Motion-Based Representations For Activity Recognition

Carlos Caetano, Jefersson A. dos Santos, William Robson Schwartz
Smart Sense Laboratory, Department of Computer Science
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
{carlos.caetano, jefersson, william}@dcc.ufmg.br

***Abstract*—This work[1] addresses the activity recognition problem. We propose two different representations based on motion information for activity recognition. The first representation is a novel temporal stream for two-stream Convolutional Neural Networks (CNNs) that receives as input images computed from the optical flow magnitude and orientation to learn the motion in a better and richer manner. The method applies simple non-linear transformations on the vertical and horizontal components of the optical flow to generate input images for the temporal stream. The second representation is a novel skeleton image representation to be used as input of CNNs. The approach encodes the temporal dynamics by explicitly computing the magnitude and orientation values of the skeleton joints. Experiments carried out on challenging well-known activity recognition datasets (UCF101, NTU RGB+D 60 and NTU RGB+D 120) demonstrate that the proposed representations achieve results in the state of the art, indicating the suitability of our approaches as video representations.**

I. Introduction

Human activity recognition has been used in many real-world applications, such as surveillance systems, video retrieval systems and health care. Over the last decade, a significant portion of the progress on the activity recognition task has been achieved with the design of discriminative representations known as handcrafted feature descriptors employed with a machine learning classifier. Moreover, with the development of cost-effective RGB-D sensors (e.g., Kinect), it became possible to employ different types of data such as human skeleton joints to perform 3D activity recognition. Nowadays, large efforts have been dedicated to the employment of deep Convolutional Neural Networks (CNNs) as representation learning. These approaches learn hierarchical layers of representations to perform pattern recognition and have achieved effective results on the activity recognition task [2]–[6].

The temporal component of videos provides an important clue for activity recognition, as a number of activities can be reliably recognized based on motion information. Hence, a significant portion of the progress on the activity recognition task has been achieved with the design of discriminative representations exploring temporal information based on motion analysis [2], [4], [5]. In view of that, in this work we explore the use of motion information based on optical flow and the motion extracted from skeleton joints to compute our representations. Our representations are based on the assumption that the motion information on a video sequence can be described by the spatial relationship contained on the local neighborhood of magnitude and orientation extracted from the optical flow or from skeleton information. More specifically, we assume that the motion information is adequately specified by fields of magnitude and orientation.

Considering the image-domain, our first representation is a novel feeding scheme for CNNs based on images computed from the optical flow to learn the motion in a better and richer manner, named Magnitude-Orientation Stream (MOS). The method applies simple nonlinear transformations (magnitude and orientation) on the vertical and horizontal components of the optical flow to generate the input images. Regarding the skeleton-domain, we propose the SkeleMotion representation. The proposed approach encodes temporal dynamics by explicitly using motion information computing the magnitude and orientation values of the skeleton joints. Moreover, different temporal scales are used to aggregate more temporal dynamics to the representation making it able to capture long-range joint interactions involved in activities.

II. Proposed Approaches

A. Image Domain-based Approach

1) Magnitude-Orientation Stream (MOS): The two-stream network is composed of two different networks receiving distinct flows of data, spatial and temporal. The spatial stream receives as input the RGB frames while the temporal stream receives optical flow images as input.

The process for computing the optical flow images is the following. For each frame F at time t, the optical flow O_t is computed considering F_t and F_{t+1}. The resulting optical flow O_t is composed of two channels: (i) $\mathcal{O}_t^x$, denoting an image containing the x (horizontal) displacement field; and (ii) $\mathcal{O}_t^y$, denoting an image containing the y (vertical) displacement field. Moreover, to avoid storing the displacement fields as floats, the horizontal and vertical components of the flow are linearly rescaled to a [0, 255].

Our MOS follows the same fundamentals as the two-stream networks. However, aiming at extracting more information from the optical flow, MOS captures the displacement information by using the orientation of the optical flow and the velocity of the movement considering the optical flow magnitude. The spatial relationship contained on local neighborhoods of magnitude and orientation captures not only displacement by using orientation, but also magnitude, providing information

[1] This work corresponds to a PhD dissertation [1].

regarding the velocity of the movement. The method is based on non-linear transformations on the optical flow components to generate input images for the temporal stream.

To incorporate such information on the temporal stream, we compute the dense optical flow as Wang et al. [7]. For each video composed of n frames, we compute $n-1$ optical flows $\mathcal{O}$. Once the optical flow is available, we compute the magnitude and orientation information as $M_{i,j} = \sqrt{(\mathcal{O}^x_{i,j})^2 + (O^y_{i,j})^2}$ and $\theta_{i,j} = tan^{-1}\left(\frac{\mathcal{O}^y_{i,j}}{\mathcal{O}^x_{i,j}}\right)$, where M and θ are magnitude and orientation information, respectively.

Since the values obtained in M and θ are composed of real numbers, they are linearly rescaled to a [0, 255]. Moreover, since the orientation values are estimated for every pixel of the optical flow, they can generate noisy values from regions of the image without any movement. Therefore, we perform a filtering on θ based on the values of M as

$$\theta^{'}_{i,j} = \left\{ \begin{array}{ll} 0, & \text{if } M_{i,j} < m \\ \theta_{i,j}, & \text{otherwise} \end{array} \right. ,$$

where m is a magnitude threshold value.

With the rescaled magnitude and orientation information, which can be seen as two image channels, it can be used as input to CNNs.

B. Skeleton Domain-based Approach

1) SkeleMotion: As the forerunner of skeleton image representations, Du et al. [3] 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 $[0, 255]$). 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.

Motivated by our MOS approach, we propose a novel skeleton image representation (named SkeleMotion), based on magnitude and orientation of the joints to explore the temporal dynamics. Our approach expresses the displacement information by using orientation encoding (direction of joints) and magnitude to provide information regarding the velocity of the movement. Furthermore, due to the successful results achieved by the skeleton image representations ([3], [6], [8]–[12]), our approach follows the same fundamentals by representing the skeleton sequences as matrices. First, we apply the depth-first tree traversal order [6] to the skeleton joints to generate a pre-defined chain order C that best preserves the spatial relations between joints in original skeleton structures[2]. Afterwards, we compute a matrix S that corresponds to a chain of concatenated skeleton joint coordinates from the frame t. In view of that, each column of the matrix corresponds to the temporal evolution of the arranged chain joint c. At this point, the size of matrix S is $C \times T \times 3$, where C is the number of joints of the chain, T is the total frame number of the video sequence and 3 is the number joint coordinate axes (x, y, z). Then, we create the motion structure $\mathcal{D}$ as $D_{c,t} = S_{c,t+d} - S_{c,}$, where each matrix cell is composed of the temporal difference computation of each joint between two frames of d distance, resulting in a $C \times T - d \times 3$ matrix.

We build two different representations using the proposed motion structure $\mathcal{D}$: one based on the magnitudes of joint motions and another one based the orientations of the joint motion. We compute both representations using $M_{c,t} = \sqrt{(D^x_{c,t})^2 + (D^y_{c,t})^2 + (D^z_{c,t})^2}$ and

$$\theta_{c,t} = \text{stack}(\theta^{xy}_{c,t}, \theta^{yz}_{c,t}, \theta^{zx}_{c,t}),$$

$$\theta^{xy}_{c,t} = \tan^{-1}\left(\frac{\mathcal{D}^y_{c,t}}{\mathcal{D}^x_{c,t}}\right),$$

$$\theta^{yz}_{c,t} = \tan^{-1}\left(\frac{\mathcal{D}^z_{c,t}}{\mathcal{D}^y_{c,t}}\right),$$

$$\theta^{zx}_{c,t} = \tan^{-1}\left(\frac{\mathcal{D}^x_{c,t}}{\mathcal{D}^z_{c,t}}\right),$$

where M is the magnitude skeleton representation of size $J \times T - d \times 1$ and θ is the orientation skeleton representation of size $J \times T - d \times 3$ (composed of 3 stacked channels).

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

To capture long-range joint interactions involved in activities, we pre-compute the motion structure $\mathcal{D}$ considering different d distances, which we called *Temporal Scale Aggregation* (TSA). For each of the motion structures $\mathcal{D}$, we compute its respective magnitude skeleton representation M and then stack them all into one single representation. The same idea is applied to compute the orientation skeleton representation θ, however since the orientation values are estimated for every joint, it might generate noisy values for joints without any movement. Therefore, we perform a filtering on θ based on the values of M with a weighting scheme, as

$$\theta^{'}_{c,t} = \left\{ \begin{array}{ll} 0, & \text{if } M_{c,t} < m \times d \\ \theta_{c,t}, & \text{otherwise} \end{array} \right. .$$

where m is a magnitude threshold value. Such technique adds more temporal dynamics to the representation by explicitly showing temporal scales to the network. Therefore, the network can learn which movements are relevant for the activity and also is able to capture long-range joint interactions.

III. Experimental Analysis

In this section we present the experimental results obtained with our proposed representations. The results will be

[2] Chain C 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 [6].

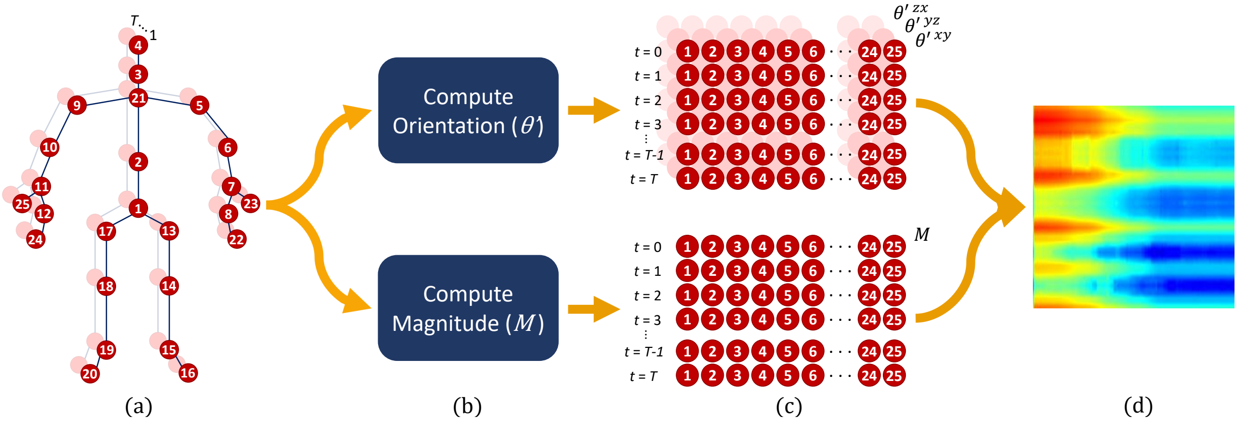


Fig. 1. SkeleMotion representation. (a) Skeleton data sequence of T frames. (b) Computation of the magnitude and orientation from the joint movement. (c) $\theta^{'}$ and M arrays: each row encodes the spatial information (relation between joint movements) while each column describes the temporal information for each joint movement. (d) Skeleton image after resizing and stacking of each axes.

presented in two different groups, each one regarding the proposed representation and its experimental setup.

A. Image Domain-based Approach Evaluation

We first introduce the implementation details regarding MOS and then we compare the proposed approach to other CNN methods in the literature. We used Very Deep Two-Stream network (VD2S) [7] with VGG-16 and Temporal Segment Networks (TSN) with Inception [4] as baseline comparisons. To isolate only the contributions brought by our method, the baselines were evaluated with ImageNet model as pre-training, the same splits of training and testing data. The evaluations are performed considering a well-known dataset for the activity recognition problem, UCF101 [13].

1) Implementation Details: Training: Following our baselines [4], [7], we set the learning rate initially to 0.005. For the VD2S [7], the learning rate decreases at every $5,000$ iterations dividing it by 10. The maximum number of iterations was set to $15,000$. We followed a similar scheme for the TSN [4] reducing the learning rate after $12,000$ and $18,000$ iterations. For the TSN, the number of iterations was set to $20,000$. We kept the same schedule for all training sets. The weights are learned using the mini-batch stochastic gradient descent with a momentum set to 0.9 and weight decay of 0.0005. We also set high dropout ratio for the fully-connected (FC) layers (0.9 and 0.8).

We employed the same data augmentation techniques used by our baselines [4], [7]. Thus, we cropped and flipped four corners and the center of the frame. In addition, we applied a multi-scale cropping method and randomly sampled the cropping width and height from $\{256, 224, 192, 168\}$ (finally, we resize the cropped regions to 224×224).

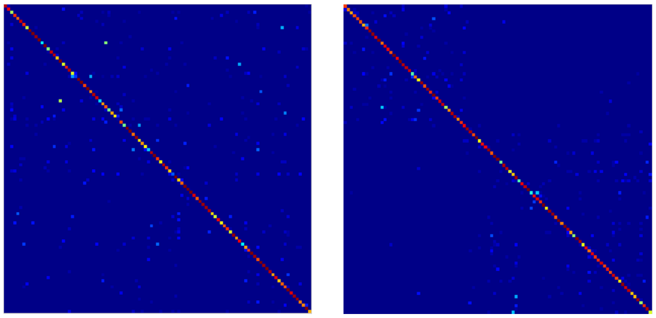
Test: To perform a fair comparison, we applied the same test scheme used by our baseline [7], described as follows. First, we sample 25 magnitude/orientation flow images for testing. Then, from each of these, we obtain 10 convolutional network inputs (by cropping and flipping four corners and the center). Finally, the prediction score for the input video is obtained by averaging the sampled images' scores and their crops. For the fusion of streams, we use a non-weighted linear fusion that consists in a combination of their prediction scores.

Optical Flow Extraction: The magnitude/orientation images are computed from the optical flow information. For the sake of comparison, we used the same optical flow algorithm as our baselines (TVL1 algorithm [14]). To obtain the magnitude and orientation image information we empirically set the parameter$m = 128$ to compute $\theta^{'}$.

2) Results and Comparisons: We report the activity recognition performance of our MOS with VGG-16 architecture and the VD2S baseline [7] on the UCF101 dataset in Table I. A considerable improvement was obtained with MOS when compared to the baseline single streams, reaching 90.5% of average accuracy over the 3 splits of UCF101 dataset. There is an improvement of 3.5 p.p. when compared to Very Deep Temporal Stream (VDTS) [7] and 12.1 p.p. when compared to Very Deep Spatial Stream (VDSS) [7]. This shows that the optical flow preprocessing (i.e., extraction of magnitude and orientation information) brings improvement over using raw optical flow information. Furthermore, it is worth noting that our best result using MOS is close to the best one reported (VD2S), which is obtained by using a combination of two different streams (spatial and temporal information), while we only used our single MOS (temporal information). Therefore, such results can be considered remarkably good and shows that preprocessing the inputs helps on guiding the network to extract certain information.

Figure 2 shows the confusion matrices of VDTS and our MOS for the UCF101 split 1. Our approach fails on classes that are more semantically closer to each other[3], whereas VDSS and VDTS fail in a random manner. In addition, the three methods produce false positives and false negatives different from each other, indicating the possibility of fusion.

[3] Since the activities on the confusion matrices are sorted according to its labels (e.g., ApplyEyeMakeup, ApplyLipstick, or BaseballPitch, Basketball, BasketballDunk), near regions denote semantically closer activities.



(a) VDTS

(b) MOS

Fig. 2. Confusion matrices on UCF101 split 1. False positives and false negatives were highlighted to show where each method fails.

To exploit a possible complementarity of the three approaches (VDSS, VDTS and our MOS), we combined the different streams by employing a late fusion technique using a weighted linear combination of their prediction scores. According to the results showed in Table I, any type of combination performed with our MOS provides better results than VD2S [7], with the best result achieving an improvement of 2.4 p.p. over VD2S [7].

TABLE I
ACTIVITY RECOGNITION RESULTS (AVERAGE ACCURACY % OVER 3 SPLITS AND STANDARD DEVIATION) OF MOS WITH VGG-16 ARCHITECTURE AND VD2S [7] BASELINE ON THE UCF101 [13]

	Approach	**Avg. Acc. (%)**
Baseline	VDSS [7]	78.4 ± 1.1
	VDTS [7]	87.0 ± 1.0
	VD2S [7]	**91.4** ± 0.3
Our results	MOS (VGG-16)	90.5 ± 0.9
	MOS + VDSS [7]	**92.5** ± 0.5
	MOS + VDTS [7]	**92.4** ± 0.9
	MOS + VD2S [7]	**93.8** ± 0.7

We also report the activity recognition performance of our MOS with Inception architecture in comparison with the TSN [4] baseline in Table II. According to the results, a considerable improvement was achieved with MOS when compared to the TSN [4] baseline single streams, reaching 92.4% of accuracy on UCF101. We can note an improvement of 7.3 p.p. when compared to Spatial Segment Stream (SSS) [4] and 2.7 p.p. when compared to Temporal Segment Stream (TSS) [4]. Once more, such results confirm that preprocessing the optical flow inputs helps guiding the network to extract a better information and, although temporal evolution patterns can be learned implicitly with CNNs, an explicit modeling is preferable and is able to achieve better results.

We also exploited a possible complementarity of the spatial and temporal streams from TSN and our MOS approach. Here, we applied the same late fusion technique used on VGG-16 architecture experiments, which consists of a weighted linear combination of the prediction scores. Last line of Table II shows the combination results, with the best result improving 2.7 p.p when compared to TSN [7].

TABLE II
ACTIVITY RECOGNITION RESULTS (AVERAGE ACCURACY % OVER 3 SPLITS AND STANDARD DEVIATION) OF MOS WITH INCEPTION ARCHITECTURE AND TSN [4] BASELINE ON THE UCF101 [13].

	Approach	**Avg. Acc. (%)**
Baseline	SSS [4]	85.1 ± 0.4
	TSS [4]	89.7 ± 1.6
	TSN [4]	**94.0** ± 0.4
Our results	MOS (Inception)	92.4 ± 0.7
	MOS + SSS [4]	**96.3** ± 0.3
	MOS + TSS [4]	**94.3** ± 0.6
	MOS + TSN [4]	**96.7** ± 0.2

Table III presents results for many works on the UCF101 dataset. According to the results, by only using our MOS, we outperform many methods [2], [4], [15]–[18]. It is worth mentioning that we also improved the results achieved by the original two-stream from Simonyan and Zisserman [2]. Using the VGG-16 architecture, we outperform it by 2.5 p.p. (temporal stream) and by 5.8 p.p. (combining it with VD2S). Further more, using the Inception architecture, we outperform it by 4.4 p.p. (temporal stream) and by 8.7 p.p. (combining it with TSN). Finally, we can observe that our best result did not outperform only the I3D method from Carreira et al. [5]. However, it is important to emphasize that they used the huge Kinetics Dataset [5] for pre-training. Nevertheless, we believe our results are remarkably good since 3D convolutional operations are more computationally expensive than the 2D convolutional operations used in our approach. For instance, the Two-Stream I3D network used by Carreira et al. [5] has 25 million parameters, while the 2D Two-Stream employed by us has less than half (12 million parameters).

TABLE III
ACTIVITY RECOGNITION ACCURACY COMPARISON ON THE UCF101 [13].

	Approach	**Avg. Acc. (%)**
Literature Results	Deep Networks [15]	65.4
	Composite LSTM [16]	75.8
	C3D [17]	85.2
	Factorized CNN [18]	88.1
	Two-Stream [2]	88.0
	Two-Stream F [19]	92.5
	KVMF [20]	93.1
	TSN (3 modalities) [4]	94.2
	R-STAN-101 (RGB+FLOW) [21]	94.5
	STM ResNet-50 [22]	96.2
	Two-Stream I3D [5]	**98.0**
Our Results	MOS (VGG-16)	90.5
	MOS (VGG-16) + VD2S	93.8
	MOS (Inception)	92.4
	MOS (Inception) + TSN	**96.7**

To verify the statistical significance of our results, a statistical test for the differences between the means was performed using a Student t-test [23], paired over the dataset splits. Thus, at 95% confidence level, we can conclude that the difference is significant for our combination results.

3) Discussion: To better analyze our proposed approach, we take a closer look at the activities from UCF101 that our method achieved higher performance than the baseline approaches. For instance, some activities that were most correctly classified by MOS and misclassified by the baselines include activities with movements on very similar areas (*apply lip stick* and *shaving beard* with *brushing teeth* or *apply eye makeup*) and classes that are more semantically closer to each other, such as *rafting* with *kayaking* and *basketball* and *volleyball spiking*. The correct classification of these activities by our MOS approach shows that feeding the network with explicit orientation information instead of x and y displacements might improve the classification of activities with movements on very close areas or even with similar movements. Besides, we might note the importance of using magnitude information (velocity) since the velocity information can be used to distinguish between similar activities with different velocities.

The analysis of the misclassified videos revealed that our method had trouble classifying activities that are only distinguishable by the object used as they have very similar movements, such as playing instrument activities (cello, guitar and sitar; or daf, dhol and tabla). The same difficulties were also noted on the baseline methods. Another misclassification of our approach is *walking with dog* with *horse riding*. Such analysis indicates the use of object information could help enhancing the classification.

B. Skeleton Domain-based Approach Evaluation

This section describes the experimental results obtained with the proposed SkeleMotion approach. We first introduce the implementation details and then we compare our proposed approaches to other CNN methods in the literature on the RGB+D 60 [24] dataset as well as to state-of-the-art methods on the NTU RGB+D 120 [25] dataset, in which we applied the same split of training and testing data and we employ the evaluation protocols and metrics proposed by their authors.

1) Implementation Details: To isolate only the contributions brought by the proposed representation, all compared skeleton image representations were implemented and tested on the same datasets and used the same network architecture. We adopted a smaller version of the CNN architecture proposed by Li et al. [11] to learn the features of the generated skeleton image representations, which consists of three convolution layers and only two FC layers.

To cope with activities 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. To obtain the orientation skeleton image representation $\theta^{'}$ we empirically set the parameter $m = 0.004$, as described in Section II-B1.

2) Results and Comparisons: We used a subset of the NTU RGB+D 60 [24] training set (considering cross-view protocol) to set the number of temporal scales of our SkeleMotion approach. We empirically varied it from two to four temporal scales considering 20 frames in total. The best result is obtained by using three temporal scales for both magnitude and orientation. Also, we noticed that the performance tends to saturate or drop when considering more temporal scales.

TABLE IV
ACTIVITY RECOGNITION ACCURACY (%) RESULTS ON THE NTU RGB+D 60 [24].

	Approach	Cross-subject Acc. (%)	Cross-view Acc. (%)
Baselines	Du et al. [3]	68.7	73.0
	Wang et al. [8]	39.1	35.9
	Ke et al. [10]	**70.8**	75.5
	Li et al. [11]	56.8	61.3
	Yang et al. [6]	69.5	**75.6**
Our results	SkeleMotion (Ori.)	65.3	73.2
	SkeleMotion (Mag)	69.6	80.1
	SkeleMotion (Mag.Ori.)	72.2	81.7
	SkeleMotion (Mag.Ori.) + [26]	**77.9**	**86.1**

Table IV compares our approach with other skeleton image representations. The methods that have more than one “image” per representation ([4] and [10]) were stacked to be used as input to the network. The same was performed for our approach, considering magnitude and orientation. Regarding the cross-subject protocol, the best result was obtained by our SkeleMotion (Mag.Ori.) representation with 72.2% of accuracy. There is an improvement of 1.4 (p.p.) when compared to Ke et al. [10], which was the best baseline result. It is worth noting that there is a considerable improvement of 15.4 (p.p.) obtained by SkeleMotion (Mag.Ori.) when compared to [11] baseline, which also explicitly encodes motion information. On the cross-view protocol, the best results was also achieved by our SkeleMotion (Mag.Ori.) representation, with 81.7% of accuracy. There is an improvement of 6.1 (p.p.) when compared to the Tree Structure Skeleton Image (TSSI) [6], which was the best baseline result. Again, there is a considerable improvement of 20.4 (p.p.) when compared to [11].

To exploit a possible complementarity of temporal and spatial skeleton information, we combined our approach with Tree Structure Reference Joints Image (TSRJI) [26] by employing a late fusion technique (non-weighted linear combination of the prediction scores of each method). According to the results showed in Table IV, the combination our approaches achieves the best results. Detailed improvements are shown in Figure 3.

Finally, Table V presents the experiments of our proposed skeleton image representation on the NTU RGB+D 120 [25] dataset. Again, when combining our representation we achieve state-of-the-art results, outperforming the best reported method (Body Pose Evolution Map [27]) by up to 6.5 p.p. on cross-subject protocol and achieve competitive results on cross-setup protocol (up to 0.8 p.p. better).

In comparison with LSTM approaches, we outperform the best reported method (Two-Stream Attention LSTM) by 9.9 p.p. on cross-subject protocol. Regarding the cross-setup protocol, we outperform them by 4.4 p.p. using our combined skeleton image representation. This indicates that, our skeleton image representation approach used as input for CNNs leads

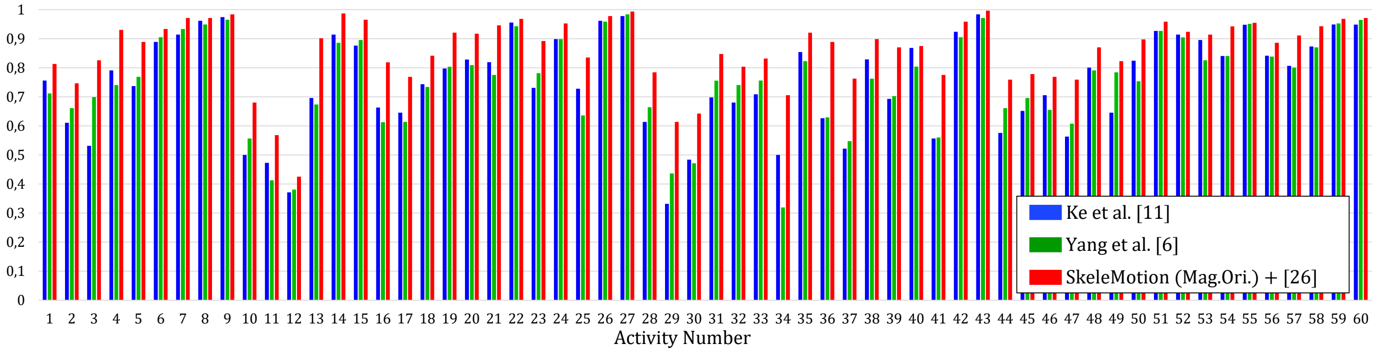


Fig. 3. Comparison of SkeleMotion (Magnitude-Orientation) with Ke et al. [10] and Yang et al. [6] on the NTU RGB+D 60 [24] for crossview protocol.

to a better learning of temporal dynamics than the approaches that employs LSTM.

TABLE V
ACTIVITY RECOGNITION ACCURACY (%) RESULTS ON THE NTU RGB+D 120 [25].

	Approach	**Cross-subject Acc. (%)**	**Cross-setup Acc. (%)**
Literature results	Part-Aware LSTM [24]	25.5	26.3
	Dynamic Skeleton [28]	50.8	54.7
	Internal Feat. Fusion [29]	58.2	60.9
	GCA-LSTM [30]	58.3	59.2
	Multi-Task Learning [10]	58.4	57.9
	FSNet [31]	59.9	62.4
	Two-Stream LSTM [32]	61.2	63.3
	Multi-Task CNN [33]	62.2	61.8
	Body Evolution Map [27]	**64.6**	**66.9**
Our results	SkeleMotion (Ori.)	52.2	54.1
	SkeleMotion (Mag.)	57.6	60.4
	SkeleMotion (Mag.Ori.)	62.9	63.0
	SkeleMotion (Mag.Ori.) + [26]	**71.1**	**67.7**

3) Discussion: We better analyze our achieved results by taking a closer look at the activities from the NTU RGB+D 60 dataset that our method achieved higher performance than [10] and [6]. The activities that were most correctly classified by our representation and misclassified by the baselines are activities involving arm and hand movements, such as *writing; playing with phone;* and *handshaking*. We note that the baselines usually confused such activities, which are activities involving arm and hand movements.

The correct classifications of the aforementioned activities by our representation show that feeding the network with explicit motion information can be used to distinguish between similar activities with different velocities. 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.

The analysis of the misclassified videos revealed that the method had trouble classifying activities that are only distinguishable by the object used as they have very similar movements (e.g., the activities *writing* is confused with *reading*, *typing on a keyboard* and *playing with phone*). This was expected given that our approach is based only on skeleton joints and does not encodes any appearance information.

IV. Conclusions

In this work, we have presented novel representation methods for the activity recognition problem and evaluated in three datasets from the literature. Regarding our image domain CNN-based method, MOS, the representation uses non-linear transformations on the optical flow to generate magnitude-orientation input images for a temporal stream. MOS has the advantage of capturing displacement information by using orientation of the optical flow and velocity of the movement considering the optical flow magnitude. We showed that our approach provides better recognition accuracy than other isolate streams - i.e., only using spatial stream or temporal stream of the literature compared to our MOS temporal stream.

We also introduced a skeleton domain CNN-based method, SkeleMotion. SkeleMotion is based on temporal dynamics encoding and explicitly uses motion information (magnitude and orientation) of skeleton joints. It also takes advantage of a structural organization of joints that preserves spatial relations of more relevant joint pairs and also incorporates different temporal relationships between the joints (TSA). Again, the performed experiments confirmed the relevance of the use of magnitude and orientation information to be used for motion learning for the activity recognition task.

V. Awards & Publications

Contributions of this work correspond to the main achievements during the doctorate research. Methods were published on the International Conference on Pattern Recognition (ICPR) 2016 [34] (*awarded with an IAPR Travel Stipend*), Conference on Graphics, Patterns and Images (SIBGRAPI) 2017 [35] *(awarded as the best Computer Vision/Image Processing/Pattern Recognition main track paper award)*, International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP) 2018 [36], IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) 2019 [37], Conference on Graphics, Patterns and Images (SIBGRAPI) 2019 [26] and a publication in the Journal of Visual Communication and Image Representation (JVCI) 2019 [38].

ACKNOWLEDGMENTS

The authors would like to thank the National Council for Scientific and Technological Development – CNPq (Grants 438629/2018-3, 204952/2017-4 and 311053/2016-5), 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). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior - Brasil (CAPES) - Finance Code 001.

REFERENCES

[1] C. Caetano, J. A. dos Santos, and W. R. Schwartz, "Motion-based representations for activity recognition," Ph.D. dissertation, Department of Computer Science, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil, Jan. 2020.

[2] K. Simonyan and A. Zisserman, "Two-stream Convolutional Networks for Action Recognition in Videos," in *NIPS*, 2014.

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

[4] 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 *ECCV*, 2016.

[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the Kinetics dataset," in *CVPR*, 2017.

[6] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with spatio-temporal visual attention on skeleton image sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[7] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards Good Practices for Very Deep Two-Stream ConvNets," *CoRR*, 2015.

[8] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *ACM International Conference on Multimedia (MM)*, 2016.

[9] M. Liu, C. Chen, and H. Liu, "3d action recognition using data visualization and convolutional neural networks," in *IEEE International Conference on Multimedia Expo Workshops (ICME)*, 2017.

[10] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

[12] ——, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[13] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild," CRCV-TR, Tech. Rep., 2012.

[14] C. Zach, T. Pock, and H. Bischof, "A Duality Based Approach for Realtime TV-L1 Optical Flow," in *Proceedings of the 29th DAGM Conference on Pattern Recognition*, 2007.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," in *CVPR*, 2014.

[16] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised Learning of Video Representations Using LSTMs," in *ICML*, 2015.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features With 3D Convolutional Networks," in *ICCV*, 2015.

[18] L. Sun, K. Jia, D. Y. Yeung, and B. E. Shi, "Human Action Recognition Using Factorized Spatio-Temporal Convolutional Networks," in *ICCV*, 2015.

[19] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *CVPR*, 2016.

[20] W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao, "A Key Volume Mining Deep Framework for Action Recognition," in *CVPR*, 2016.

[21] Q. Liu, X. Che, and M. Bie, "R-stan: Residual spatial-temporal attention network for action recognition," *IEEE Access*, 2019.

[22] B. Jiang, B. Jiang, W. Gan, W. Wu, and J. Yan, "STM: SpatioTemporal and Motion Encoding for Action Recognition," in *ICCV*, 2019.

[23] R. Jain, *The art of computer systems performance analysis - techniques for experimental design, measurement, simulation, and modeling.* Wiley, 1991.

[24] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] 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," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[26] C. A. Caetano, F. Brémond, and W. R. Schwartz, "Skeleton image representation for 3d action recognition based on tree structure and reference joints," in *2019 32th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2019.

[27] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[28] J. Hu, W. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[29] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[30] J. Liu, G. Wang, P. Hu, L. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[31] J. Liu, A. Shahroudy, G. Wang, L. Duan, and A. Kot Chichung, "Skeleton-based online action prediction using scale selection network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[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," *IEEE Transactions on Image Processing*, 2018.

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

[34] C. Caetano, J. A. dos Santos, and W. R. Schwartz, "Optical Flow Co-occurrence Matrices: A novel spatiotemporal feature descriptor," in *International Conference on Pattern Recognitio (ICPR)*, 2016.

[35] C. A. Caetano, V. H. C. D. Melo, J. A. dos Santos, and W. R. Schwartz, "Activity recognition based on a magnitude-orientation stream network," in *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2017.

[36] C. Caetano, J. A. dos Santos, and W. R. Schwartz, "Statistical measures from co-occurrence of codewords for action recognition," in *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, 2018.

[37] C. A. 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)*, 2019.

[38] C. Caetano, V. H. de Melo, F. Brémond, J. A. dos Santos, and W. R. Schwartz, "Magnitude-orientation stream network and depth information applied to activity recognition," *Journal of Visual Communication and Image Representation*, 2019.