performance (up to 3.3 percentage points when compared to the best skeleton based method reported to date).

The code of our TSRJI representation is publicly available to facilitate future research[1].

II. Related Work

In this section, we present a literature review of works that employ 3D action recognition based on skeleton image representations in conjunction with CNNs.

As one of the earliest works on skeleton image representations, Du et al. [12] represent the skeleton sequences as a matrix. Each row of such matrix corresponds to a chain of concatenated skeleton joint coordinates from the frame t. Hence, each column of the matrix corresponds to the temporal evolution of the joint j. At this point, the matrix size is $J \times T \times 3$, where J is the number of joints for each skeleton, T is the total frame number of the video sequence and 3 is the number coordinate axes (x, y, z). The values of this matrix are quantified into an image (i.e., linearly rescaled to a $[0, 255]$) and normalized to handle the variable-length problem. In this way, the temporal dynamics of the skeleton sequence is encoded as variations in rows and the spatial structure of each frame is represented as columns. Finally, the authors use their representation as input to a CNN model composed by four convolutional layers and three max-pooling layers. After the feature extraction, a feed-forward neural network with two fully-connected layers is employed for classification.

Wang et al. [13], [17] present a skeleton representation to represent both spatial configuration and dynamics of joint trajectories into three texture images through color encoding, named Joint Trajectory Maps (JTMs). The authors apply rotations to the skeleton data to mimicking multi-views and also for data enlargement to overcome the drawback of CNNs usually being not view invariant. JTMs are generated by projecting the trajectories onto the three orthogonal planes. To encode motion direction in the JTM, they use a hue colormap function to "color" the joint trajectories over the action period. They also encode the motion magnitude of joints into saturation and brightness claiming that changes in motion results in texture in the JMTs. Finally, the authors individually fine-tune three AlexNet [22] CNNs (one for each JTM) to perform classification.

Representations based on heat map to encode spatialtemporal skeleton joints were also proposed by Liu et al. [14]. Their approach considers each joint as 5D point space (x, y, z, t, j) and expresses them as a 2D coordinate space on a 3D color space. Thus, they permute elements of the 5D point space. Nonetheless, such permutation can generate very similar representations which may contain redundant information. To that end, they use ten types of ranking to ensure that each element of the point (x, y, z, t, j) can be assigned to the color 2D coordinate space. After that, the ten skeleton representations are quantified and treated as a color image. Finally, the authors employ a multiple CNN-based model, one for each of the representations. They used the AlexNet [22] architecture and fused the posterior probabilities generated from each CNN for the final class score.

To overcome the problem of the sparse data generated by skeleton sequence video, Ke et al. [15] represent the temporal dynamics of the skeleton sequence by generating four skeleton representation images. Their approach is closer to Du et al. [12] method, however they compute the relative positions of the joints to four reference joints by arranging them as a chain and concatenating the joints of each body part to the reference joints resulting onto four different skeleton representations. According to the authors, such structure incorporates different spatial relationships between the joints. Finally, the skeleton images are resized and each channel of the four representations is used as input to a VGG19 [23] pre-trained architecture for feature extraction.

To encode motion information on skeleton image representation, Li et al. [16], [19] proposed the skeleton motion image. Their approach is created similar to Du et al. [12] skeleton image representation, however each matrix cell is composed by joint difference computation between two consecutive frames. To perform classification, the authors used Du et al. [12] approach and their proposed representation independently as input of a neural network with a two-stream paradigm. The CNN used was a small seven-layer network consisting of three convolution layers and four fully-connected layers.

Yang et al. [18] claim that the concatenation process of chaining all joints with a fixed order turns into lack of semantic meaning and leads to loss in skeleton structural information. To that end, Yang et al. [18] proposed a representation named Tree Structure Skeleton Image (TSSI) to preserve spatial relations. Their method is created by traversing a skeleton tree with a depth-first order algorithm with the premise that the fewer edges there are, the more relevant the joint pair is. The generated representation is then quantified into an image and resized before being presented to a ResNet-50 [24] CNN

[1]https://github.com/carloscaetano/skeleton-images

architecture.

As it can be seen from the reviewed methods, most of them are improved versions of Du et al. [12] skeleton image representation focusing on spatial structural of joint axes while the temporal dynamics of the sequence is encoded as variations in columns. Despite the aforementioned methods produce promising results, we believe that performance can be improved by explicitly employing joints relationships, which enhances the temporal dynamics encoding. In view of that, our approach takes advantage of combining a structural organization that preserves spatial relations of more relevant joint pairs by using the skeleton tree with a depth-first order algorithm from Yang et al. [18] and also by incorporating different spatial relationships by using the reference joints technique from Ke et al. [15].

III. Proposed Approach

In this section, we introduce our proposed skeleton image representation based on reference joints and a tree structure skeleton, named Tree Structure Reference Joints Image (TSRJI). Finally, we present the CNN architecture employed in our approach.

A. Tree Structure Reference Joints Image (TSRJI)

As reviewed in Section II, a crucial step to achieve good performance using skeleton image representations is to define how to build the structural organization of the representation preserving the spatial relations of relevant joint pairs. In view of that and due to the successful results achieved by the skeleton image representations, our approach follows the same fundamentals by representing the skeleton sequences as a matrix. Furthermore, our method is based on two premises of successful representations of the literature: (i) the fewer edges there are, the more relevant the joint pair is [18]; and (ii) different spatial relationships between the joints leads to less sparse data [15].

To address the first premise, we apply the depth-first tree traversal order [18] to each skeleton data from frame t to generate a pre-defined chain order C^t that best preserves the spatial relations between joints in original skeleton structures (see Figure 1). The basic assumption here is that the spatially related joints in original skeletons have direct graph links between them [18]. The less edges required to connect a pair of joints, the more related is the pair. In view of that, with the C^t chain order, the neighboring columns in skeleton images are spatially related in original skeleton structures.

To address the second premise, we apply the reference joints technique [15] to each generated C^t chain. To that end, four reference joints are respectively used to compute relative positions of the other joints: (a) the left shoulder; (b) the right shoulder; (c) the left hip; and (d) the right hip. Thus, at this point, we have four C chains for each skeleton of each frame (i.e., C_a^t, C_b^t, C_c^t, C_d^t). The hypothesis here, introduced by Ke et al. [15], is that relative positions between joints provide more useful information than their absolute locations. According to Ke et al. [15], these four joints are selected as

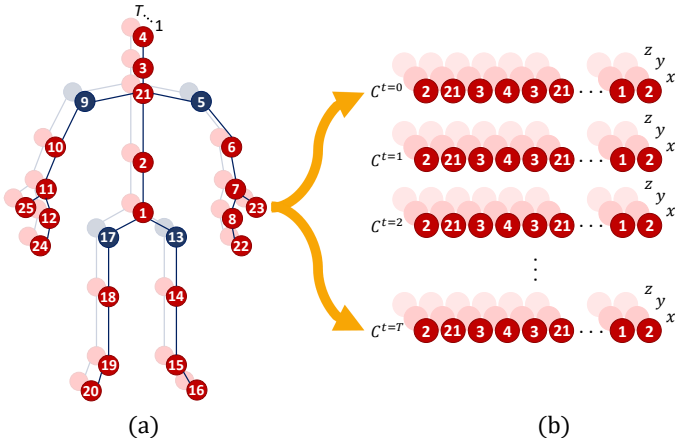

Fig. 1. Depth-first tree traversal order applied to skeleton data. (a) Skeleton data sequence of T frames. (b) Chains C^t considering 25 Kinect joints: [2, 21, 3, 4, 3, 21, 5, 6, 7, 8, 22, 23, 22, 8, 7, 6, 5, 21, 9, 10, 11, 12, 24, 25, 24, 12, 11, 10, 9, 21, 2, 1, 13, 14, 15, 16, 15, 14, 13, 1, 17, 18, 19, 20, 19, 18, 17, 1, 2], as defined in [18].

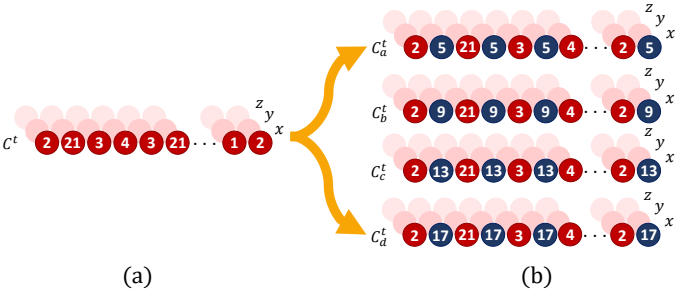

Fig. 2. Reference joints technique applied to skeleton data. (a) Chain C^t considering 25 Kinect joints. (b) Generated chains C_a^t, C_b^t, C_c^t, C_d^t considering the reference joints (dark painted joints).

reference joints due to the fact that they are stable in most actions, thus reflecting the motions of the other joints. Figure 2 illustrates the reference joints technique computation.

After dealing with the aforementioned premises, we compute four matrices S (one for each reference joint) that correspond to the concatenation of the chains C^t from a video (i.e., S_a, S_b, S_c, S_d), where each column of each matrix denotes the temporal evolution of the arranged chain joint c. At this point, the size of matrix S is $J \times T \times 3$, where J is the number of joints of the any reference joint chain C^t, T is the total frame number of the video sequence and 3 is the number joint coordinate axes (x, y, z).

Finally, the generated matrices are normalized into [0, 1] and empirically resized into a fixed size of $J \times 100$ to be used as input to CNNs, since number of frames may vary depending on the skeleton sequence of each video. Figure 3 gives an overview of our method for building the skeleton image representation.

B. Convolutional Neural Network Architecture Employed

To learn the features of the generated skeleton image representations, we adopted a modified version of the CNN architecture proposed by Li et al. [16]. They designed a small

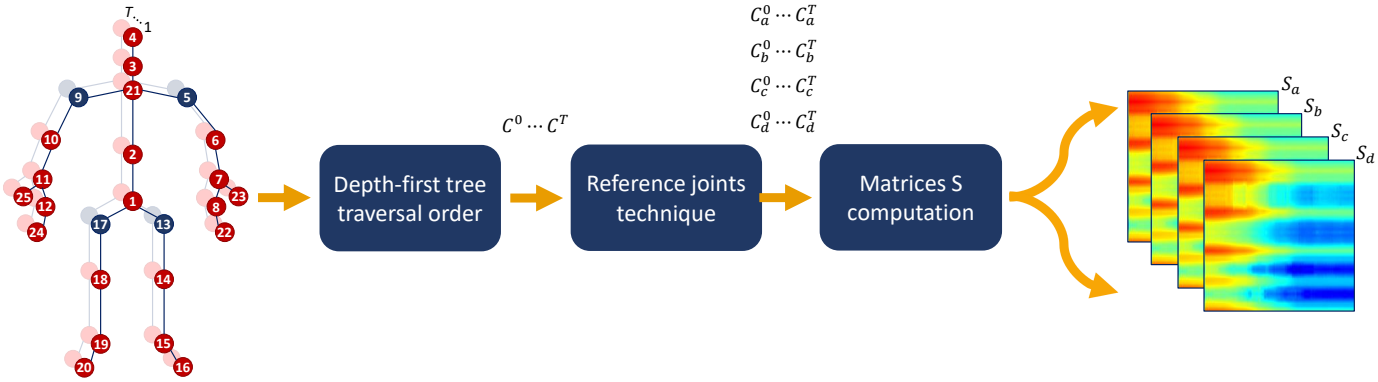

Fig. 3. Proposed skeleton image representation.

convolutional neural network which consists of three convolution layers and four fully-connected (FC) layers. However, we modified it to a tiny version, employing the convolutional layers and only two FC layers. All convolutions have a kernel size of 3×3, the first and second convolutional layers with a stride of 1 and the third one with a stride of 2. Max pooling and ReLU neuron are adopted and the dropout regularization ratio. We opted for using such architecture since it demonstrated good performance and, according to the authors, it can be easily trained from scratch without any pre-training and is superior on its compact model size and fast inference speed as well. Figure 4 presents an overview of the employed architecture.

To cope with actions involving multi-person interaction (e.g., shaking hands), we apply a common choice in the literature which is to stack skeleton image representations of different people as the network input.

IV. Experimental Results

In this section, we present the experimental results obtained with the proposed Tree Structure Reference Joints Image (TSRJI) for the 3D action recognition problem. To prove that a good structural organization of joints is important to preserve the spatial relations of the skeleton data, we compare our approach with a baseline employing random joints order when creating the representation (i.e., the creation of the chains' order C^t does not take into account any semantic meaning of adjacent joints). Moreover, we also compare with the classical skeleton image representations used by state-of-the-art approaches [12], [13], [15], [16], [18], [19] as well as to estate-of-the-art methods on the NTU RGB+D 120 [21].

A. Datasets

1) NTU RGB+D 60 [9]: it is a publicly available 3D action recognition dataset consisting of 56,880 videos from 60 action categories which are performed by 40 distinct subjects. The videos were collected by three Microsoft Kinect sensors. The dataset provides four different data information: (i) RGB frames; (ii) depth maps; (iii) 395 infrared sequences; and (iv) skeleton joints. There are two different evaluation protocols: cross-subject, which split the 40 subjects into training and testing; and cross-view, which uses samples from one camera for testing and the other two for training. The performance is evaluated by computing the average recognition across all classes.

2) NTU RGB+D 120 [21]: is the most recent large-scale 3D action recognition dataset captured under various environmental conditions and consists of 114,480 RGB+D video samples captured using the Microsoft Kinect sensor. As in NTU RGB+D 60 [9], the dataset provides RGB frames, depth maps, infrared sequences and skeleton joints. It is composed by 120 action categories performed by 106 distinct subjects in a wide range of age distribution. There are two different evaluation protocols: cross-subject, which split the 106 subjects into training and testing; and cross-setup, which divides samples with even setup IDs for training (16 setups) and odd setup IDs for testing (16 setups). The performance is evaluated by computing the average recognition across all classes.

B. Implementation Details

To isolate only the contribution brought by the proposed representation to the action recognition problem, all compared skeleton image representations were implemented and tested on the same datasets and using the same network architecture. We also applied the same split of training and testing data and employed the evaluation protocols and metrics proposed by the creators of the datasets.

For the network architecture employed, we used a dropout regularization ratio set to 0.5. The learning rate is set to 0.001 and batch size is set to 1000.

C. Evaluation

In this section, we present experiments for our proposed TSRJI representation and report a comparison with skeleton images baselines and methods of the literature.

Table I presents a comparison of our approach with skeleton image representations of the literature. For the methods that have more than one "image" per representation ([13] and [15]), we stacked them to be used as input to the network. The same was performed for our TSRJI (Stacked) approach considering the images for each reference joint (i.e., S_a,

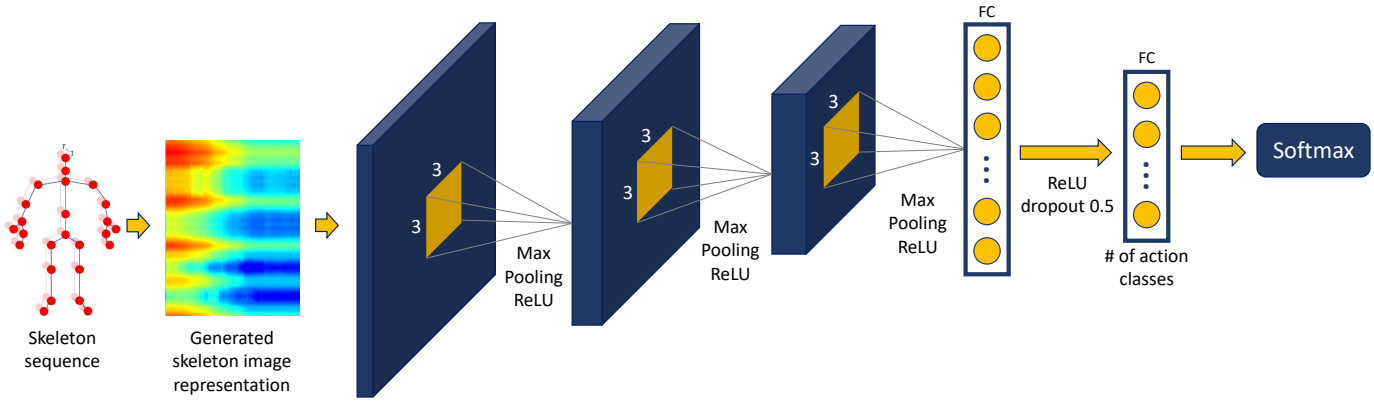

Fig. 4. Network architecture employed for 3D action recognition.

TABLE I
ACTION RECOGNITION ACCURACY (%) RESULTS ON NTU RGB+D 60 [9] DATASET. RESULTS FOR THE BASELINES WERE OBTAINED RUNNING EACH METHOD IMPLEMENTATION.

	Approach	Cross-subject Acc. (%)	Cross-view Acc. (%)
Baseline results	Random joints order	67.8	74.2
	Du et al. [12]	68.7	73.0
	Wang et al. [13]	39.1	35.9
	Ke et al. [15]	**70.8**	75.5
	Li et al. [19]	56.8	61.3
	Yang et al. [18]	69.5	**75.6**
Our results	TSRJI (Stacked)	69.3	76.7
	TSRJI (Late Fusion)	**73.3**	**80.3**

S_b, S_c, S_d). Regarding the cross-subject protocol, the best results were obtained by Reference Joints technique from Ke et al. [15] achieving 70.8% of accuracy and the Tree Structure Skeleton Image (TSSI) from Yang et al., [18] achieving 70.8% of accuracy 69.5%. However, it is worth noting that we achieved a close competitive accuracy of 69.3% with our TSRJI (Stacked) approach. On the other side, the best result on cross-view protocol was obtained by our TSRJI (Stacked) approach achieving 76.7% of accuracy. Compared to Ke et al. [15], we achieved an improvement of 1.2 percentage points (p.p.). Moreover, there is an improvement of 1.1 p.p. when compared to the Tree Structure Skeleton Image (TSSI) from Yang et al., [18], which was the best baseline result on this protocol. Detailed improvements are shown in Figure 5.

Comparing to the random joints order baseline (Table I), it is worth noting an improvement of 1.5 p.p. on cross-subject protocol and 1.5 p.p. on cross-view protocol obtained by our TSRJI (Stacked). This shows the importance of keeping a structural organization of joints that preserves spatial relations of relevant joint pairs, bringing semantic meaning of adjacent joints to the representation.

We also employed experiments by employing a late fusion technique with our proposed skeleton image representation. To that end, each reference isolate joint image S is used as input to a CNN. The late fusion technique applied was a non-weighted linear combination of the prediction scores generated by each CNN output. Table I presents a comparison of our TSRJI (Late Fusion) with skeleton image representations of the literature. Here, our proposed representation achieved the best results in both protocols of the NTU RGB+D 60 [9] dataset. We achieved 73.3% of accuracy on cross-subject protocol, with an improvement of 2.5 p.p over the best baseline method (Ke et al. [15]). Furthermore, we achieved an accuracy of 80.3% on the cross-view protocol with an improvement of 4.7 p.p. when compared to Yang et al., [18].

Finally, Table II presents the experiments of our proposed skeleton image representation on the recent proposed NTU RGB+D 120 [21] dataset. Based on the results achieved in Table I, we employed the late fusion scheme for our approach.

According to Table II, we achieved good results with our TSRJI (Late Fusion) representation outperforming many skeleton based methods [9], [15], [25]–[34]. We achieved state-of-the-art results, outperforming the best reported method (Body Pose Evolution Map [34]) on cross-subject protocol (accuracy of 65.5%). On the other hand, the best result on cross-setup protocol is obtained by Liu et al. [34] achieving 66.9 of accuracy.

To exploit a possible complementarity of the temporal (motion) and spatial skeleton image representations, we employed the late fusion combination scheme of our approach and Li et al. [19] method that explicitly provides motion information on the representation. With such combination, we achieved state-of-the-art results outperforming the best reported method (Body Pose Evolution Map [34]) by up to 3.3 p.p. on cross-subject protocol.

In comparison with LSTM approaches, we outperform the best reported method (Two-Stream Attention LSTM [32]) by 4.3 p.p. using our TSRJI representation and 6.7 p.p. when combining it with Li et al. [19] method on cross-subject protocol. Regarding the cross-setup protocol, we obtained similar comparative accuracy (62.8) using our TSRJI fused

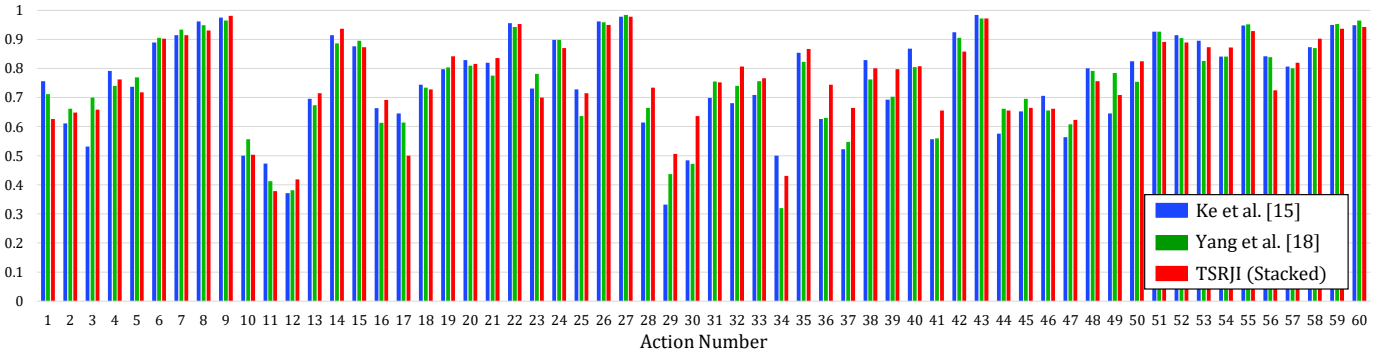

Fig. 5. Comparison of TSRJI (Stacked) with Ke et al. [15] and Yang et al. [18] on NTU RGB+D 60 [9] dataset for cross-view protocol. Best viewed in color.

TABLE II
ACTION RECOGNITION ACCURACY (%) RESULTS ON NTU RGB+D 120 [21] DATASET. RESULTS FOR LITERATURE METHODS WERE OBTAINED FROM [21].

	Approach	Cross-subject Acc. (%)	Cross-setup Acc. (%)
Literature results	Part-Aware LSTM [9]	25.5	26.3
	Soft RNN [25]	36.3	44.9
	Dynamic Skeleton [26]	50.8	54.7
	Spatio-Temporal LSTM [27]	55.7	57.9
	Internal Feature Fusion [28]	58.2	60.9
	GCA-LSTM [29]	58.3	59.2
	Multi-Task Learning Network [15]	58.4	57.9
	FSNet [30]	59.9	62.4
	Skeleton Visualization (Single Stream) [31]	60.3	63.2
	Two-Stream Attention LSTM [32]	61.2	63.3
	Multi-Task CNN with RotClips [33]	62.2	61.8
	Body Pose Evolution Map [34]	**64.6**	**66.9**
Our results	TSRJI (Late Fusion)	**65.5**	59.7
	TSRJI (Late Fusion) + Li et al. [19]	**67.9**	62.8

with Li et al. [19]. This indicates that our skeleton image representation approach used as input for CNNs leads to a better learning of joint spatial relations than the approaches that employs LSTM.

D. Discussion

Since our proposed TSRJI representation is based on the combination of the tree structural organization from Yang et al. [18] and the reference joints technique from Ke et al. [15], we better analyze our achieved results by taking a closer look at actions from NTU RGB+D dataset that our method achieved higher performance than Ke et al. [15] and Yang et al. [18]. Figure 5, presents the detailed improvements of our TSRJI (Stacked) representation. The actions that were most correctly classified by TSRJI (Stacked) and misclassified by the baselines are: *standing up (9); writing (12); tear up paper (13); wear jacket (14); wear a shoe (16); take off glasses (19); take off a hat cap (21); make a phone call (28); playing with phone (29); typing on a keyboard (30); taking a selfie (32); check time (33); nod head bow (35); shake head (36); wipe face (37); sneeze or cough (41); point finger at the other person (54); touch other person pocket (57); and handshaking (58)*[2]. We note that the baselines usually confused such actions, which are actions involving arm and hand movements.

We also analyze our achieved results with the employed late fusion scheme. To better perform such comparison, we combined Ke et al. [15] and Yang et al. [18] representations with the same late fusion scheme employed by us. Figure 6, presents the detailed improvements of our TSRJI (Late Fusion) representation. For instance, some actions that were most correctly classified by TSRJI (Late Fusion) and misclassified by the baseline are: *brushing teeth (3); make a phone call (28); playing with phone (29); typing on a keyboard (30); wipe face (37); and sneeze or cough (41)*. We note that the baseline confused actions involving arm and hand movements. It shows that our proposed representation performs better and provides a richer discriminability than a simply combination of the based methods.

The correctly classifications of the aforementioned actions by our TSRJI representation show that feeding the network with explicit structural organization of relevant joint pairs might improve the classification. We believe that the reference

[2]The number in parentheses represents the action index.

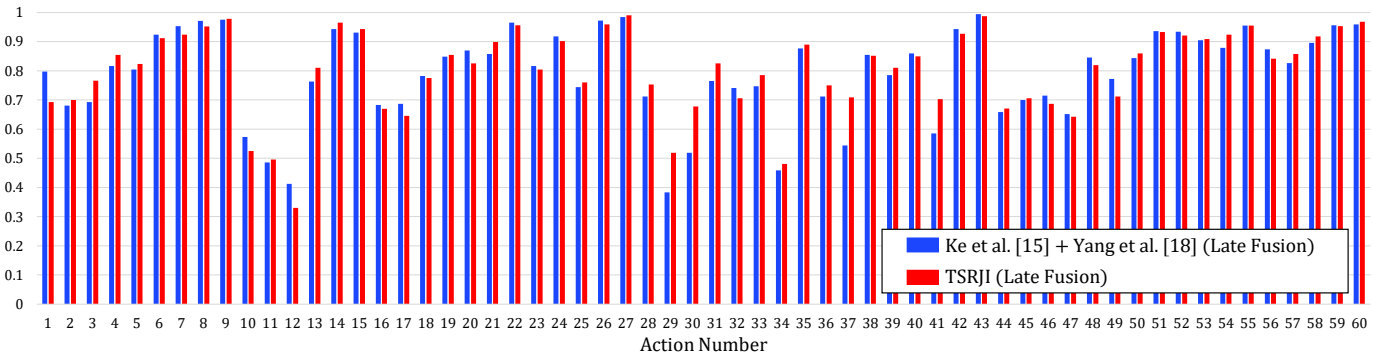

Fig. 6. Comparison of TSRJI (Late Fusion) with Ke et al. [15] + Yang et al. [18] (Late Fusion) on NTU RGB+D 60 [9] dataset for cross-view protocol. Best viewed in color.

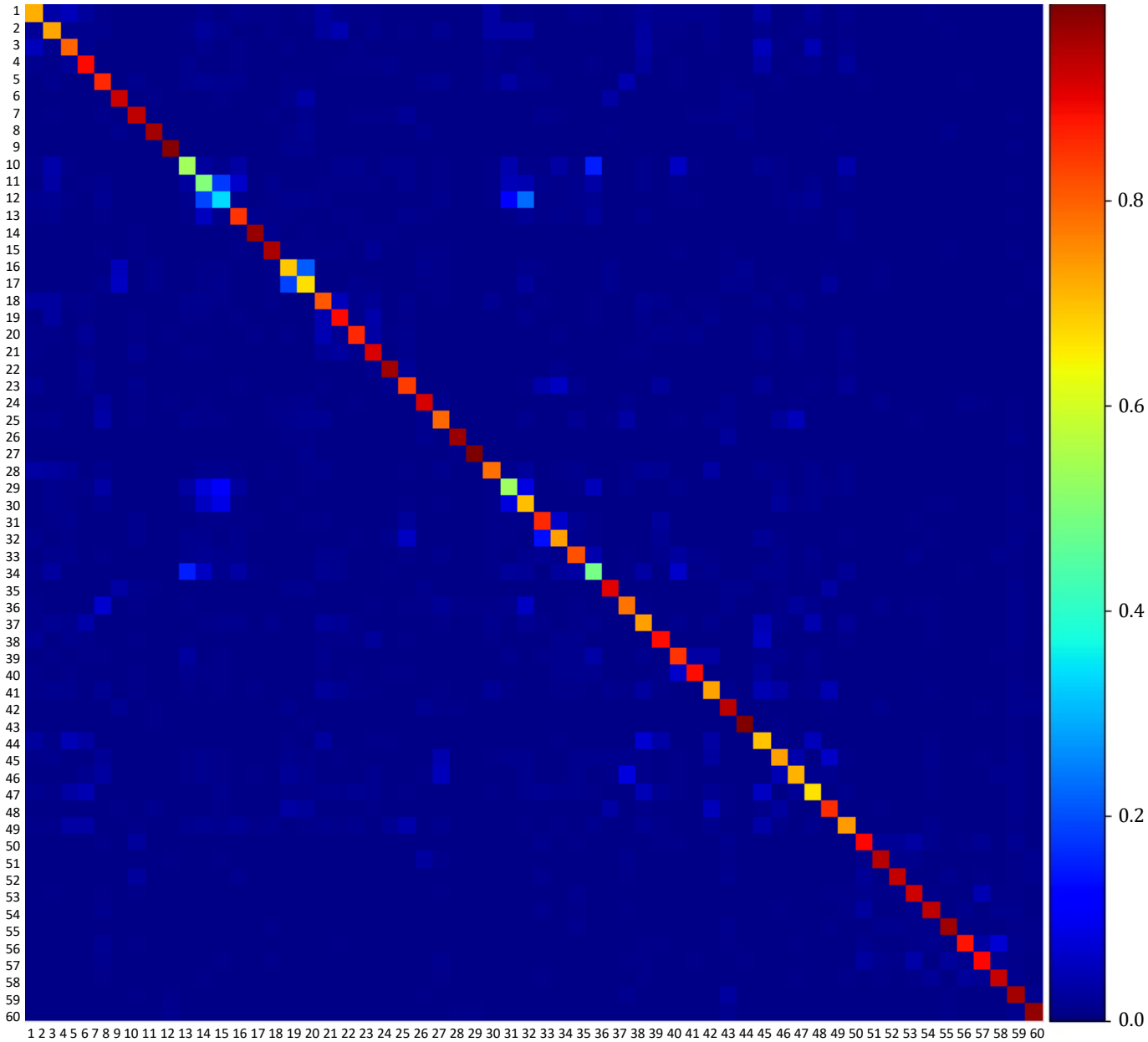

Fig. 7. Confusion matrix of TSRJI (Late Fusion) on NTU RGB+D 60 [9] dataset. Best viewed in color.

joints technique helped on improving such actions since the shoulders were two of the reference joints. Thus, since such joints are stable they could reflect the motions of the arms and hand joints. Furthermore, the spatial relations of adjacent joint pairs were preserved by the use of the depth-first tree traversal order algorithm bringing more semantic meaning to the representation.

We also investigated the cases where our method failed. The most misclassified actions correspond to cases, such as *clapping (10), rub two hands together (34), reading (11), writing (12), typing on a keyboard (30), wear a shoe (16) and take off a shoe (17)*. Our method confused *clapping (10)* with *rub two hands together (34)*, in which both actions are composed by closer movements with the hands. Furthermore, the analysis of the misclassified videos revealed that the method presented difficulties with actions with very similar movements differentiating by the object used (e.g., the action *writing (12)* is confused with *reading (11)*, *typing on a keyboard (30)* and *playing with phone (29)*). Another misclassification of our approach is *wear a shoe (16)* with *take off a shoe (17)*. Such analysis indicates that the use of explicit motion information could help enhancing the classification. Figure 7 illustrates the confusion matrix of our TSRJI representation.

V. Conclusions and Future Works

In this work, we proposed a novel skeleton image representation to be used as input of CNNs. The method takes advantage of a structural organization of joints that preserves spatial relations of more relevant joint pairs and also by incorporating different spatial relationships between the joints. Experimental results on two publicly available datasets demonstrated the excellent performance of the proposed approach. Another interesting finding is that the combination of our representation with explicitly motion method of the literature improves the 3D action recognition outperforming the state-of-the-art on NTU RGB+D 120 dataset.

Directions to future works include the evaluation of the proposed representation with other distinct architectures. Moreover, we intend to evaluate its behavior on 2D action datasets with skeletons estimated by methods of the literature.

Acknowledgments

The authors would like to thank the National Council for Scientific and Technological Development – CNPq (Grants 311053/2016-5, 204952/2017-4 and 438629/2018-3), the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00540-17) and the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project).

References

[1] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3d skeletal data," *CVIU*, 2017.

[2] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*, 2012.

[3] X. Yang and Y. L. Tian, "Eigenjoints-based action recognition using nave-bayes-nearest-neighbor," in *CVPRW*, 2012.

[4] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection," in *ICCV*, 2013.

[5] M. A. Gowayyed, M. Torki, M. E. Hussein, and M. El-Saban, "Histogram of oriented displacements (hod): Describing trajectories of human joints for action recognition," in *IJCAI*, 2013.

[6] P. Wang, W. Li, P. Ogunbona, Z. Gao, and H. Zhang, "Mining mid-level features for action recognition based on effective skeleton representation," in *DICTA*, 2014.

[7] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-d human action recognition by shape analysis of motion trajectories on riemannian manifold," *IEEE Transactions on Cybernetics*, 2015.

[8] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *ICCV*, 2015.

[9] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *CVPR*, 2016.

[10] 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 *AAAI Conference on Artificial Intelligence*, 2017.

[11] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *ICCV*, 2017.

[12] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *ACPR*, 2015.

[13] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *MM*, 2016.

[14] M. Liu, C. Chen, and H. Liu, "3d action recognition using data visualization and convolutional neural networks," in *ICME*, 2017.

[15] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *CVPR*, 2017.

[16] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *ICMEW*, 2017.

[17] P. Wang, W. Li, C. Li, and Y. Hou, "Action recognition based on joint trajectory maps with convolutional neural networks," *Knowledge-Based Systems*, 2018.

[18] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with spatio-temporal visual attention on skeleton image sequences," *TCSVT*, 2018.

[19] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *IJCAI*, 2018.

[20] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "Potion: Pose motion representation for action recognition," in *CVPR*, 2018.

[21] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding," *TPAMI*, 2019.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[25] J. Hu, W. Zheng, L. Ma, G. Wang, J. Lai, and J. Zhang, "Early action prediction by soft regression," *TPAMI*, 2018.

[26] J. Hu, W. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," *TPAMI*, 2017.

[27] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *ECCV*, 2016.

[28] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," *TPAMI*, 2018.

[29] J. Liu, G. Wang, P. Hu, L. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *CVPR*, 2017.

[30] J. Liu, A. Shahroudy, G. Wang, L. Duan, and A. Kot Chichung, "Skeleton-based online action prediction using scale selection network," *TPAMI*, 2019.

[31] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recogn.*, 2017.

[32] J. Liu, G. Wang, L. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks," *TIP*, 2018.

[33] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3d action recognition," *TIP*, 2018.

[34] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *CVPR*, 2018.