REGINA—Reasoning Graph Convolutional Networks in Human Action Recognition

Bruno Degardin, Vasco Lopes, and Hugo Proença, *Senior Member, IEEE*

Abstract—It is known that the kinematics of the human body skeleton reveals valuable information in action recognition. Recently, modeling skeletons as spatio-temporal graphs with Graph Convolutional Networks (GCNs) has been reported to solidly advance the state-of-the-art performance. However, GCN-based approaches exclusively learn from raw skeleton data, and are expected to extract the inherent structural information on their own. This paper describes *REGINA*, introducing a novel way to *RE*asoning *G*raph convolutional networks *IN* Human *A*ction recognition. The rationale is to provide to the GCNs additional knowledge about the skeleton data, obtained by hand-crafted features, in order to facilitate the learning process, while guaranteeing that it remains fully trainable in an end-to-end manner. The challenge is to capture complementary information over the dynamics between consecutive frames, which is the key information extracted by state-of-the-art GCN techniques. Moreover, the proposed strategy can be easily integrated in the existing GCN-based methods, which we also regard positively. Our experiments were carried out in well-known action recognition datasets and enabled to conclude that REGINA contributes for solid improvements in performance when incorporated to other GCN-based approaches, without any other adjustment regarding the original method. For reproducibility, the REGINA code and all the experiments carried out will be publicly available at https://github.com/DegardinBruno.

Index Terms—Action recognition, graph convolutional networks, human behavior analysis, skeleton-based action recognition.

I. INTRODUCTION

THE remarkable effectiveness of the human visual perception system to recognize motion patterns has been inspiring decades of research in the computer vision community
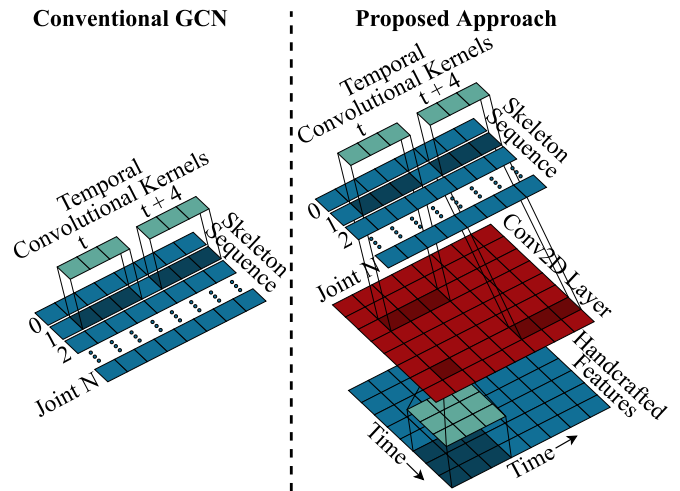
Fig. 1. Rationale for the strategy proposed to **improve temporal graph convolution** strategies. When compared to the conventional scheme (at left), we consider two additional components: 1) a set of handcrafted features (lower layer on the right side); that are integrated into the temporal convolutional kernel patches (light blue), according to 2) a learnable temporal convolutional layer (in red). This information is further fused to the skeleton sequence information, according to the conventional temporal convolutional kernels.

[1]–[6]. In particular, the ability to instinctively recognize other humans actions has been vital for the survival of the specie itself. However, action recognition can be particularly hard to perform automatically, due to the sophistication and complexity of movements, the inter-dissimilarity between humans, the background clutter and viewpoints [7]–[9]. This is an extremely crowded topic in the research community, where skeleton-based approaches have been gaining popularity, due to their ability to partially handle these challenges: they are known to faithfully handle variations in appearance and cluttered backgrounds, which allowed the pioneering skeleton-based action recognition techniques [10]–[13] to learn and extract relevant patterns.

The appearance of data-driven approaches (i.e., deep learning-based) facilitated the perception of human poses on image data, through efficient pose estimation algorithms [14]–[17], which has propelled a growing number of skeleton-based techniques and large-scale datasets [18]–[20]. This ability to extract high-level features with deep learning-based architectures has paved the way to structured skeleton-based approaches, such as convolutional neural networks (CNNs) with pseudo-images [21]–[23] and recurrent neural networks (RNNs) with sequence coordinate vectors [24]–[26].

More recently, graph convolutional networks (GCNs) were introduced [27], to model skeleton data by means of spatiotemporal graph convolutions [6]. Due to its intrinsic ability to exploit the skeleton inherent graph structure, several variants were subsequently introduced, typically proposing additional modules [28]–[31] and multi-streams [1], [3], [32] strategies.

In spite of the rapid progresses observed in recent years, two major shortcomings remain: 1) the high-level formulation of temporal graph convolutions constrains its design in computing them over a root node temporal neighbor set. Despite its neighborhood importance, this results in losing the skeleton global view, in particular concerning its temporal evolution; and 2) current methods use a spatial adjacency matrix to define the graph spatial topology and only assume its natural temporal adjacency, pre-defining the temporal receptive fields heuristically. Even though [1], [3], [29] present a learnable spatial adjacent matrix and graph shifting, our experiments point for local and temporal limitations of the encoding process, losing the holistic perspective in their temporal convolutions.

As a response to the above problems, in this paper we describe REGINA, a novel graph convolution architecture that can be easily integrated into the existing GCN-based techniques and enables solids improvements in performance with respect to the baselines. REGINA extends GCNs by employing shallow temporal representations to obtain a learnable temporal adjacency configuration, adjusting the temporal receptive fields. The proposed architecture is composed of a new graph convolution layer that learns the temporal importance of actions, which is trainable in an end-to-end manner, exploiting the skeleton data in a complementary way apart motion between frames.

The proposed strategy addresses naturally the previously mentioned problems: instead of using GCNs that only consider the temporal adjacency, we integrate a handcrafted self-similarity matrix (SSM) to the temporal graph convolution part, aimed to enhance the global connectivity across the temporal axis. Specifically, by extracting the SSM with point-wise convolutional kernel patches, the temporal graph receptive fields can be weighted based in the skeleton corresponding temporal differences, assigning local weights in direct proportion to the difference values and, consequently, considering an action *holistic view*. As illustrated in Fig. 1, the SSM is fed into a 2-dimensional convolutional layer before weighting each sample from the temporal convolution kernel. As different graph convolutional layers may require different temporal receptive fields [33], [34], the temporal configuration is learned and adjusted based on the SSM descriptor. Therefore, the proposed scheme will behave similarly to a self-attention module computed over time, which takes into account the temporal differences and gets self-optimized through the 2-dimensional convolutional layer by adjusting those weights. The self-similarity descriptor choice is justified empirically, based in the experiments carried out, which showed the robustness of the SSM descriptor to pose and camera variance. Additionally, as described in Section IV-B, REGINA outperforms the conventional temporal convolutions strategies and can be easily integrated to state-of-the-art GCN-based techniques, enabling consistent improvements in performance

in well-known datasets. A cohesive view of the proposed REGINA strategy, incorporated within a spatiotemporal graph convolutional network is given in Fig. 2.

In summary, our main contributions are three-fold: 1) We propose a graph convolution architecture that fuses shallow temporal representations to spatiotemporal GCNs, enabling to retain the skeleton global information, while extracting other higher-level spatiotemporal semantic features; 2) we experimentally show the ease of incorporating REGINA into the state-of-the-art GCN-based approaches, which does not prevent them from being trained in an end-to-end manner, enabling solid improvements in performance over the baselines; 3) we report competitive results with respect to the state-of-the-art techniques in large-scale datasets (NTU RGB+D and Kinetics-Skeleton) for skeleton-based action recognition.

The remainder of this paper is organized as follows: Section II summarizes the related work and in Section III we describe in detail the REGINA strategy. Section IV discusses our experiments and results and, finally, Section V concludes this paper.

II. Related Work

Due to the rise of deep learning-based architectures over the last decade, an evolution in the action recognition challenges has been witnessed. From the temporal information extraction breakthrough in video data, using 3-dimensional convolutional networks (3D ConvNets) [2], [4], [5], to the successful application of graph convolutional networks (GCNs) [1], [3], [6] to model skeleton-based data as a graph over the spatial and temporal domains.

A. Skeleton-Based Action Recognition

Human pose estimation [17], [35], [36] is acknowledged as one of the most important cues in behavior analysis/action recognition topics. Typically, pose estimation techniques provide 3D representations of the subject skeleton, attenuating the variations in appearance that RGB and depth data representations contain, while obtaining semantically rich and very descriptive representations that drive the learning process exclusively over human dynamics. Consequently, skeleton data ignited a variety of scientific works in action recognition, from the pioneering techniques based in handcrafted features to the current data-driven state-of-the-art approaches (deep learning-based).

Handcrafted feature methods capture the actions dynamics patterns by designing algorithms based on physical human foreknowledge. Such patterns can be self-similarity descriptors [10] to capture the temporal evolution of actions in a nearly view-independent paradigm, or Lie group curves [12], [13], to model the relative geometry between different body parts (providing a more meaningful representation than their absolute locations), or even covariance matrixes of joint trajectories [37], that hierarchically encode the temporal dependencies.

The existing deep learning-based architectures can be further classified into three categories. CNN-based methods, that represent the skeleton sequence data as pseudo-images through pre-processing transformations, such as temporal cylindrical
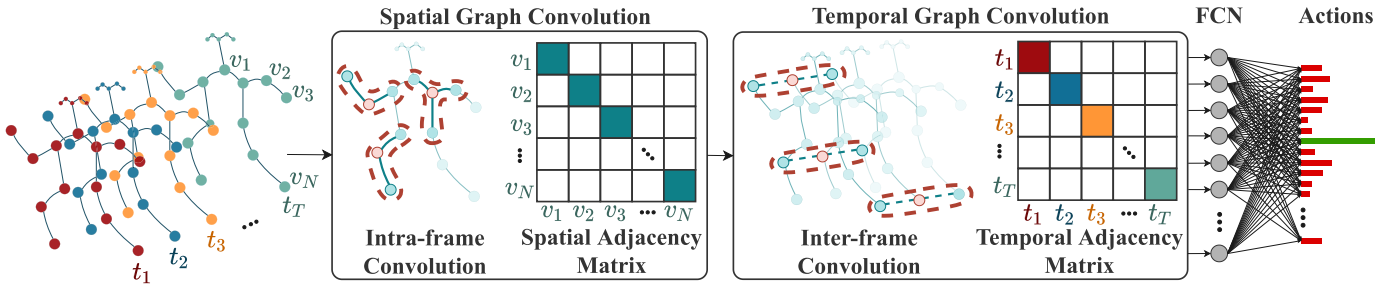
Fig. 2. **Cohesive view of the REGINA incorporation in a spatio-temporal graph convolutional network (ST-GCN)** [6]. A skeleton sequence is used as input of the network, where a spatio-temporal graph is constructed over the sequence. The spatial adjacency matrix handles the intra-frame (spatial) convolutions, while our temporal adjacency matrix handles the inter-frame (temporal) convolutions. The temporal graph convolution will consider the skeleton joint information (from the spatial graph convolution) and its global information in terms of temporal progression through the optimized temporal adjacency matrix. Finally, a standard Softmax layer predicts the corresponding action class.

coordinates [21], color series representing the skeletons [22], or temporal concatenation of the skeleton sequence raw coordinates [23]. RNN-based approaches, which are commonly applied over each human body joint coordinate vector, model the skeleton data as a sequence to learn its long-term contextual information, with a focus on resorting to LSTMs [24]–[26]. More recently, spatiotemporal GCNs successfully generalized the convolution from image to graph [1], [3], [6], [38], [39], being justified by their improved representation of structural information embedded in skeleton data, modeling the data as a graph with vertices (joints) and edges (bones), rather than vector sequences or 2D grids. This kind of models has been consistently advancing the state-of-the-art performance over the last years. Later on, Plizzari *et al.* [40] proposed to use the leading neural architecture for Natural Language Processing, Transformers, based on self-attention. The authors substitute regular graph convolutions with a two-stream transformer self-attention operator, which is spatially and temporally applied directly on nodes. However, similar to conventional graph convolutions, the temporal stream transformer (T-TR) analyses the same joint independently across time.

B. Graph Convolutional Networks

Considering that skeleton information can be naturally embedded in graphs, the work by Yan *et al.* [6] inspired several variants of spatiotemporal GCNs, by modeling skeletons as graph structures. The basis is that, for each node (joint), a layer-wise update is performed with a heuristically predefined neighborhood (adjacent joints). Additional modules were proposed, such as graph directions [29] and attention mechanisms [30]. Observing that the human body physical connections tend to constraint the learned graph structure, Shi *et al.* [3] proposed an adaptive adjacency matrix. More recently, spatial and temporal shift convolutions were also considered [1], having as goal to increase the flexibility of the resulting graph structure.

Given the important role of GCNs in skeleton-based action recognition, we observed that most state-of-the-art approaches underestimate the importance of handcrafted-based features, due to the challenge in representing the whole human skeleton dynamic complexity. Notwithstanding, we consider that

handcrafted-based techniques still provide valuable information about the skeleton spatial arrangement and temporal kinematics. In a way similar to the well-known optical flow technique [41], there is no doubt that CNN-based methods significantly overcome the performance of optical flow-based methods. However, fusion optical-flow to CNN-based features [42]–[45] not only advanced the state-of-the-art but has also inspired future works. Our proposal, REGINA, is the first to use handcrafted features at the core of GCNs. In opposition to previously published approaches, it is the first of its kind to provide an easy integration to other methods, by not modeling the raw skeleton information from scratch, and - instead - providing foreknowledge to GCNs as a way to conduct their learning process.

III. REGINA: PROPOSED GRAPH CONVOLUTIONAL ARCHITECTURE

As stated above, our rationale is that the richness of skeleton-based data expedites the understanding of human actions. In this section, we describe in detail our proposed graph convolutions architecture that fuses skeleton-based shallow representations into graph convolutional units, which enables to analyze complementary information and enhances the effectiveness of GCNs on this task.

A. Reasoning Skeleton Information

The pioneering action recognition techniques were able to extract consistent patterns of the human kinematics by resorting to foreknowledge formulations that analyse the natural body dynamics and extract skeleton information.

Having noticed the constrained structure of typical GCN-based techniques due to their high-level formulation, we propose to consider self-similarity descriptors between representations of time-frame pairs. Such kind of information, combined with skeleton-based sequences, explicitly describes the temporal difference evolution for each time frame skeleton pose, which provides additional information to the networks that should useful in the optimization process (learning). The underlying idea is that - this way - the graph is not required anymore to implicitly estimate the global skeleton information, and can analyze other types of (orthogonal) features, making the recognition of human actions an easier task.
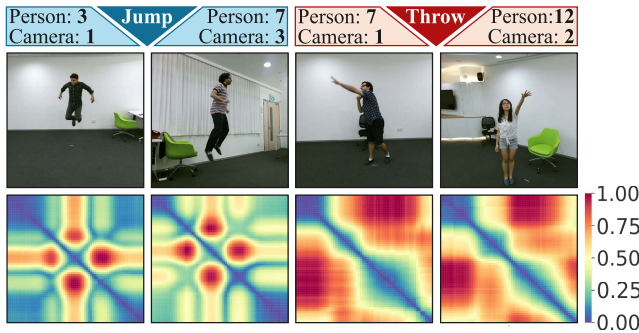
Fig. 3. **Example of the Self-Similarity Matrix (SSM) intra and inter-action variance over the NTU RGB-D** [20] dataset. The first-row illustrates two actions done by different people and under different pose (*Jump Up*, in the leftmost columns and *Throw*, in the rightmost columns). The bottom row provides the corresponding SSM yielding from the corresponding skeleton sequences.
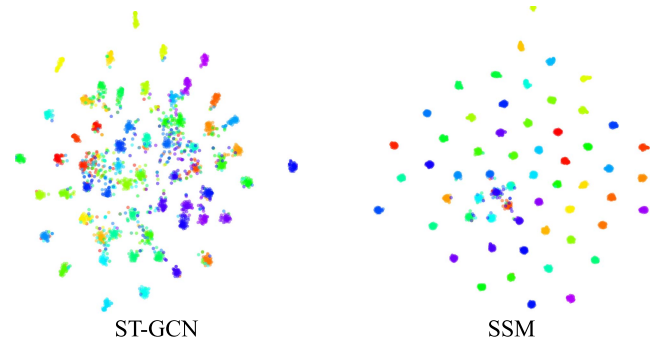

Fig. 4. **Feature Embedding**. Illustration of the Feature embedding (using t-SNE [46]) obtained for ST-GCN [6] and SSM features on the cross-view validation benchmark of the NTU RGB+D [20] dataset. Each point represents one skeleton video sequence, with elements belonging to the same class (total of 60 classes) appearing in the same color.

Having observed the SSM features permanence with respect to variations in pose and identity, we decided to use this kind of handcrafted features in our action recognition experiments. However, it should be noted that the proposed strategy can use other types of handcrafted information, that can seemingly integrated in our architecture without any other adaptions required. However, for the proof-of-concept reported in this paper, we resorted to the use of SSM in all experiments. As illustrated in Fig. 3, the SSM patterns can easily provide discriminating/permanent auxiliary information to the GCN, enabling the network to analyze other kinds of information. In this example, regarding the action "jump up", the two *hotspots* (e.g., in lower diagonal) reflect the body's movement's flexion (before jumping) and the highest attainable height from jumping up, the differences in the resulting SSM values are obvious.

Aiming at perceiving the robustness of SSM descriptor as generic features, we qualitatively evaluated them by visualizing the feature embedding yielding from a state-of-the-art action recognition technique (ST-GCN)and SSM. We randomly selected 50 skeleton sequences of each class from the cross-view validation benchmark of the NTU RGB+D [20] dataset, and then extracted the features of the skeleton sequences using our primary baseline pre-trained on the respective training benchmark, spatiotemporal graph convolutional networks (ST-GCN) [6], and the SSM descriptor used in our experiments. These features were then projected into the 2-dimensional space using t-SNE [46] technique. Figure 4 provides an illustration of the resulting embedding, where colors represent the different actions (classes), and each point represents one instance of the corresponding class.

Formally, let $\mathcal{V} \in \mathbb{R}^{n \times t \times c}$ denote the skeleton sequence representation with n joints, t frames, and c dimensions of joint coordinates. Given a set of skeleton points $\mathcal{V}_j = \{v_{ij}\}$, $i \in \{1, \ldots, n\}$ at time t, in this paper, we obtain the self-similarity matrix $\mathcal{S}$ using the mean Euclidean distance ($\ell - 2$ norm) between skeleton poses at any moments p and q of the sequence:

$$d_{pq} = \frac{1}{N} \sum_i \|v_{ip} - v_{iq}\|_2, \tag{1}$$

where v_{ip} and v_{iq} denote corresponding skeleton joints at times p and q on the sequence. Consequently, by obtaining a distance value between representations for all time-frame pairs, the self-similarity matrix descriptor yields a hollow square symmetric matrix of size $t \times t$.

B. REGINA + Graph Convolutional Networks

The recent appearance of Graph Convolutional Networks (GCNs) enabled the extraction of embedded patterns over the spatial and temporal axes of a skeleton sequence, essentially by generalizing convolutions from images to graphs. However, due to their high-level formulation, graph convolutions are performed locally by taking into account exclusively a root node neighboring set. Recent studies reported that considering only local connections might be sub-optimal for recognizing skeleton-based actions [1], [3], [29], [38]. These findings provided our motivations to take advantage of the valuable hard-coded knowledge from handcrafted shallow representations and design a graph convolution that not only considers an action holistic perspective in conventional graph convolutions, but can also be easily integrated into the current GCN-based approaches.

Let the skeleton-based data used in GCN be denoted as a spatiotemporal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with n skeleton joints and t frames. Therefore, the skeleton sequence's feature map can be denoted as $\mathbf{X} \in \mathbb{R}^{n \times t \times c}$, with c channels representing the joints coordinates. Typically, in spatial graph convolutions, one adjacency matrix $\mathbf{A}$ and one identity matrix $\mathbf{I}$ are used to define the intra-body joints connections, which can be dismantled into three partitions p (regarding the set of neighbors resulting from the spatial configuration proposed in [6]) where $\mathbf{A} + \mathbf{I} = \sum_p \mathbf{A}_p$. In a single frame, the graph convolution is given by:

$$\mathbf{Y} = \sum_{i=1}^{p} \Lambda_i^{-\frac{1}{2}} \mathbf{A}_i \Lambda_i^{-\frac{1}{2}} \mathbf{X} \mathbf{W}_i, \tag{2}$$

where the degree matrix $\Lambda_p^{ii} = \sum_j (\mathbf{A}_p^{ij})$ denotes the number of edges attached to each joint node, to normalize the adjacency matrix $\mathbf{A}_p$. $\mathbf{W}_p$ are the stacked weight vectors for each partition group p.

Over the temporal axis, graphs are defined over consecutive frames, where most skeleton-based GCN approaches [3], [6], [29], [31] use 1-dimensional kernels as the conventional temporal graph convolution operators, where the receptive fields are defined beforehand. Despite the temporal neighborhood importance, it is obvious that the skeleton global perspective is lost in this operation.

Considering the above weaknesses, we propose an adaptive and learnable graph convolution network that mainstreams the temporal evolution expressiveness (from the skeleton holistic perspective) over the corresponding temporal receptive fields from the temporal graph convolutions, that can also be interpreted as a learnable temporal adjacency matrix. As stated above, the self-similarity matrix $\mathcal{S}$ explicitly describes the temporal difference for each time-frame with respect to every other time-frames. However, since multiple graph convolutional layers are used, different layers contain multilevel semantic information, and the raw integration of $\mathcal{S}$ in (2) would result in the same rigid and pre-defined temporal configuration to every layer. For our purposes, it was considered essential to adapt the handcrafted features to different semantic levels. Hence, we resorted to a convolutional layer over the self-similarity descriptor, in order to adaptively learn the temporal configuration that fits optimally the hierarchical structure of GCNs. An illustration of this idea is provided in our reasoning graph convolutional unit in Fig. 1. Hence, our major contribution can be defined as:

$$\mathcal{R}_k = \left((\mathcal{S}_k \mathbf{W})^\top \mathcal{J}_k \right)^\top, \tag{3}$$

where $\mathcal{S}_k$ is the representation of the temporal evolution from the SSM comprised by the kernel size of the temporal convolutional kernel patch k. $\mathcal{J}_k$ is an all-ones vector of size c, and $\mathbf{W} \in \mathbb{R}^{t \times t \times C}$ denotes the filter weights from the temporal convolutional layer with t as the kernel size. This way, we are able to adaptively optimize our handcrafted descriptor, and the reasoning term is obtained for all feature channels in the graph convolution by incorporating (3) in (2), according to the Hadamard product:

$$\mathbf{Y} = \sum_{p=1}^{P} \Lambda_p^{-\frac{1}{2}} \mathbf{A}_p \Lambda_p^{-\frac{1}{2}} (\mathbf{X} \odot \mathcal{R}) \mathbf{W}_p \tag{4}$$

By incorporating the 2D convolution matrix directly into the skeleton receptive fields of graph convolutions, we are able to provide global information of the corresponding skeleton without dismissing its partition strategies applied to the neighbor set of each node. Furthermore, due to its simple formulation, our contribution in (3) is able to be easily integrated into other GCN-based methods, without preventing them from being trained in an end-to-end manner, as described in Section IV-B.

IV. EXPERIMENTS AND DISCUSSION

In this section, we start by comparing the effectiveness of two state-of-the-art GCN-based techniques with/without our contribution (REGINA). Then, we compare our best performing models to the state-of-the-art approaches over both benchmarks of the well-known NTU RGB+D [20] dataset and Kinetics-Skeleton [18].

A. Datasets, Evaluation Metrics and Experimental Settings

1) Datasets: Our experiments were conducted on two well-known datasets NTU RGB-D [20] and Kinetics-Skeleton [18]. NTU RGB-D [20] is currently the most widely used human labelled set for evaluating skeleton-based action recognition methods in 3-dimensional space. This set is composed of 56,880 video samples, with 60 action classes. Data were captured from highly restricted camera views providing 3D skeleton and RGB-D data from 40 volunteers for each action sample, with 25 joints for each skeleton. The authors have recommended two benchmarks: 1) cross-subject, where models are trained with 20 subjects and tested with the remaining 20 subjects; and the 2) cross-view setting: where models are trained with camera views 2 and 3 and tested on camera view 1.

Kinetics-Skeleton is a human action dataset with 300,000 video clips in 400 action classes. Since it only provides RGB videos without skeleton data, we obtain skeleton data by human pose estimation with OpenPose toolbox [15]. The human poses generated are in 2-dimensional space coordinates (x, y) with 18 joints from the RGB videos. However, we also use as input the confidence score c associated with each joint, such as the joint feature vector becomes (x, y, c). We select the two skeleton sequences with the highest average joint confidence scores when multiple persons are present in the scenes.

2) Evaluation Metrics: We followed the NTU RGB-D [20] and Kinetics-Skeleton [18] convention and evaluated the recognition performance by reporting the Top-1 and Top-5 recognition accuracy on both benchmarks.

3) Experimental Settings: The skeleton sequences temporal length were normalized to $t = 300$ frames in both datasets. Therefore, each self-similarity descriptor corresponding to one skeleton sequence is defined by a 300×300 matrix. For a fair comparison, the skeleton data was pre-processed as described in [6] and [3] for the ST-GCN and 2s-AGCN experiments, respectively. The methods that were used as baseline comparisons with/without the proposed REGINA were not adapted in any manner. The unique adaptation was that, when considering our proposal, instead of the one-dimensional graph convolutions, we used our temporal graph convolution scheme. All experiments reported in this paper follow these guidelines and settings.

B. Baselines Comparison

We evaluated the REGINA effectiveness with respect to two state-of-the-art GCN approaches over both evaluation protocols from NTU RGB-D [20] dataset, and the large-scale benchmark Kinetics-Skeleton [18]. Furthermore, we combined our approach with multi-stream architectures to illustrate the robustness of our proposal to different kinds of features.

1) GCN-Based + REGINA Fusion: In order to verify REGINA effectiveness when fused to GCN-based techniques, we considered two baselines: the spatiotemporal GCN [6], ST-GCN, and a two-stream adaptive GCN (2s-AGCN) [3], being both described in Section II. As these baselines use one-dimensional kernels in their temporal convolutions, some

TABLE I

COMPARISON OF **REGINA + GCN-BASED RESULTS**. PERFORMANCE COMPARISON WHEN FUSING THE PROPOSED REGINA ARCHITEC-TURES TO TWO STATE-OF-THE-ART METHODS. RESULTS ARE PROVIDED FOR BOTH BENCHMARKS OF THE NTU RGB+D [20] SET

	Method	Top-1	Top-5	Top-1	Top-5
		Cross-Subject		Cross-View	
NTU RGB-D	ST-GCN [6]	81.51	96.92	88.34	98.33
	ST-GCN + REGINA	**83.62**	**97.71**	**90.12**	**99.00**
	2s-AGCN [3]	**88.51**	**98.43**	95.12	99.14
	2s-AGCN + REGINA	88.44	98.22	**95.21**	**99.38**

TABLE III

COMPARISON OF **REGINA EFFECTIVENESS IN MULTI-STREAM ARCHITECTURES**. PERFORMANCE COMPARISON REGARDING THE INTEGRATION OF REGINA IN DIFFERENT FEATURE STREAMS (JOINTS AND BONES) OVER BOTH BENCHMARKS OF THE NTU RGB+D [20] SET

	Method	Top-1	Top-5	Top-1	Top-5
		Cross-Subject		Cross-View	
NTU RGB-D	Js-AGCN [3]	85.33	97.04	93.72	99.12
	Js-AGCN + REGINA	**86.62**	**97.51**	**93.91**	**99.20**
	Bs-AGCN [3]	86.21	97.09	93.18	98.93
	Bs-AGCN + REGINA	**86.23**	**97.77**	**93.32**	**99.41**

TABLE II

COMPARISON OF **REGINA + GCN-BASED RESULTS**. PERFORMANCE COMPARISON WHEN FUSING THE PROPOSED REGINA ARCHITECTURES TO TWO STATE-OF-THE-ART METHODS OVER THE KINETICS-SKELETON [18] DATASET

	Method	Top-1	Top-5
Kinetics	ST-GCN [6]	30.68	52.82
	ST-GCN + REGINA	**30.74**	**53.05**
	2s-AGCN [3]	36.11	58.70
	2s-AGCN + REGINA	**36.13**	**58.81**

TABLE IV

COMPARISON OF **REGINA EFFECTIVENESS IN MULTI-STREAM ARCHITECTURES**. PERFORMANCE COMPARISON REGARDING THE INTEGRATION OF REGINA IN DIFFERENT FEATURE STREAMS (JOINTS AND BONES) OVER THE KINETICS-SKELETON [18] DATASET

	Method	Top-1	Top-5
Kinetics	Js-AGCN [3]	35.08	57.09
	Js-AGCN + REGINA	**35.13**	**57.19**
	Bs-AGCN [3]	33.30	55.64
	Bs-AGCN + REGINA	**33.31**	**55.69**

adaptations were carried out in the original implementations to add our temporal graph convolutions scheme in each method, without any further changes in each original framework. Noting that REGINA does not prevent any of those methods from being trained in an end-to-end manner, we also used the default learning settings, reporting the obtained performance in Tables I and II. Results show that the REGINA scheme was able to improve the overall GCN effectiveness, contributing for improvements in performance. This is most evident in the "ST-GCN + REGINA" configuration in both datasets, where an improvement of 2.1 percentual points (i.e., decreasing the Top-1 error rates over 12%) in the Cross-Subject evaluation benchmark. With respect to the 2s-AGCN network, the performance improvements were not as much significant, and were mostly observed in the Cross-View setting. In our viewpoint, this was due to the learnable graph topology of that technique, which has additional flexibility to better suit the hierarchical structure of GCNs, reducing the benefits due to the inclusion of the SSM-based handcrafted features. Also, the REGINA scheme's improvements over the Kinetics-Skeleton were not as notable as for the NTU RGB+D dataset. This is justified by the quality of the skeleton pose from the data, and 2-dimensional space skeleton poses are less informative than 3-dimensional space. While NTU RGB+D was human labelled in 3-dimensional space, the Kinetics-Skeleton was obtained with a pose estimation algorithm from RGB images only, and the errors are much more frequent.

2) Multi-Stream Architectures + REGINA Fusion: Most of the state-of-the-art methods utilize multi-stream fusion strategies. As a baseline to verify whther REGINA also contributes for improvements in performance for such kind of architectures, we considered a two-stream adaptive GCN

(2s-AGCN) [3], which resorts not only to joints information (first-order) but also the second-order information (bones) of the skeleton-based data. Each stream of this framework is trained individually, combining both features to obtain the final class prediction.

The bones information is found by the difference between coordinates of two connected joints. For instance, for a bone with its source joint $v_i = (x_i, y_i, z_i)$ and target joint $v_j = (x_j, y_j, z_j)$, the corresponding vector is given by $b_{v_i, v_j} = (x_j - x_i, y_j - y_i, z_j - z_i)$. Considering that the human skeleton is an acyclic graph, we can assign each bone to a unique target joint. Moreover, as the skeleton central joint cannot be assigned to any bone, there is one bone less than joins in a skeleton, and - thus - an *empty bone* was added to the central joint to simplify the network design. We then obtain the self-similarity descriptor for the second-order information, similarly to the joints graph described in (1).

The results are shown in Tables III and IV. As can be seen, the REGINA scheme has - again - contributed for improvements in the results on both datasets. This is more evident in the joint-stream than in the bone-stream, which was justified by the fact that joints are less informative and discriminative than bones, where we already have the length and direction information.

C. State-of-the-Art Comparison

To objectively perceive whether REGINA provides consistent advances over the state-of-the-art, in this section we compare the results attained by REGINA with/without the ST-GCN and 2s-AGCN architectures in both benchmarks of the NTU RGB-D [20] dataset and on Kinetics-Skeleton [18] dataset. As previously stated, the state-of-the-art can be considered

TABLE V

NTU RGB+D Recognition Results. Performance Summary of the REGINA Architecture Fused to the ST-GCN and 2s-AGCN Architectures, With Respect to the State-of-the-Art Methods in Skeleton-Based Action Recognition Over Both Benchmarks From the NTU RGB+D [20] Dataset

	Method	Type	Top-1	Top-5	Top-1	Top-5
			Cross-Subject		Cross-View	
NTU RGB-D	Lie Group [12]	Handcrafted	50.1	—	52.8	—
	Deep LSTM [20]	RNN	60.7	—	67.3	—
	PA-LSTM [20]	RNN	62.9	—	70.3	—
	2s 3DCNN [47]	CNN	66.8	—	72.6	—
	ST-LSTM [24]	RNN	69.2	—	77.7	—
	TCN [23]	CNN	74.3	—	83.1	—
	VA-LSTM [26]	RNN	79.4	—	87.6	—
	Syn-CNN [22]	CNN	80.0	—	87.2	—
	ST-GCN [6]	Graph	81.51	96.92	88.34	98.33
	ST-GCN + REGINA	Graph + Handcrafted	**83.62**	**97.71**	**90.12**	**99.00**
	DPRL+GCNN [39]	Graph	83.5	—	89.8	—
	AS-GCN [38]	Graph	86.8	—	94.2	—
	2s-AGCN [3]	Graph	**88.51**	**98.43**	95.12	99.14
	2s-AGCN + REGINA	Graph + Handcrafted	88.44	98.22	**95.21**	**99.38**

to comprise three categories: 1) handcrafted feature methods; 2) manually structured skeleton data approaches (CNNs and RNNs); and 3) and graph-based architectures. Our results are summarized in Tables V and VI, being the methods grouped according to this criterion, in the "Type" column. As can be seen, REGINA is the first attempt to combine methods of two of out of these families: handcrafted features-based and graph-based networks.

The first observation was the overwhelming performance of graph-based architectures, when compared to the manually structured-based skeleton ones, such as CNNs and RNNs. Even though the latter approaches can be considered to still successfully extract structural information according to their data-driven architectures, it was evident solid improvements can be attained when modeling the skeleton data as a graph. Furthermore, despite the relatively poor performance of standalone handcrafted-based shallow representations, an overall improvement in performance can be obtained if handcrafted-based features are seemingly integrated more sophisticated Graph-based techniques, as the results of "ST-GCN + REGINA" (for both *Cross-Subject*, *Cross-View* settings and Kinetics-Skeleton) and "2s-AGCN + REGINA" (for the *Cross-View* setting and Kinetics-Skeleton).

The results obtained also turn evident the relatively poor performance of the ST-GCN model, when compared to the deep reinforcement learning GCN (DPRL+GCNN) [39], where this method resorts to reinforcement learning to select keyframes through a frame distillation network to refine the GCN. However, the ST-GCN enhanced by our proposal is able to overcome DPRL-GCNN performance over NTU RGB-D [20], which we also consider to provide a sifgnificant cue over the improvements that would also be attained if fused to other similar techniques. In our view, this is mostly due to the fact that the optimized temporal adjacency matrix provides self-attention information to specific frames, while

TABLE VI

Kinetics-Skeleton Recognition Results. Performance Summary of the REGINA Architecture Fused to the ST-GCN and 2s-AGCN Architectures, With Respect to the State-of-the-Art Methods in Skeleton-Based Action Recognition Over the Kinetics-Skeleton [18] Dataset

	Method	Type	Top-1	Top-5
Kinetics-Skeleton	Feature Enc. [6]	Handcrafted	14.9	25.8
	Deep LSTM [20]	RNN	16.4	35.3
	TCN [23]	CNN	20.3	40.0
	ST-GCN [6]	Graph	30.68	52.82
	ST-GCN + REGINA	Graph + Handcrafted	**30.74**	**53.05**
	AS-GCN [38]	Graph	34.8	56.5
	2s-AGCN [3]	Graph	36.11	58.70
	2s-AGCN + REGINA	Graph + Handcrafted	**36.13**	**58.81**

keeping the property of automated training in an end-to-end manner. Overall, our experiments point out that using graph convolutional networks fused to handcrafted-based feature layer can be considered an improvement for action recognition. Specifically, REGINA is able to achieve competitive results, bringing gains up to 2.1 and 1.3 percentile points within the ST-GCN and 2s-AGCN, respectively on NTU RGB-D, and overcoming the state-of-the-art methods on both datasets.

D. Ablation Studies

In order to perceive the most important REGINA components and evaluate the overall robustness of our method to changes in parameterization, we varied the distance metrics used to obtain the self-similarity descriptor, the convolutional kernel size used in the handcrafted-feature fusion. Also, we performed additional experiments that show the importance

TABLE VII

SELF-SIMILARITY'S DISTANCE METRICS RESULTS. PERFORMANCE SUMMARY OF DIFFERENT DISTANCE METRICS OVER THE CROSS-VIEW BENCHMARK OF NTU RGB+D [20]

	Method	Top-1	Top-5
NTU RGB-D		*Cross-View*	
	ST-GCN [6]	88.34	98.33
	ST-GCN + REGINA ($\ell - 1$)	89.33	98.62
	ST-GCN + REGINA ($\ell - 2$)	**90.12**	**99.00**

TABLE VIII

EFFECTIVENESS OF THE TEMPORAL ADJACENCY MATRIX. PERFORMANCE COMPARISON OF THE RESULTS OBTAINED WITH/WITHOUT THE CONVOLUTIONAL LAYER. RESULTS ARE GIVEN FOR THE *Cross-View* BENCHMARK OF NTU RGB+D [20] SET

	Method	Top-1	Top-5
NTU RGB-D		*Cross-View*	
	ST-GCN [6]	88.34	98.33
	ST-GCN + REGINA (w/o Conv-Layer)	88.62	98.43
	ST-GCN + REGINA (w Conv-Layer)	**90.12**	**99.00**

TABLE IX

IMPORTANCE OF THE KERNEL SIZE FROM THE LEARNABLE TEMPORAL ADJACENCY MATRIX OVER THE HANDCRAFTED FEATURE. RESULTS REGARD THE RECOGNITION PERFORMANCE OVER THE *Cross-View* BENCHMARK OF THE NTU RGB+D [20] SET, WITH DIFFERENT KERNEL SIZES (k) OF THE CONVOLUTIONAL LAYER ON TOP OF THE HANDCRAFTED FEATURE

	Method	Top-1	Top-5
NTU RGB-D		*Cross-View*	
	ST-GCN + REGINA ($k = 1$)	88.93	98.61
	ST-GCN + REGINA ($k = 3$)	**90.12**	**99.00**
	ST-GCN + REGINA ($k = 5$)	89.31	98.63
	ST-GCN + REGINA ($k = 7$)	89.12	98.51
	ST-GCN + REGINA ($k = 9$)	89.10	98.50

of the learnable temporal adjacency importance over the handcrafted features.

1) Role of the Distance Metrics: We compare two different distance metrics between time-frame pairs in obtaining each skeleton sequence self-similarity matrix: a) the Manhattan distance, also known as $\ell - 1$ norm, which expresses the sum of individual coordinates shifting difference between two skeleton poses from different temporal instances; and b) the Euclidean distance, also known as $\ell - 2$ norm, that expresses the direct distance from one pose to another from the temporal skeleton difference perspective. As shown in Table VII, the self-similarity descriptor using with the $\ell - 1$ norm and fused to the ST-GCN [6] network is already able to improve the baseline performance. However, since the similarity between two time-frame pairs only considers the distance between the same joints, without any structural information, directions or angles, better expressiveness of the temporal kinematics can be captured by the $\ell - 2$ norm, which justifies the optimal performance observed for that distance function.

2) Learnable Temporal Adjacency Matrix Effectiveness: As described in Section III-B, the temporal adjacency matrix consists of a convolutional layer on top of the handcrafted feature, which then becomes a learnable matrix that controls the GCN temporal receptive fields through the self-similarity matrix. For ablation purposes, we manually removed the convolutional layer and observed the resulting REGINA performance, with results shown in Table VIII. The obtained values show that the learnable temporal adjacency matrix has the ability to adapt to different semantical levels, which is beneficial for the hierarchical GCN as different layers contain multilevel semantic information. This way, REGINA architecture without the convolutional layer can be seen as a rigid and pre-defined temporal configuration (self-similarity descriptor) to every layer. Despite still contributing for improvements in performance than the ST-GCN [6] alone, this fusion scheme is clearly sub-optimal. Consequently, the graph convolution architecture with the learnable temporal adjacency matrix achieves consistently higher performance than its counterpart.

3) Exploring the Convolutional Kernel Size: Finally, we conducted some experiments on REGINA, when varying the kernel size (k) of the learnable adjacency matrix (*Conv2D Layer* in Fig. 1). Based in the results reported in Table IX, we can see that our model performance reaches a peak when $k = 3$, meaning that we have nine temporal comparisons (from the self-similarity matrix) being fed into the convolutional kernel. For smaller kernel sizes ($k = 1$), we only perform one comparison per convolution operation, which decreases the information feed to the subsequent layer of the network. In opposition, for larger $k > 3$ values, the performance was observed to decrease again, which we justified to the too high number of temporal comparisons fed into the convolutional kernel, that affects the discriminability of the descriptor itself (i.e., it is a weighted mean of too many values). This has a negative effect in the extraction of short-term temporal adjacency features, which results in lower performance.

V. CONCLUSION AND FURTHER WORK

The automated and reliable recognition of human actions has been the focus of substantial research efforts, being accepted that it will support various applications in the security/forensics domains. In this context, Graph Convolutional Networks-based (GCNs) approaches have been advancing consistently the state-of-the-art performance, being currently the most popular kind of approaches for this problem. In this paper, we proposed REGINA, an architecture that can be faithfully integrated into the state-of-the-art GCNs techniques and enables consistent improvements in performance over the baselines. The idea is to fuse the information naturally extracted by GCNs to *particularly interesting* handcrafted representations (SSM), which provide complementary information to the networks and intuitively enables that the networks focus in extracting alternate (orthogonal) information

to the SSM during the optimization process. As a result, the network has access to additional cues that were observed to contribute for improvements in performance. Also, the proposed REGINA scheme is able to overcome the limitations of conventional temporal convolutions, by capturing discriminative features over the temporal dynamics. Our experiments pointed out for highly encouraging results, where REGINA + state-of-the-art GCN techniques - is able to push forward their performance on two challenging benchmarks. Our efforts are currently focused in finding/describing automatically other kinds of handcrafted features that will also be positively combined to the automatically extracted by GCN-based networks.

REFERENCES

[1] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 183–192.

[2] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5533–5541.

[3] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12026–12035.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[5] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.

[6] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 2–11.

[7] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 1–43, 2011.

[8] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, 2010.

[9] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vis. Image Understand.*, vol. 115, no. 2, pp. 224–241, Feb. 2011.

[10] I. N. Junejo, E. Dexter, P. Laptev, and I. Perez, "View-independent action recognition from temporal self-similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 172–185, Jan. 2011.

[11] E. Ohn-Bar and M. M. Trivedi, "Joint angles similarities and HOG$_2$ for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 465–470.

[12] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 588–595.

[13] R. Vemulapalli and R. Chellappa, "Rolling rotations for recognizing human actions from 3D skeletal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4471–4479.

[14] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[15] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[16] N. Neverova, D. Novotny, and A. Vedaldi, "Correlated uncertainty for learning dense correspondences from noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.

[17] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, Feb. 2012.

[18] W. Kay *et al.*, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[19] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[20] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1010–1019.

[21] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3288–3297.

[22] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017.

[23] T. S. Kim and A. Reiter, "Interpretable 3D human action analysis with temporal convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 20–28.

[24] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 816–833.

[25] 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 *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, pp. 4263–4270, 2017.

[26] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2117–2126.

[27] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[28] B. Li, X. Li, Z. Zhang, and F. Wu, "Spatio-temporal graph routing for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8561–8568.

[29] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 7912–7921.

[30] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1227–1236.

[31] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 1112–1121.

[32] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 143–152.

[33] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 305–321.

[34] C. Li, Q. Zhong, D. Xie, and S. Pu, "Collaborative spatiotemporal feature learning for video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7872–7881.

[35] R. A. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7297–7306.

[36] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose flow: Efficient online pose tracking," 2018, *arXiv:1802.00977*.

[37] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *Proc. 33rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 2466–2472.

[38] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3595–3603.

[39] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5323–5332.

[40] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Comput. Vis. Image Understand.*, vols. 208–209, Jul. 2021, Art. no. 103219.

[41] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.

[42] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," 2014, *arXiv:1406.2199*.

[43] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4305–4314.

[44] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 20–36.

[45] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, "Real-time action recognition with enhanced motion vector CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2718–2726.

[46] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[47] H. Liu, J. Tu, and M. Liu, "Two-stream 3D convolutional neural network for skeleton-based action recognition," 2017, *arXiv:1705.08106*.

Vasco Lopes received the B.Sc. and M.Sc. degrees in computer science and engineering from the University of Beira Interior (UBI) in 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree in computer vision and neural architecture search. In the past, he has worked on research projects, such as uPATO, RobotChain, INDTech 4.0, and CovidSight. Since 2019, he has been collaborating as a Teaching Assistant at UBI, and since 2020, he has been the CEO and the Co-Founder of DeepNeuronic, a company that develops computer vision solutions to automate daily processes. Since 2021, he has been collaborating with Huawei Research and Development Center, Paris, as a Research Assistant, where he is developing neural architecture search methods for network reliability. He received several awards, such as the APRP Best M.Sc. Dissertation in Pattern Recognition 2019 Award and the ICANN21 "1st Springer & ENNS Best Paper Award."

Bruno Degardin received the B.Sc. and M.Sc. degrees from the Universidade da Beira Interior (UBI) in 2018 and 2020, respectively, where he is currently pursuing the Ph.D. degree with a focus on computer vision, emphasizing biometrics, human behavior analysis, and video understanding with already accepted and published works on international conferences and journals (*IJCB* and *WACV*). In the past, he has worked on research projects, such as BIODI, CovidSight, and PAPSE-UBI. Since 2020, he has been a Teaching Assistant at UBI and the Co-Founder and the CTO of DeepNeuronic (computer vision solutions Startup). He received several awards, such as the Best M.Sc. Thesis Fraunhofer Portugal Challenge (2020—3rd place) and research of practical utility (1st place—Biomedical World Hackathon 2020 and 1st place—UBIMedical Health Cup 2020).

Hugo Proença (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from the University of Beira Interior, Portugal, in 2001, 2004, and 2007, respectively. He is currently a Full Professor at the Department of Computer Science, UBI. He is also a Senior Researcher of the Instituto de Telecomunicações (IT) and a member of the SOCIA Laboratory—Soft Computing and Image Analysis Group. His research interests are artificial intelligence, computer vision, pattern recognition, biometrics, and visual surveillance. He was the Coordinating Editor of the IEEE Biometrics Council Newsletter and the Area Editor (ocular biometrics) of the IEEE BIOMETRICS COMPENDIUM journal. He is a member of the editorial boards of the *Image and Vision Computing*, IEEE ACCESS, and *International Journal of Biometrics*. Also, he served as a Guest Editor for Special Issues of the *Pattern Recognition Letters*, *Image and Vision Computing*, and *Signal, Image, and Video Processing* journals.