

Late Fusion of Graph Convolutional Networks for Action Recognition Using UAV Bone Data

Dinh-Tan Pham(✉)

International School, Vietnam National University, Hanoi, Vietnam
tanpd@vnu.edu.vn

Abstract. Unmanned Aerial Vehicle (UAV) is becoming popular for surveillance, search, and rescue operations. Human action recognition (HAR) using UAV data is an important task that allows human behavior to be detected instantly from the visual data. UAVs' attitudes and motions make HAR for UAVs challenging. In recent years, Graph Convolutional Network (GCN) architectures have been mainstream for skeleton-based HAR. This paper proposes a Late Fusion GCN (LF-GCN) framework by applying late fusion to three graph-based deep learning architectures. The framework combines the advantages of different graph-based deep-learning models. The proposed model outperforms existing GCN schemes on the large-scale UAV-Human dataset.

Keywords: action recognition · bone data · deep learning

1 Introduction

Nowadays, Unmanned Aerial Vehicle (UAV) equipped with cameras is used to examine significant areas for surveillance, search, and rescue. Understanding human behavior using UAV-captured data has emerged recently. While there has been immense work on ground-camera data, study on UAV data is limited. For ground-camera scenes, human body size takes up a high portion of the scenes. However, compared to ground-camera data, human action recognition (HAR) from UAV data is more complicated. The difficulty arises from various UAV conditions, including altitude, motion, weather, and geography. A large-scale benchmark dataset, namely UAV-Human, is introduced in [6] to facilitate the development and evaluation of deep learning models. Various modalities, including RGB, optical flows, and skeleton data, are collected for HAR. Skeleton-based HAR has drawn much interest lately because of its compactness and robustness [3]. A variety of skeleton-based methods for HAR have been proposed [8].

While many methods for HAR on datasets captured by ground cameras have been proposed, the performance of these methods on UAV-Human has yet to be thoroughly studied. Because of their quick movements and constant changes in height and attitude while in flight, UAVs tend to capture video sequences with more diverse and unique viewpoints, noticeable motion blurs, and a more comprehensive range of target resolutions than ground cameras. Changes in UAV

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024, LNCS 14814, pp. 408–417, 2024.
https://doi.org/10.1007/978-3-031-64608-9_27



Fig. 1. UAV-Human action samples.

motion can easily cause image blurring. Moreover, UAV data are typically collected from a high altitude, so images have varying resolutions and are of low quality, as shown in Fig. 1. Pose estimation from UAV images is prone to errors, which makes skeleton-based HAR a challenging issue.

This paper makes two contributions.

- First, the performance of graph-based learning models for HAR is evaluated on the large-scale UAV-Human dataset.
- Second, a late-fusion framework is proposed to benefit from the advantages of these three graph-based learning models to improve the performance of HAR on UAV-Human.

The remaining sections of the document are arranged as follows: While Sect. 2 reviews deep learning models for HAR, Sect. 3 describes the proposed method. The dataset and experimental findings are thoroughly explained in Sect. 4. Section 5 concludes the paper.

2 Related Work

Deep learning methods for skeleton-based HAR can be categorized into three types: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Graph Convolutional Network (GCN). GCN has emerged as the mainstream for skeleton-based HAR recently. This paper will mainly focus on GCN models, as these are the building blocks of the proposed method.

Graph-based deep learning for HAR is first introduced in ST-GCN [18]. The network models skeletal dynamics using a graph. Every node in the graph represents a joint. Temporal edges link the same joints over successive frames, whereas

spatial edges reflect the skeleton's structure. In [10], a Directed Graph Neural Network (DGNN) is presented by modeling the joint relationship using networks with directed connections. Double-Feature Double-Motion network (DD-Net) is proposed in [20]. The DD-Net's feature shape and graph structure are both reduced. A straightforward but efficient semantics-guided neural network (SGN) is proposed in [22]. The authors explicitly include the high-level semantics to improve the graph's capacity to represent features. There is a hierarchical link between joints using two modules: one at the joint level, which models the correlations between joints within the same frame, and another at the frame level, which models the dependencies between frames by considering all the joints. HARD-Net is proposed in [5] to target early HAR to forecast the action label before the action is completed. Shift-GCN, a lightweight GCN model, is presented in [1] using the shift convolution operator for the 2D convolution procedure. In [13], a multi-stream spatial-temporal relative transformer (MSST-RT) is proposed. The transformer breaks the inherent graph topology in spatial and temporal dimensions.

In [16], TE-GCN is proposed. A unique temporal-enhanced graph convolution network is described to consider temporal aspects. To prevent future information leakage and maintain the ordering information of the inputs, a Causal Convolution layer is added at each time step. In [17], Interacted Object SGN (IO-SGN) is proposed to support the accurate localization of interacted objects, aiding in action identification. The authors investigate the initial action cues and the characteristics of interacted objects. In [15], a lightweight channel-topology-based AGCN (LC-AGCN) is suggested as a solution. A multi-scale spatial-temporal CNN (MSSTNet) is proposed in [2] to capture high-level features for HAR. Rather than employing sophisticated graph convolution, the authors fully utilize multi-scale convolutions to capture the relationships of joints. In [21], combinatorial attention-enhanced adjacency matrix-based GCN (CA-EAMGCN) is proposed. First, an upgraded adjacency matrix is built to increase the model's perceptual field of features. Second, a feature fusion module is added to offer the best fusion ratio possible for the model's various input characteristics.

Since the skeleton model is a graph, representing skeleton data using nodes and edges is efficient and natural. The literature has proposed a variety of graph-based architectures. Among many GCN models for HAR, three models are selected for UAV-Human: AGCN, AAGCN, and MS-G3D. These are state-of-the-art methods for HAR on ground-camera datasets like Kinetics-Skeleton [4] and NTU RGB+D [9].

Three skeleton-based deep learning models, AGCN, AAGCN, and MS-G3D, for HAR on ground-camera data are presented in this section. Using data from ground cameras, these models represent HAR's state-of-the-art deep learning models. While these models perform well on ground-camera data, their accuracy on UAV data can be low. Human shapes only occupy a small portion of scenes. As skeleton data collected by UAVs suffers from extreme noise and varying resolution owing to UAV motion, HAR on UAV data is a complex issue.

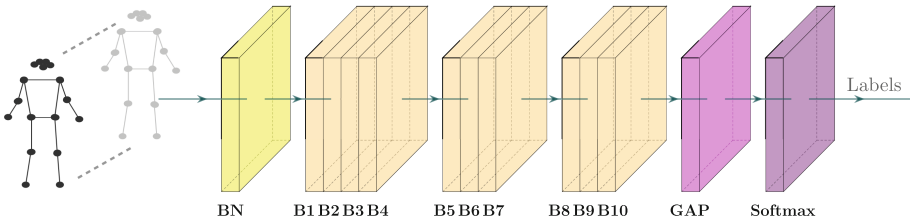

Fig. 2. Block diagram of AGCN and AAGCN. AAGCN differs from AGCN by the attention mechanism in the basic blocks.

2.1 Adaptive GCN (AGCN)

The framework involves cascading layers of graph convolution to analyze data in two dimensions: space and time. Adaptive GCN (AGCN) is introduced in [11]. Unlike ST-GCN, which uses a fixed adjacency matrix, AGCN adaptively learns the graph's topology. The model's generality is increased by this data-driven approach, which makes it more flexible and outperforms ST-GCN on large-scale datasets. AGCN is introduced in [11]. It is mentioned that only some samples of every action class will benefit equally from a single fixed graph topology. The model parameterizes two different graph types, whose structures are trained and updated with the model's convolutional parameters. A global graph is one kind that shows the overall pattern in all the data. An individual graph is another kind that shows a distinct pattern for every data sample. Because each of the two graph types is specifically tuned for a distinct layer, the hierarchical structure of the model can be better fitted. This data-driven approach gives the model more generality to adapt to different data samples and boosts its flexibility for graph creation. As seen in Fig. 2, the stack of fundamental blocks represents the AGCN system diagram. At the input, a Batch Normalization (BN) layer is applied. Ten fundamental blocks with varying numbers of output channels are present. There are 64 channels in the first four blocks. There are 128 channels in the next three blocks and 256 in the final three. The basic blocks are numbered from B1 to B10. A residual link is employed for every basic block to prevent gradient vanishing. The block stack's output is sent to a layer known as Global Average Pooling (GAP). A softmax classifier is utilized to forecast the action label.

2.2 Attention-Enhanced Adaptive GCN (AAGCN)

AAGCN is proposed in [12] to allow the model to pay attention to important data. The block diagram of AAGCN is the same as that of AGCN in Fig. 2. AAGCN differs from AGCN in that each basic block contains a Spatial-Temporal-Channel (STC) attention module, which can aid in teaching the model to focus on discriminative joints, frames, and channels selectively.

2.3 Multi-scale Graph 3D (MS-G3D)

MS-G3D is introduced in [7]. The system diagram of MS-G3D is shown in Fig. 3. The system is the concatenation of three Spatial-Temporal Graph Convolutional (STGC) blocks, a GAP, and a Fully Connected (FC)/Softmax layer. The original MS-G3D paper provides information on STGC blocks. To mitigate the biased weighting, a proposed multi-scale aggregation strategy involves cutting down on extra connections between neighborhood disentangling features. Consequently, more potent multi-scale operators may represent cooperative interactions at any distance.

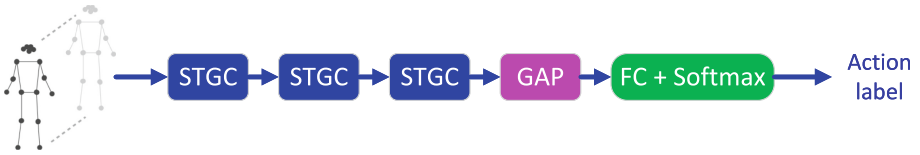

Fig. 3. Block diagram of MS-G3D.

3 Proposed Method

Each graph-based deep learning model has advantages in inference for certain action categories. Every graph-based model makes use of a distinct skeleton data feature. Bone data are defined as the spatial offsets between adjacent joint pairs. Bone data are used in this work since they contain offset values between adjacent joint pairs. Figure 4 shows a schematic diagram of our proposed method. The framework comprises three streams that use bone data as the input to AGCN, AAGCN, and MS-G3D. Softmax scores of three streams are summed up for late fusion to calculate the final score.

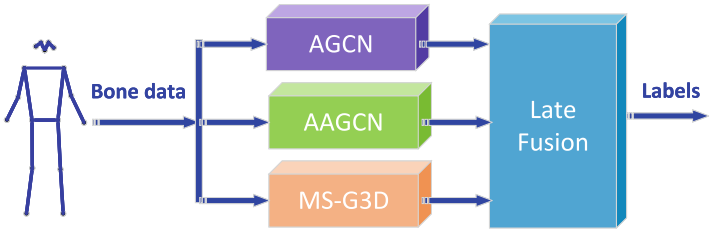

Fig. 4. System diagram of the proposed method.

– C: the number of channels, $C = 2$ for UAV-Human.
– T: the maximum frame length, $T = 263$ for UAV-Human.

– N: the number of joint in the skeleton model, $N = 17$ for UAV-Human.
– M: the maximum number of subjects in each frame, $M = 2$ for UAV-Human

Let $S^i_{\dots}(v)$ be the score that each stream returned for sample v, and let v be the action sample. The following is the expected class of v:

$$i^* = \arg\max_i((S^i_{AGCNbone}(v) + S^i_{AAGCNbone}(v) + S^i_{MSG3Dbone}(v)) \tag{1}$$

where argmax$(\cdot)$ returns the index of the class with the largest value.

4 Experimental Results

4.1 Dataset

The UAV-Human, a large-scale benchmark dataset, is introduced in [6]. The dataset includes skeletal, color, infrared, and depth data, among other data modalities. The dataset poses significant challenges for action detection for the following reasons: It was collected from 45 distinct locations, encompassing a wide range of indoor and outdoor environments, including forests, squares, streets, gyms, university campuses, farms, and many scenarios within buildings. It was gathered over three months by a drone flying in both rural and urban areas, days and nights. As a result, it includes a wide range of subjects, lighting, weather, and camera motion.

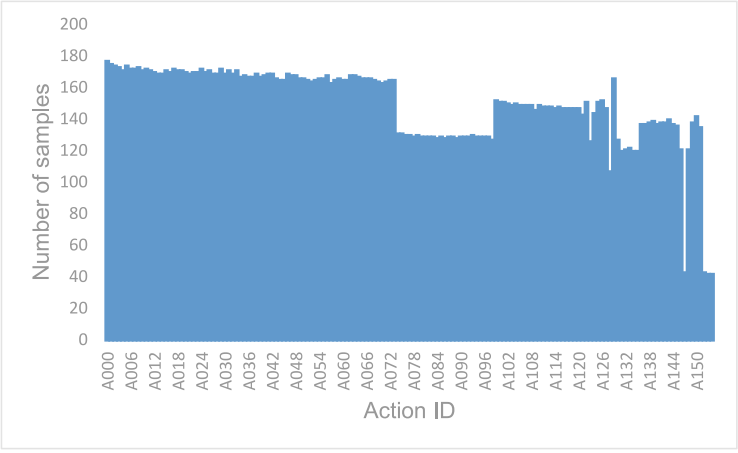

Fig. 5. Number of samples per action category.

Figure 6 depicts the UAV-Human skeletal structure. The construction of the skeleton has 17 joints, also known as joints. UAV-Human is a challenging dataset since it has 155 classes with action IDs ranging from A000 to A154. The list of 155 action classes can be found at [14]. Figure 5 shows the number of samples per action category.

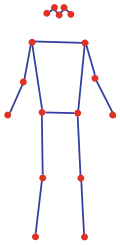

Fig. 6. The skeleton model of the UAV-Human dataset.

4.2 Results and Discussions

All of the GCN schemes are considered to benefit from late fusion regarding HAR performance. With late fusion applied to all three streams, our proposed method achieves up to 43.90% accuracy on UAV-Human. Table 1 shows performance comparison with state-of-the-art methods. The proposed method outperforms the state-of-the-art methods MSSTNet and CA-EAMGCN with accuracies of 43.00% and 43.11%, respectively.

The top 10 action classes with the best accuracy are shown in Table 2. The list of 10 challenging action classes for the proposed method is shown in Table 3. This is an ordinary situation since having a model that outperforms other models in all action classes is difficult, especially for a large-scale dataset like UAV-Human with 155 action classes. The small number of samples for training partially causes low accuracies on action IDs A152, A153, and A154. Action *mow* is mainly confused with *dig a hole* and *set on fire*; action *cough* is confused with *blow nose* as these actions are of high similarity even for human eyes.

Table 1. Comparison with state-of-the-art methods by accuracy (%).

No.	Method	Year	Accuracy [%]
1	ST-GCN [18]	2018	30.25
2	DGNN [10]	2019	29.90
3	DD-Net [20]	2019	31.47
4	2 s-AGCN [11]	2019	34.84
5	SGN [22]	2020	36.37
6	HARD-Net [5]	2020	36.97
7	Shift-GCN [1]	2020	37.98
8	AAGCN [12]	2020	34.45
9	MS-G3D [7]	2020	40.51
10	MSST-RT [13]	2021	41.22
11	TE-GCN [16]	2022	42.50
12	IO-SGN [17]	2022	39.99
13	LC-AGCN [15]	2023	40.49
14	MSSTNet [2]	2023	43.00
15	CA-EAMGCN [21]	2023	43.11
16	ViA [19]	2024	42.60
17	**Proposed**	2024	**43.90**

Table 2. Top 10 action classes with the best accuracies using the proposed method.

Action ID	Action	# Training	# Testing	Accuracy [%]
A092	move something with someone	93	35	82.86
A007	stand up	126	45	75.56
A134	comfort someone	85	34	73.53
A011	put on a coat	125	45	64.44
A023	jump on two legs	125	44	63.64
A112	rear right turn	106	41	63.41
A090	high five	93	35	62.86
A097	lend an arm to support someone	93	35	62.86
A132	wave a goodbye	86	34	61.76
A101	land at designated locations	109	41	60.98

Table 3. Top 10 action classes with the lowest accuracies using the proposed method.

Action ID	Action	# Training	# Testing	Accuracy [%]
A123	mow	87	38	21.05
A062	cough	122	45	20.00
A017	throw away a hat	127	44	18.18
A152	walk	20	22	18.18
A041	make a victory sign	123	45	17.78
A086	hit someone with something	92	35	17.14
A040	make an OK sign	122	45	15.56
A125	smoke	109	41	14.63
A154	close an umbrella	20	21	14.29
A153	open an umbrella	20	21	9.52

5 Conclusions

UAVs' automatic HAR is crucial for surveillance and rescue operations. This paper assessed the performance of various graph-based deep learning models on the large-scale UAV-Human dataset. Bone data are used in this work since they contain offset values between adjacent joint pairs. A late fusion framework is proposed to benefit the advantages of AGCN, AAGCN, and MS-G3D using UAV bone data. According to evaluation results, the proposed method outperforms existing methods on UAV-Human. Future studies should focus on improving pose estimation from UAV data, which will help improve HAR results.

Acknowledgments. Supported by International School, Vietnam National University, Hanoi, Vietnam.

References

1. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
2. Cheng, Q., Cheng, J., Ren, Z., Zhang, Q., Liu, J.: Multi-scale spatial-temporal convolutional neural network for skeleton-based action recognition. Pattern Analysis and Applications, pp. 1–13 (2023)
3. Degardin, B., Proença, H.: Human behavior analysis: a survey on action recognition. Appl. Sci. **11**(18), 8324 (2021)
4. Kay, W., et al.: The Kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
5. Li, T., Liu, J., Zhang, W., Duan, L.: HARD-Net: hardness-aware discrimination network for 3D early activity prediction. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision - ECCV 2020, pp. 420–436. Springer International Publishing, Cham (2020)

6. Li, T., Liu, J., Zhang, W., Ni, Y., Wang, W., Li, Z.: UAV-Human: a large benchmark for human behavior understanding with unmanned aerial vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021)
7. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 143–152 (2020)
8. Pham, D.T., et al.: Deep learning models for skeleton-based action recognition for uavs. In: 2022 IEEE Ninth International Conference on Communications and Electronics (ICCE), pp. 343–348. IEEE (2022)
9. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: a large scale dataset for 3D human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)
10. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with directed graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7912–7921 (2019)
11. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12026–12035 (2019)
12. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. IEEE Transactions on Image Processing, pp. 9532–9545 (2020)
13. Sun, Y., Shen, Y., Ma, L.: Msst-rt: multi-stream spatial-temporal relative transformer for skeleton-based action recognition. Sensors **21**(16), 5339 (2021)
14. SUTDCV: UAV-Human. https://github.com/SUTDCV/UAV-Human (2021). Accessed 02 May 2024
15. Wang, K., Deng, H., Zhu, Q.: Lightweight channel-topology based adaptive graph convolutional network for skeleton-based action recognition. Neurocomputing **560**, 126830 (2023)
16. Xie, Y., Zhang, Y., Ren, F.: Temporal-enhanced graph convolution network for skeleton-based action recognition. IET Comput. Vision **16**(3), 266–279 (2022)
17. Xu, L., Lan, C., Zeng, W., Lu, C.: Skeleton-based mutually assisted interacted object localization and human action recognition. IEEE Transactions on Multimedia (2022)
18. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI Conference on Artificial Intelligence (2018)
19. Yang, D., Wang, Y., Dantcheva, A., Garattoni, L., Francesca, G., Brémond, F.: View-invariant skeleton action representation learning via motion retargeting. Int. J. Comput. Vision pp. 1–16 (2024)
20. Yang, F., Wu, Y., Sakti, S., Nakamura, S.: Make skeleton-based action recognition model smaller, faster and better. In: Proceedings of the ACM multimedia Asia, pp. 1–6. ACM (2019)
21. Zhang, D., Deng, H., Zhi, Y.: Enhanced adjacency matrix-based lightweight graph convolution network for action recognition. Sensors **23**(14), 6397 (2023)
22. Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N.: Semantics-guided neural networks for efficient skeleton-based human action recognition. In: proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1112–1121 (2020)