AN EFFICIENT FRAMEWORK FOR HUMAN ACTION RECOGNITION BASED ON GRAPH CONVOLUTIONAL NETWORKS

Nikolaos Kilis *Christos Papaioannidis* *Ioannis Mademlis* *Ioannis Pitas*

Department of Informatics, Aristotle University of Thessaloniki, Greece

ABSTRACT

This paper presents a novel framework for skeleton-based Human Action Recognition (HAR) based on Graph Convolution Networks (GCNs). The proposed framework aims to increase human action recognition performance of GCN-based methods by incorporating a missing-joint-handling pre-processing step and a novel adjacency matrix construction method in a single human action recognition pipeline. The missing-joint-handling pre-processing step is utilized to infer missing data in the input sequence, which may occur due to imperfect skeleton extraction, based on imputation methods. The novel adjacency matrix construction method is executed offline to compute an improved weighted adjacency matrix specifically designed for HAR, which is utilized in every layer of the employed GCN. Moreover, both the pre-processing step and the adjacency construction method can be utilized along with any GCN architecture, allowing any GCN-based HAR method to be employed in the proposed framework. Experimental evaluation on two public datasets indicate favorable human action classification scores compared to the employed baseline and all competing methods both for 2D and 3D skeleton-based human action recognition, while using a GCN architecture with less learnable parameters.

***Index Terms*—** Skeleton-based human action recognition, Graph Convolutional Networks, graph node clustering, feature imputation.

1. INTRODUCTION

Human Action Recognition (HAR) objective is to identify human actions that are depicted in videos [1, 2]. In general, HAR research involves enhancing monitoring processes from environmental, spatial and temporal data. However, due to background variation and illumination changes, human actions are typically represented by skeleton data sequences that derive by extracting the 2D or 3D human skeleton spatial coordinates from video frames using 2D/3D human pose estimation methods [3, 4]. Skeleton-based HAR methods [5, 6, 7, 8] analyze these sequences to classify each sequence to the corresponding action that is performed in that sequence.

However, skeleton-based HAR also faces some challenges that hinder skeleton-based HAR methods, such as camera viewpoint variations, potential body parts occlusions and noisy human skeleton data. Many of these issues can be alleviated by representing actions as graphs that change over time. This can be achieved by representing a human skeleton as a graph: the human skeleton joints serve as graph nodes and together with edges connecting those nodes, they constitute the human skeleton graph.

Graph Convolution Networks (GCNs) [9] were successfully utilized to process human skeleton graphs for human action recognition [5, 10, 11, 12]. One example is [5], which proposed a Spatio-Temporal Graph Convolutional Network (ST-GCN) that acts on human graph sequences to compute spatio-temporal features that enable human action recognition. An important component of such GCN-based HAR methods is the adjacency matrix, which encodes the human graph structure and is utilized in each graph convolutional block of GCNs. Different ways of constructing the adjacency matrix can have a great impact on GCN-based HAR performance [12]. Moreover, another issue that may cause the degradation of GCN-based and skeleton-based HAR performance is the fact that some skeleton joints may be often missing from the 2D/3D skeletons (e.g., due to occlusions or imperfect skeleton extraction from the 2D/3D human pose estimation methods) that are used to construct the input human graph sequences.

In this direction, a GCN-based human action recognition framework is proposed, which aims to increase GCN-based HAR performance by addressing both the missing joints and the adjacency matrix construction problems. More specifically, the missing joints problem is addressed by introducing a pre-processing step to effectively infer missing data in the 2D/3D skeletons, which are used to construct the human graphs, based on feature imputation (FI) algorithms. Furthermore, a novel adjacency matrix construction method is utilized to construct an improved, weighted adjacency matrix for GCN-based HAR which considers the human skeleton/graph as a set of joint/node clusters. This is inspired from the fact that a human action can be modeled from the actions of small groups/clusters (e.g., arms, legs), each of which has a dif-

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871479 (AERIAL-CORE).

ferent contribution to the complete action. Finally, both the missing-joint-handling pre-processing step and the adjacency matrix construction method are compatible with any GCN architecture, allowing any GCN-based HAR method to be utilized in the proposed framework.

The superiority of the proposed framework is verified by conducting experiments both on 2D and 3D skeleton-based HAR datasets, demonstrating increased performance compared to the baseline and all competing HAR methods, in both cases.

2. SKELETON-BASED HUMAN ACTION RECOGNITION

In skeleton-based Human Action Recognition, deep learning approaches [5, 6] outperformed methods that used handcrafted features to model the natural connections between the skeleton graph nodes [13, 14]. Early deep learning methods represented skeletons as 2D or 3D Euclidean grids, which were fed into Recurrent Neural Networks (RNNs) [15, 16] or Convolutional Neural Networks (CNNs) [17, 18]. For example, DD-Net [6] utilized 2D human skeleton data in a lightweight CNN architecture to encode slow and fast body joint movements in an action and compute pairwise body joint distances, which were exploited to improve action recognition performance.

With the rise of Graph Convolutional Networks (GCNs) [9], 2D or 3D skeletons could also be represented as graphs and thus, perform the operation of convolution on non-Euclidean grids, resulting in more flexible and efficient models [5, 10, 11, 12]. For example, the human skeleton was represented as a directed acyclic graph in [12], which was subsequently utilized in a Directed Graph Neural Network (DGNN) to compute features for action recognition. Spatio-Temporal Graph Convolutional Network (ST-GCN) [5] utilized GCNs to encode both spatial and temporal information from human graph sequences. More specifically, its input consists of two parts: the adjacency matrix and the human graph sequence that represents a human action. The human graphs in the sequence are connected via undirected and unweighted edges that bridge the same nodes between consecutive human graphs. As a result, the entire action is represented by one spatio-temporal graph. The input action graph is then passed through a number of spatio-temporal graph convolutional layers to obtain spatio-temporal human action features. These features are finally utilized by a simple classifier to predict human action classes.

In an orthogonal research direction, the proposed HAR framework introduces a missing-joint-handling pre-processing step and an adjacency matrix construction method to complement existing GCN-based HAR methods, towards increasing their performance.

3. EFFICIENT GCN-BASED HAR RECOGNITION

This work introduces an efficient framework that aims to improve GCN-based human action recognition. The proposed framework consists of three building blocks: the missing-joint-handling pre-processing step, the proposed adjacency matrix construction method and the human action classification/recognition GCN.

Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be a human graph, where $\mathcal{V}$ is a set of K body joints/nodes and $\mathcal{E}$ is a set of M bones/edges. Also let $\mathbf{S} \in \mathbb{R}^{N \times K \times D}$ be a tensor representing an input human graph sequence, where N is the number of human graphs comprising the sequence, K is the number of the human graph nodes (body joints) and D is the graph nodes' feature vector (spatial coordinates) dimensionality ($D = 2$ and $D = 3$ in the 2D and 3D skeleton cases, respectively). The pre-processing step is applied to the input skeleton/graph sequence $\mathbf{S}$ to fill any missing data before given to the employed GCN for action classification/recognition. In addition, an improved weighted adjacency matrix is calculated offline using the proposed adjacency matrix construction method and is subsequently utilized in every graph convolution block of the employed GCN, which can be any human action recognition GCN from the literature.

3.1. Missing Joint Handling with Feature Imputation

In order to obtain the input sequence $\mathbf{S}$ that is fed to the GCN for human action recognition/classification, the 2D/3D human skeletons $\mathbf{X} \in \mathbb{R}^{K \times D}$ need to be extracted using sensors or human pose estimation methods [3, 4]. This skeleton extraction process often leads to missing data (body joint 2D/3D coordinates) in the input sequence, e.g., due to body joint occlusions.

Missing joints are typically handled by setting their respective coordinates to zero. However, this is not an optimal approach, since these joints may be crucial for recognizing specific actions. In contrast, the proposed method handles missing joints by utilizing a pre-processing step that performs feature imputation (FI). That is, feature imputation is performed by separately processing each input sequence $\mathbf{S}$, which may contain human graphs with missing node features (body joint spatial coordinates), using the Multivariate Imputation by Chained Equations (MICE) [19] algorithm.

Initially, all missing graph node features in the input sequence are labeled as "missing" and are replaced with the mean feature of this specific node in the sequence. After this initialization, each human graph that contains node features labeled as "missing" is processed sequentially. For each node feature labeled as "missing" in the human graph that is currently being processed, a linear regression model is first fitted by utilizing all remaining human graphs in the sequence. The fitted model is subsequently used to infer the feature values for this specific node. This process is repeated for all human

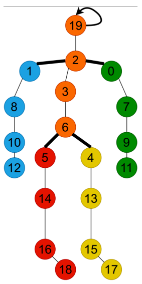

Fig. 1. Louvain [20] output at a lower hierarchical level. Different colors indicate different node clusters.

graphs with missing node features in the sequence, until all node features labeled as "missing" are inferred.

3.2. Weighted Adjacency Matrix Construction Method

Defining the adjacency matrix is an important step of GCN-based methods. The proposed adjacency matrix construction method aims to construct an optimal adjacency matrix for HAR, based on the idea that human actions can be more efficiently modeled from the movements of many small body joint clusters.

In this direction, the first step is to define the body joint clusters. Given that the node features of the human graph $\mathcal{G}$ are the 2D/3D spatial coordinates of the corresponding body joints, node/body joint clustering is performed by utilizing the human graph in a node clustering algorithm. Specifically, the Louvain [20] algorithm is employed, which can find similar groups of nodes in a hierarchical manner, offering multiple node clustering options. For example, a higher clustering level defines two clusters on the human graph that correspond to the upper and lower body, while a lower clustering level results in groups of nodes that corresponds to human body parts (e.g., arms, legs, etc.) as illustrated in Fig. 1. In the proposed adjacency matrix construction method the lower clustering level was selected, as it can provide richer information for action recognition compared to the higher clustering level. As a result, five graph node clusters were selected, representing left/right hand, left/right leg and body, respectively.

After the definition of the node clusters, the improved weighted adjacency matrix $\mathbf{A}_w \in \mathbb{R}^{K\times K}$ is constructed as follows: a group of weights are set to the edges connecting nodes that belong to the same node cluster, while a separate group of weights are set to edges connecting nodes that belong to different node clusters. The improved weighted adjacency matrix is subsequently equipped with self-connections, $\tilde{\mathbf{A}}_w = \mathbf{A}_w + \mathbf{I}$, before being normalized using:

$$\hat{\mathbf{A}}_w = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}_w\tilde{\mathbf{D}}^{-1/2}, \quad (1)$$

where $\tilde{\mathbf{D}} \in \mathbb{R}^{K\times K}$ is the diagonal degree matrix of $\tilde{\mathbf{A}}_w$, $\tilde{\mathbf{D}}_{i,i} = \sum_j \tilde{\mathbf{A}}_{w_{i,j}}$.

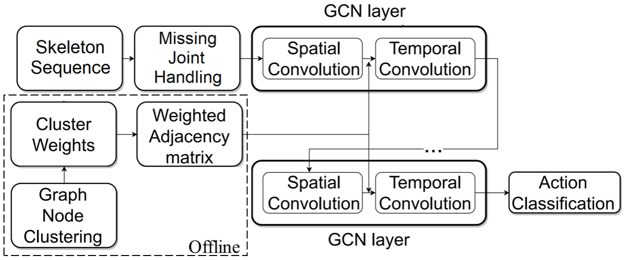

Fig. 2. The proposed human action recognition framework.

3.3. Human Action Recognition

The improved weighted adjacency matrix $\hat{\mathbf{A}}_w$ obtained by the proposed adjacency matrix construction method, along with the human graph sequence $\mathbf{S}$ are given to the employed GCN for action classification/recognition. The employed GCN architecture is the one used in [5], however any GCN architecture could be utilized in its place. The general scheme of the proposed framework is shown in Figure 2. Moreover, since different node/body joints clusters may have different levels of importance for specific actions, a learnable mask $\mathbf{M} \in \mathbb{R}^{K\times K}$ is also used in each layer of the employed GCN to scale the contribution of each node feature. If $\mathbf{X}_t^l$ represents the human graph at time t and GCN layer l, each GCN layer performs spatio-temporal graph convolution, formulated as:

$$\mathbf{X}_t^{l+1} = \sigma\Bigg(\sum_{\tau=t-\mu}^{t+\mu} \mathbf{W}_{t-\tau}^l \sigma(\mathbf{M}_l \otimes \hat{\mathbf{A}}_w \mathbf{X}_\tau^l \mathbf{W}_s^l)\Bigg), \quad (2)$$

where $\mathbf{W}_s^l$ is the spatial convolution kernel and $\mathbf{W}_t^l$ is the temporal convolution kernel of length $2\mu + 1$ at layer l, while σ is a non-linear activation function and $\otimes$ denotes the Hadamard product. Moreover, it can be easily seen that $\mathbf{X}^0 = \mathbf{S}$. Before passing $\mathbf{X}^{l+1}$ to the next GCN layer, a residual connection is also added. Therefore, the input of the $(l+2)^{th}$ layer equals to $\mathbf{X}^{l+1} + f(\mathbf{X}^l)$, where f is represented by a single convolutional layer.

Finally, after a specified number of GCN layers, graph classification is performed using a simple SoftMax classifier, which outputs the final action classes.

4. EXPERIMENTAL EVALUATION

The employed GCN architecture consists of three GCN layers, resulting in a very lightweight architecture with 0.53M learnable parameters. The number of filters of the first, second and third layer is 90, 100 and 180, respectively. The temporal kernel size was set to 9 in all cases. Dropout with 0.5 probability is also added after each GCN layer to avoid overfitting. The strides of the first and second temporal convolution layers was set to 1, while for the last one it was set to

Table 1. Evaluation using the 1^{st} protocol of JHMDB [22] dataset.

Method	FI	#Parameters	Acc (%)
ST-GCN [5]	no	0.60M	65.45
ST-GCN [5]	yes	0.60M	66.21
DD-Net [6]	—	1.82M	77.16
Proposed	no	0.53M	77.05
Proposed	yes	0.53M	**79.89**

2. The non-linear activation function σ is the rectified linear unit (ReLU). The model was trained using the Adam [21] optimizer. Note that all experiments were conducted using a GeForce RTX 2060 graphics card.

The proposed framework was evaluated on two skeleton-based HAR datasets: a) the Joint-annotated Human Motion Data Base (JHMDB) dataset [22] and b) the MSR Action 3D dataset [23]. JHMDB dataset consists of 928 video clips of 21 human actions, where each video frame is annotated with the corresponding 2D human skeletons. Two evaluation protocols were defined depending on the availability of the 2D skeleton sequences during the inference stage. In the first protocol, the 2D skeleton sequence that contains the full action (from the start till the end) is available before inference, while in the second protocol only a part of the action takes place in the available sequence. Note that in both cases, the GCN models are trained with 2D skeleton sequences that contain full actions.

The MSR Action 3D dataset contains 20 human actions performed by 10 different subjects. The valid number of skeleton data sequence is 557. The available 3D skeleton annotations are used to evaluate the proposed framework also on 3D skeleton-based HAR. The dataset is divided into 3 subsets (AS1, AS2, AS3), each one containing 8 of the 20 human actions [23, 24]. The training and the test set splits contain the 1/3 and 2/3 of the total number of actions, respectively. Both sets include actions performed by all 10 subjects.

The proposed method was compared against the baseline ST-GCN [5] and the state-of-the-art DD-Net [6] using the first evaluation protocol of JHMDB dataset. The comparison results presented in Table 1 show that the proposed method outperformed both ST-GCN and DD-Net, improving HAR accuracy up to 14% and 2.7%, respectively, despite the fact that the GCN architecture employed in the proposed framework has less learnable parameters. Furthermore, it is evident that the effect of feature imputation (FI) is positive both for the proposed method, as well as for ST-GCN baseline.

The proposed method was also evaluated using the second evaluation protocol of JHMDB dataset, which simulates a real-world scenario where 2D skeletons are obtained by processing an RGB camera feed using real-time 2D human pose estimation methods [3]. This scenario was implemented using a sliding window for partitioning the 2D skeleton sequences. Two cases were explored, having 0% and 30% overlapping between consecutive sliding windows, respectively. The results reported in Table 2 show that the proposed method again outperformed the ST-GCN baseline by a considerable margin in all cases.

Table 2. Evaluation using the 2^{nd} protocol of JHMDB [22] dataset.

Method	FI	Overlap	Acc (%)
ST-GCN [5]	no	0%	59.00
ST-GCN [5]	yes	0%	59.38
Proposed	no	0%	65.45
Proposed	yes	0%	**66.77**
ST-GCN [5]	no	30%	58.60
ST-GCN [5]	yes	30%	58.09
Proposed	no	30%	61.96
Proposed	yes	30%	**62.60**

Table 3. Evaluation on the MSR Action 3D dataset [23] dataset.

Method	AS1	AS2	AS3	Avg
Li et al. [23]	89.50%	89%	96.30%	91.60%
Chen et al. [24]	97.30%	96.10%	98.70%	97.40%
Ilias et al. [25]	98.68%	96.96%	96.99%	97.54%
Xu et al. [26]	98.00%	96.70%	**100**%	98.20%
Jin et al. [27]	98.00%	97.40%	99.30%	98.20%
ST-GCN [5]	98.63%	98.04%	98.66%	98.44%
Luo et al. [28]	**100**%	98.70%	**100**%	98.90%
Proposed	99.33%	**100**%	99.33%	**99.55**%

Finally, the proposed method was evaluated on 3D skeleton-based HAR using the MSR Action 3D dataset to verify its ability to successfully handle more complex data. The performance comparison between the proposed method and all directly comparable competing methods, concerning model complexity (number of trainable parameters), that used the same data splits is presented in Table 3. The proposed method manages to achieve best average HAR accuracy, outperforming both the ST-GCN baseline and all competing methods.

5. CONCLUSIONS

In this work, a skeleton-based human action recognition framework was proposed, aiming to increase human action recognition performance of GCN-based methods. It consists of three building blocks: a missing-joint-handling preprocessing step, an improved adjacency matrix construction method and a human action recognition GCN. The missing-joint-handling pre-processing step is applied on the input human graph/skeleton sequence to infer any missing node/body joints features before it is fed to the GCN. The adjacency construction method is executed offline to construct an improved weighted adjacency matrix, based on the assumption that a human action can be efficiently modeled from movements of a set of small body joint clusters. The resulting adjacency matrix is then utilized in each layer of the human action recognition GCN. By incorporating these two steps in a human action recognition pipeline, the proposed framework increased human action recognition accuracy both for 2D and 3D skeleton-based HAR, while using a lightweight GCN architecture with less learnable parameters than competing methods.

6. REFERENCES

[1] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[2] Y. Wang, Y. Xiao, F. Xiong, W. Jiang, Z. Cao, J. T. Zhou, and J. Yuan, "3dv: 3d dynamic voxel for action recognition in depth video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[3] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[4] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[5] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[6] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *Proceedings of the ACM Multimedia Asia*. 2019.

[7] C. Papaioannidis, D. Makrygiannis, I. Mademlis, and I. Pitas, "Learning fast and robust gesture recognition," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2021.

[8] D. Makrygiannis, C. Papaioannidis, I. Mademlis, and I. Pitas, "Optimal video handling in on-line hand gesture recognition using deep neural networks," in *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021.

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

[10] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[11] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[14] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[15] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] T. Soo Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

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

[19] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: what is it and how does it work?," *International Journal of Methods in Psychiatric Research*, vol. 20, no. 1, pp. 40–49, 2011.

[20] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, pp. P10008, 2008.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[23] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010.

[24] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps," *Journal of Real-time Image Processing*, vol. 12, no. 1, pp. 155–163, 2016.

[25] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "Pose-based human action recognition via sparse representation in dissimilarity space," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 12–23, 2014.

[26] H. Xu, E. Chen, C. Liang, L. Qi, and L. Guan, "Spatio-temporal pyramid model based on depth maps for action recognition," in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2015.

[27] K. Jin, M. Jiang, J. Kong, H. Huo, and X. Wang, "Action recognition using vague division dmms," *The Journal of Engineering*, vol. 2017, no. 4, pp. 77–84, 2017.

[28] J. Luo, W. Wang, and H. Qi, "Group sparsity and geometry constrained dictionary learning for action recognition from depth maps," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.