Human Action Recognition using BlazePose Skeleton on Spatial Temporal Graph Convolutional Neural Networks

Motasem S. Alsawadi
Department of Electronic and Electrical Engineering University College London
London, UK
motasem.alsawadi.18@ucl.ac.uk, malswadi@kacst.edu.sa

Miguel Rio
Department of Electronic and Electrical Engineering University College London
London, UK
miguel.rio@ucl.ac.uk

***Abstract*—The trend in multimedia transmission in social media has increased tremendously during the last decade and it is expected to continue growing during the next. Therefore, the need for new tools with the capacity of analyzing this kind of data grows accordingly. In this work, we implement the BlazePose skeleton topology into the ST-GCN model for action recognition. We test our experiments on the UCF-101 and HMDB-51 datasets. These are the first experiments of action recognition using the BlazePose skeleton upon these benchmarks. Moreover, we present an improved skeleton topology based on BlazePose that can enhance the performance achieved by its predecessor. By using the Enhanced-BlazePose topology presented in this study, we improved the results of the ST-GCN model on the UCF-101 benchmark more than 13% in accuracy performance. Finally, we have released the BlazePose skeleton data of the UCF-101 and HMDB-51 from our experiments to contribute future studies in the research community.**

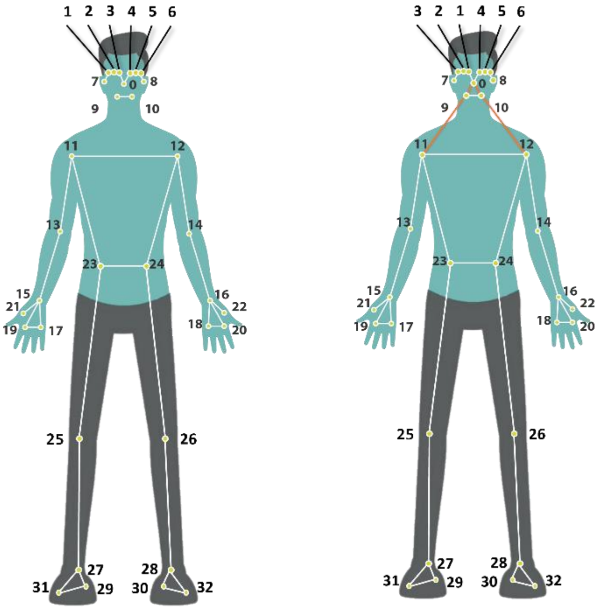
Keywords*—*BlazePose, Skeleton, Action Recognition, Graph Neural Networks, Spatial-Temporal Graph Convolutional Networks

I. Introduction

According to [1], the trend in multimedia transmission in social media has increased tremendously during the last decade and the rate of growth of visual data on the internet surpasses the rate of development of tools to interpret the information. Therefore, the need for new tools with the capacity of analyzing this kind of data grows accordingly.

In this study, we analyze the contents of videos by recognizing human actions. This approach comes with a series of difficulties. For instance, the capability of the model to represent the actions. Several proposals have been presented to address this issue. Among these, the point cloud-based, the RGB-based and the skeleton-based approaches have achieved the best performance [2]. The latter approach offers multiple advantages compared with the other alternatives. This solution represents the body of the person performing the action by the coordinates (either 2D or 3D) of a small set of landmarks located in the main joints of the body. Its background-free representation allows the classification algorithms to focus solely on the movement patterns of the main limbs during the activity. Consequently, the computational cost and the storage needed to model the actions is reduced considerably. For this reason, we chose this approach to represent the input for our system.

Multiple tools have been developed to extract the skeleton-data from images (and videos) during the last decade. The OpenPose system [3] is one of the most utilized for this aim. This tool allows to extract the skeleton of multiple persons in a single image. It is a portable solution able to run on Ubuntu, Windows, MacOS X. However, it is recommended to be used on GPU-enabled devices to achieve better performance. Alternatively, the BlazePose system [4] is a mobile-oriented skeleton extraction tool released recently by Google (shown in Fig. 1a). It offers a lightweight solution that provides a greater size of skeleton joints. The vast amount of data acquired with this tool allows an action recognition system to represent more accurately the movements of the body limbs during the activity prior the classification stage. As consequence, we have chosen the BlazePose system to extract the skeleton data from the input videos.



(a) BlazePose (b) Enhanced BlazePose

Fig. 1. Skeleton topologies used in this study. The four additional edges of the Enhanced BlazePose topology (marked in orange color) can describe more accurately the movements of a person.

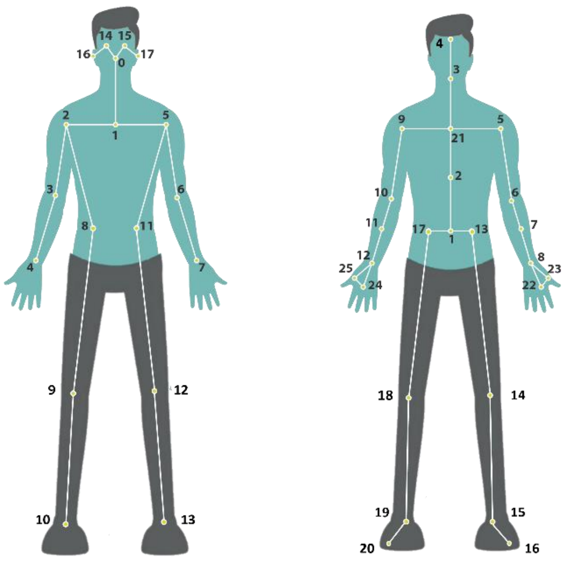
The second stage to achieve action recognition consists in the use of a classifier algorithm that can take the sequence of graphs as an input and provides the action performed in the video as an output. The most successful solutions include the use of Recurrent Neural Networks (RNNs) and Graph Neural Networks (GNNs) [5]. Given the nature of the sequence of graphs representation of the actions, it is intuitive to make use of the GNN-based solutions. To consider the relations between the joints inside the skeleton but also to analyze the patterns in the time domain, the Spatial-Temporal Graph Convolutional Neural Network (ST-GCN) has been proposed [6]. Because we intend to model the patterns in the movements over time, we find this model optimal for our aim.

In this work, we implement the BlazePose skeleton topology into the ST-GCN model for action recognition. We test our experiments on the UCF-101 [7] and HMDB-51 [8] benchmark datasets. To the knowledge of the authors, these are the first experiments of action recognition using the BlazePose skeleton upon these benchmarks. Moreover, we present an improved skeleton topology based on BlazePose (shown in Fig. 1b) that can enhance the performance achieved by its predecessor. In summary, the contributions of the present study are listed below:

- We present the first experiments of action recognition using the BlazePose skeleton upon the UCF-101 [7] and HMDB-51 [8] benchmark datasets.
- We present a novel skeleton topology, the Enhanced-BlazePose, that can capture the movements of a person more accurately that previous solutions.
- We provide a deep analysis of the BlazePose model performance for action recognition tasks.
- Additionally, we have released the skeleton data of the UCF-101 [7] and HMDB-51 [8] to contribute future studies in the research community (https://github.com/malswadi/blazepose-skeleton-hmdb-ucf).

II. Related Studies

To recognize human actions, the time dimension needs to be considered. For this reason, there have been several approaches to achieve this aim. For example, Simonyan and Zisserman [9] proposed to extract the spatial and temporal information of videos to recognize human actions using a two-stream CNN-based architecture. However, the computation needed to train these architectures is high compared with the GNN alternatives based upon skeleton data. To alleviate this problem, Li et al. [10] enhanced the 3D-CNN networks [11] with dynamic GCNs. They represent the feature maps from 3D-CNNs as interconnected nodes in a graph and learn the spatial-temporal relationship between the nodes (i.e., the edges) using a GCN. With this approach, they were able to reduce the computational cost needed considerably.



(a) OpenPose COCO

(b) NTU-RGB

Fig. 2. Reference skeleton topologies. Each alternative has a different amount of joints.

The ST-GCN model has been widely used in the recent years. Initially, it was used for action recognition with the use of the skeleton-data provided by the NTU-RGB [12] dataset and the Kinetics [13] dataset. This last information was extracted with the OpenPose system and further released by the authors as the Kinetics-skeleton dataset in [6]. The skeleton topologies provided in the Kinetics-skeleton and the NTU-RGB datasets are shown in Fig 2a and Fig 2b, respectively. The availability of this large action datasets to the public motivated the research community to continue with the improvements in this field. For instance, Shi et al. [14] used the Kinetics-skeleton dataset to propose a two- streamed architecture called 2s-AGCN (two-stream adaptive graph convolutional network) that is based upon the ST- GCN model for action recognition. Their solution can model both the bone and the joint information of the skeleton in separate streams using an attention approach. Finally, they performed late fusion to classify the activity. Inspired by this idea, Heidari and Iosifidis [15] also used this data and utilized the adaptive module presented in [14] to propose the spatio-temporal bilinear network (ST-BLN). Recently, Plizzari et al. [16] proposed the Spatial–Temporal Transformer network (ST- TR). In their work, they used the self-attention module used in Transformers [17] for natural language processing tasks and apply it for video analysis. However, all these solutions altered the ST-GCN architecture. In our previous work, we proposed a novel set of skeleton partitioning strategies for the ST-GCN model that could enhance the recognition performance of the baseline model [18]. Our approach in this work is to improve the ST-GCN model performance by changing the skeleton topology used as an input. With this solution, we aim to provide a simple and powerful alternative for all existing ST-GCN-based models improve their performance with no need of major changes in their architectures.

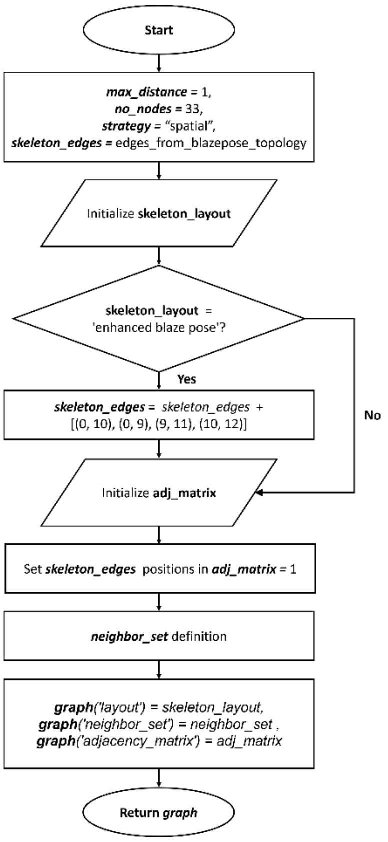
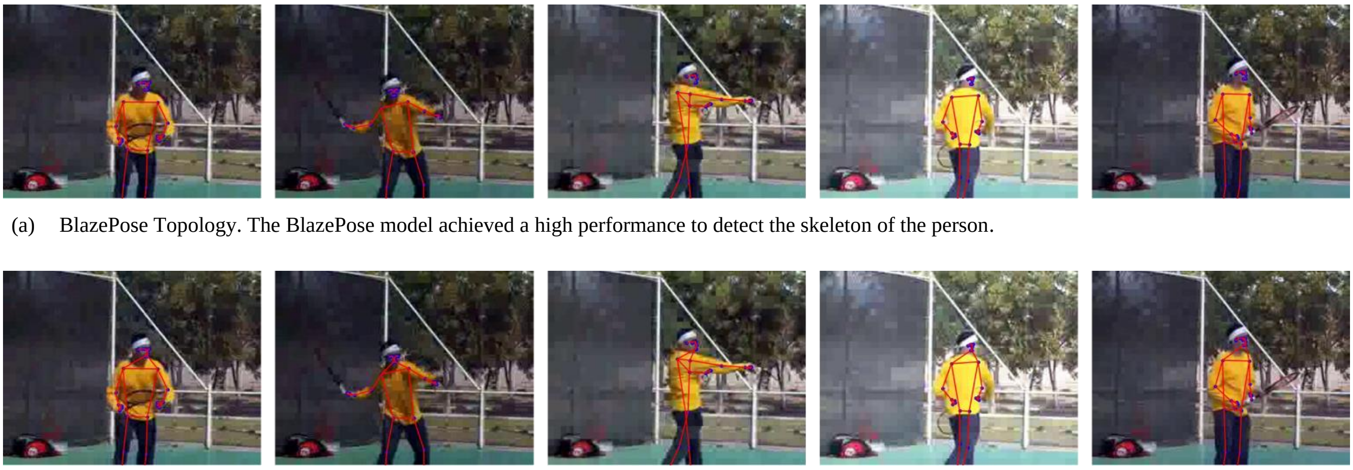

Start
max_distance = 1,
no_nodes = 33,
strategy = "spatial",
skeleton_edges = edges_from_blazepose_topology
Initialize skeleton_layout
skeleton_layout = 'enhanced blaze pose'?
Yes
No
skeleton_edges = skeleton_edges + [(0, 10), (0, 9), (9, 11), (10, 12)]
Initialize adj_matrix
Set skeleton_edges positions in adj_matrix = 1
neighbor_set definition
graph('layout') = skeleton_layout,
graph('neighbor_set') = neighbor_set ,
graph('adjacency_matrix') = adj_matrix
Return graph

Fig. 3. Flowchart for Enhanced BlazePose Topology definition



(a) BlazePose Topology. The BlazePose model achieved a high performance to detect the skeleton of the person.

(b) Enhanced-BlazePose Topology. Using four additional edges, the proposed topology is able to capture the movements of the head and torso more accurately.

Fig. 4. Skeleton output comparison with the different BlazePose-based topologies. In the figure, we selected a random sample of ’Tennis Swing’ action from the UCF-101 dataset. The frames are presented in order of appearance from left to right.

III. The Enhanced-BlazePose

The Enhanced-BlazePose topology (shown in Fig. 1b) was proposed to improve the ability of the BlazePose skeleton to represent the actions. Since the action representation has a large impact on the performance of any method for action recognition, we decided to focus on this stage to raise the performance of the ST-GCN model. The proposal has four additional edges to the existing joints that can provide a more accurate representation of the relation between the head and the torso of the person performing the action. These new edges connect the joints located in the mouth edges with those located in the shoulders.

The graph definition of the Enhanced-BlazePose topology for the ST-GCN model is described in Fig. 3. First, we set the initial graph topology variables. The max distance and the strategy variables are used to define the subsets of nodes where the convolution operation is going to be performed. Similar to the pixels of an image, these neighbor sets are used as an input to each convolution kernel. To reduce the computational load of the convolution operation, we set this value to 1. Meaning that we only have the center node (or the root node) and its direct neighbors in each subset. In the same stage, we define the skeleton edges. The skeleton edges variable is a list containing a set of tuples of two elements each. These elements correspond to the joint indexes of the tail and head of the edge of the skeleton graph. For instance, consider the skeleton edges elements for the nodes of the shoulders and hips of the BlazePose topology shown in Fig. 1a. These elements of the skeleton edges list are (11, 12), (11, 23), (12, 24) and (23, 24). Second, we define the skeleton layout name string. Third, if the topology selected corresponds to the Enhanced-BlazePose topology, then we add the four additional edges shown in Fig. 1b. Third, we initialize the square adjacency matrix adj matrix of size no nodes with zero values. For instance, the no nodes = 33 for the BlazePose topology. Fourth, we set to 1 the positions defined in the skeleton edges list of the adjacency matrix. Finally, we define the neighbor sets considering the adjacency matrix, the max distance and the strategy variables.

IV. Experimental Settings

A. Preprocessing

The experiments on both benchmarks were conducted following the same procedure. First, we extracted the 2D BlazePose skeleton data from the videos. To achieve this, we used the MediaPipe Pose Python API released publicly in [19]. This tool provides 33 landmarks of the main joint locations as they are shown in Fig. 1a. For our purposes, we utilized the x and y coordinates of the location of each joint. Additionally, we also considered the confidence score c of each joint provided by MediaPipe Pose. The values for x and y can vary from [0, 1] if the joint location is predicted inside the video frame. Otherwise, these values are outside that range depending on the region. On the other hand, the values for c strictly vary from [0, 1]. As a result, we obtained a separate file with the skeleton data (formatted as JSON) for each video.

Given that both benchmark datasets (the UCF-101 [7] and HMDB-51 [8]) have videos from different sources, the length (and the size) of the videos vary. Thus, the third stage consisted of setting a fixed length to them. To follow the baseline model experiments settings in [6], we set the video lengths to be 300 frames. When the videos length did not reach the desired length, we repeated the initial frames the times needed. Alternatively, we randomly eliminated the difference in frames.

B. Training

On the fourth and final stage, we trained the ST-GCN model with each subset using the PyTorch framework [20] for deep learning. We used the spatial configuration partitioning to map each joint with a label in the GCN (for further details, please refer to [6]). We applied stochastic gradient descent with learning rate decay as an optimization algorithm for 80 epochs. We decrease the learning rate value by 10% every 10th epoch. For regularization, we used a weight decay value of 10−4. Finally, we vary the batch size from 32, 64 and 128. These experiments were performed on 4 GPUs (NVIDIA Tesla V100) with 32GB.

C. *Evaluation*

We computed the sensitivity and accuracy metrics to assess the performance of the trained models. The first metric describes the ratio of the test set predicted as positive with respect to the positive ground truth labels. On the other hand, the accuracy describes the overall ability to classify each label to its corresponding class. The equation to compute the sensitivity metric is shown in Eq. 1.

$$Sensitivity = \frac{TP}{TP+FN} \qquad (1)$$

Where TP (true positives) is the count of samples predicted as positive with positive ground truth label and FN (false negatives) is the count of samples predicted as negative by the model, but their ground truth label is positive.

D. *Datasets*

We test our experiments on the UCF-101 [7] and HMDB-51 [8] benchmark datasets:

1) UCF-101: The UCF-101 [7] is one the most widely used for human action recognition. It consists of 13,320 clips extracted from YouTube and classified into 101 action classes. The length of the samples varies from 1.06 sec to 71.04 sec and share the same resolution of 320×240 pixels.

2) *HMDB-51:* The HMDB-51 [8] data consists in video samples gathered from YouTube, movies, Google videos, among others. It provides a total of 6,766 video clips of 51 different classes. In opposite to the UCF-101 [7] counterpart, these videos vary their resolution. For that reason, the height of all the samples was scaled to 240 pixels, and the width has been scaled to maintain the original video ratio.

V. Results and Discussion

Given that these are the first results of the ST-GCN model using the BlazePose skeleton upon the UCF-101 [7] and HMDB-51 [8], we compare our results in terms of accuracy with models that used these benchmarks for action recognition purposes. These models are the grouped attention graph convolutional networks (GAGCNs) [21] and the ST-GCN model with the COCO skeleton topology from OpenPose [22] [23] (Fig. 2a).

1) UCF-101: Using the BlazePose topology on this dataset, we achieved an accuracy performance of 59.34%. According to their sensitivity values, the classes that achieve the best performance are 'tennis swing', 'floor gymnastics' and 'playing daf'. We show a random sample from the 'tennis swing' class in Fig. 4a. As it can be noticed, the BlazePose system achieved a high performance on this class given that the whole body of the person is shown in most of the frames of the videos. On the other hand, this model has difficulties in recognizing the actions of 'swing', 'parallel bars' and 'field hockey penalty'. We show the comparison of the sensitivity performance of these classes in the confusion matrix shown in Fig. 5. In the matrix, the ground truth labels and the predicted labels are shown as rows and columns, respectively. Since the amount of action classes in the UCF-101 dataset is large (101), we reduced the matrix to a subset that can demonstrate the sensitivity performance of the overall model. Therefore, we added the 10 classes that achieved the best results along with the 10 classes that output the lowest performance. The values regarding the rest of the classes (81 classes) are represented in the class labeled as 'other'.

We improved these results to 64.2% accuracy with the use of the Enhanced-BlazePose alternative. The classes of 'tennis swing', 'playing cello', 'skateboarding', 'shotput' and 'table tennis shot' were able to rise their sensitivity to 100%. In other words, that the model was able to predict all the samples of these classes with the correct label using the additional edges added to the BlazePose topology. As it can be noticed in Fig.4b, these edges allow the ST-GCN to model the rotation of the torso more accurately while doing the swing of the racket. To provide a comparison with the BlazePose topology output, we show the sensitivity performance using this topology in Fig. 6. Similar to Fig. 5, we added the 10 classes that achieved the best results along with the 10 classes that output the lowest performance, including the class 'other'.

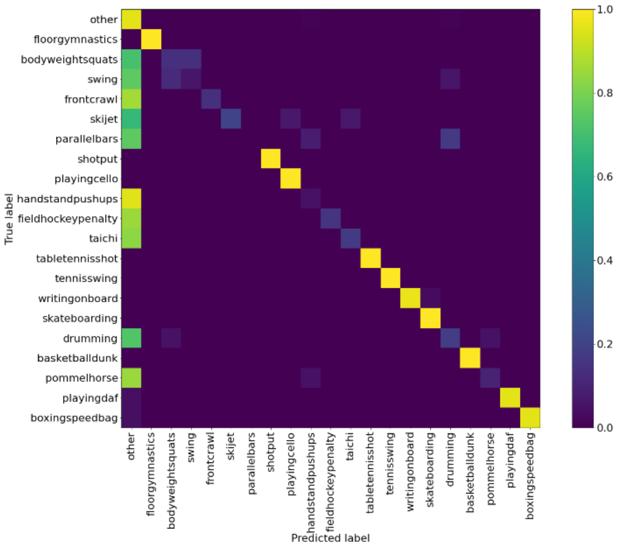


Fig. 5. Confusion matrix with the BlazePose topology performance upon the UCF-101 dataset.

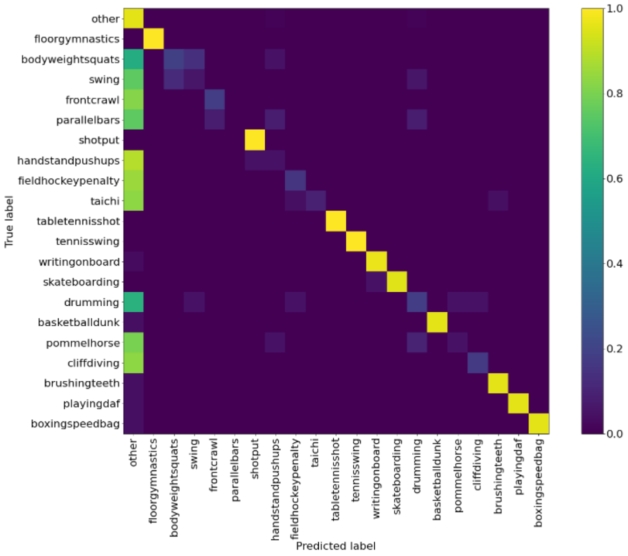


Fig. 6. Confusion matrix with the Enhanced BlazePose topol- ogy performance upon the UCF-101 dataset.

TABLE I. UCF-101 BENCHMARK ACCURACY PERFORMANCE COMPARISON

Method	Accuracy (%)
GAGCN [21]	35.6
OpenPose + ST-GCN [22]	50.53
BlazePose + ST-GCN (Ours)	59.34
Enhanced-BlazePose + ST-GCN (Ours)	64.2
OpenPose + ST-GCN + Index Split [23]	72.31

TABLE II. HMDB-51 BENCHMARK ACCURACY PERFORMANCE COMPARISON

Method	Accuracy (%)
GAGCN [21]	32.5
OpenPose + ST-GCN	36.59
BlazePose + ST-GCN (Ours)	40.07
Enhanced-BlazePose + ST-GCN (Ours)	44.1
OpenPose + ST-GCN + Index Split [23]	47.69

The comparison with other models that used the UCF- 101 for action recognition purposes is shown in Table I. The ST-GCN model using an index split strategy [23] still have the highest accuracy for this benchmark for action recognition purposes. However, our main reference is ST-GCN model on the OpenPose COCO topology using the spatial partitioning strategy in [22], since the approach to achieve the action recognition is the most similar. By changing the skeleton topology used as an input from OpenPose COCO to BlazePose, we were able to increase almost 9% the accuracy performance of the ST-GCN. Moreover, we improved this result by almost 5% with the additional edges of the Enhanced- BlazePose topology.

2) HMDB-51: Using the BlazePose topology on this dataset, we achieved an accuracy performance of 40.07%. For this model, the classes that achieve the best sensitivity performance are 'climb stairs', 'clap' and 'kiss'. Alternatively, the actions of 'smoke', 'cartwheel', 'wave', and 'sit' achieved the lowest performance.

Similar to the results obtained upon the UCF-101 dataset, we improved the overall accuracy of the model to 44.1% by choosing the Enhanced-BlazePose topology alternative upon the HMDB-51 benchmark. This topology allowed the classes 'hug', 'cartwheel' and 'kick ball' to increase its sensitivity by 40%, 35.3% and 25%, respectively.

The comparison with other models that used the HMDB-51 for action recognition purposes is shown in Table II. Similarly to the results obtained on the UCF-101 dataset, we outper- formed the outcome achieved on the ST-GCN model using the OpenPose COCO with the spatial configuring partitioning strategy.

After the experimentation process we found a shortcoming in the use of the BlazePose-based topologies for action recognition. Given that the BlazePose model utilizes a face detector to localize the person [4], it is strictly necessary for the face to appear in the frame to achieve a satisfactory outcome for action recognition. Nevertheless, these are great solutions to recognize actions that overcome this constraint.

VI. CONCLUSIONS

In this study, we present the first experiments of action recognition using the BlazePose skeleton upon the UCF-101 [7] and HMDB-51 [8] benchmark datasets. We demonstrate that the performance of the ST-GCN model can be improved solely by changing the skeleton topology in its input. Moreover, we show that the performance can be further risen if the inner skeleton topology is modified, as we proposed in the Enhanced-BlazePose topology. Given the results provided in Table I and Table II, it can be noticed that it is possible to increase the performance of the ST-GCN model by changing the partitioning strategy [6] [18] used in the convolution operation. Therefore, we foresee this opportunity as a future work.

ACKNOWLEDGMENT

The authors acknowledge the support of King Abdulaziz City of Science and Technology (KACST).

REFERENCES

[1] Y. Chen, K. Sherren, M. Smit, and K. Y. Lee, "Using social media images as data in social science research," New Media & Society, p. 14614448211038761, 2021.

[2] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," arXiv preprint arXiv:2012.11866, 2020.

[3] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "OpenPose: Real-time Multi-Person 2D Pose Estimation Using Part Affinity Fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172–186, 2021.

[4] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-time Body Pose tracking," arXiv:2006.10204, 2020. [Online]. Available: https://arxiv.org/abs/2006.10204

[5] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1227–1236.

[6] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," arXiv, 2018.

[7] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, 2012.

[8] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in 2011 International conference on computer vision. IEEE, 2011, pp. 2556–2563.

[9] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," Advances in Neural Information Processing Systems, vol. 1, Jun 2014.

[10] J. Li, X. Liu, J. Xiao, H. Li, S. Wang, and L. Liu, "Dynamic Spatio-Temporal Feature Learning via Graph Convolution in 3D Convolutional Networks," in 2019 International Conference on Data Mining Workshops (ICDMW), 2019, pp. 646–652.

[11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," Proceedings of the IEEE International Conference on Computer Vision, vol. 2015 Inter, pp. 4489–4497, 2015.

[12] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," in Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 2016, pp. 1010–1019.

[13] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, and P. Natsev, "The kinetics human action video dataset," arXiv:1705.06950, 2017. [Online]. Available: https://arxiv.org/abs/1705.06950

[14] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12 026–12 035.

[15] N. Heidari and A. Iosifidis, "On the spatial attention in spatio-temporal graph convolutional networks for skeleton-based human action recognition," in 2021 International Joint Conference on Neural Networks (IJCNN), Virtual, 2021, pp. 1–7.

[16] C. Plizzari, M. Cannici, and M. Matteucci, “Spatial temporal transformer network for skeleton-based action recognition,” in International Confer- ence on Pattern Recognition. Springer, 2021, pp. 694–701.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, 2017, pp. 5998–6008.

[18] M. S. Alsawadi and M. Rio, “Skeleton split strategies for spatial temporal graph convolution networks,” Computers, Materials & Continua, vol. 71, no. 3, pp. 4643–4658, 2022.

[19] “MediaPipe Pose.” [Online]. Available: https://google.github.io/mediapipe/solutions/pose

[20] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.

[21] Y. Xu, C. Han, J. Qin, X. Xu, G. Han, and S. He, “Transductive zero-shot action recognition via visually connected graph convolutional networks,” IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 8, pp. 3761–3769, 2020.

[22] W. Zheng, P. Jing, and Q. Xu, “Action Recognition Based on Spatial Temporal Graph Convolutional Networks,” in Proceedings of the 3rd International Conference on Computer Science and Application Engineering, 2019, pp. 1–5.

[23] M. S. Alsawadi and M. Rio, “Skeleton-Split Framework using Spatial Temporal Graph Convolutional Networks for Action Recognition,” in 2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART), Paris, France, 2021, pp. 1–5.