Deep Learning Techniques for Skeleton-Based Action Recognition: A Survey

Dinh-Tan Pham(✉)

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

Abstract. Interpreting human behavior from entirely performed actions is called human action recognition (HAR). HAR applications rapidly expand into robotics, CCTV surveillance, self-driving vehicles, gaming, and video retrieval. Among different data modalities, skeleton data offers compact representation and computational efficiency. In recent years, much work has gone into developing a robust and accurate deep-learning framework for skeleton-based HAR. The paper reviews state-of-the-art methods for skeleton-based HAR. The survey also summarizes evaluation results on a large-scale benchmark dataset. Trends in action recognition research are discussed.

Keywords: action recognition · skeleton data · deep learning

1 Introduction

In computer vision, human action refers to movements, from single-limb movements to complicated joint movements involving several limbs and the whole body. Human action recognition (HAR) is a fundamental job that identifies human activities by observing the entire execution of the action. The issue has been extensively researched for many years and remains challenging because of its numerous practical uses, such as visual surveillance and video retrieval. Much work has gone into developing an intelligent system that can recognize complicated human activities in congested situations, like humans. But an action is merely presented to a computer by a collection of pixels. The computer cannot interpret these pixels into a useful representation or deduce human behavior from that representation [7].

Psychological research shows that skeleton data are discriminative to present human actions [3]. A skeleton model with 25 joints is shown in Fig. 1. Furthermore, skeletons are invariant to changes in appearance and backdrop, facilitating the description of actions. The skeletal data for action representation contains interdependence between joints. Pixels in depth/color images do not directly provide these dependencies. Consequently, there has been a lot of research in skeleton-based HAR in recent years [19]. Noise and changes in illumination significantly impact action recognition in RGB video sets. Action recognition based

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024, LNCS 14814, pp. 427–435, 2024.
https://doi.org/10.1007/978-3-031-64608-9_29

on skeletons is resistant to these issues. Skeletal data can also enable privacy protection in the real world [32].

Motion capture (Mocap) system can be used to collect skeleton data. Since mocap devices are not affected by illumination or vision, they can provide high-quality skeleton data. Mocap systems are pricey and cumbersome, though. Therefore, skeletal data from color photographs or depth maps is used in various scenarios. Skeletal data are now readily available due to the widespread use of inexpensive depth sensors and advancements in pose estimation. Pose estimation can be used to color/depth data to provide skeleton information or obtained directly via depth sensors. Joint trajectories can be used to represent human action. Skeletal data HAR methods can prevent color data weaknesses such as background changes and lighting conditions. Additionally, skeleton data offers computation and storage efficiency.

Nevertheless, skeleton-based HAR is not the main focus of the surveys mentioned above. They draw attention to various taxonomies, uses, and theoretical stances of HAR. This serves as the driving force of this paper, providing a more thorough analysis of skeleton-based HAR. The remaining of the paper is organized as follows: Sect. 2 is the survey of skeleton-based HAR techniques. Section 3 is a comparative study on the performance of deep learning models using the benchmark dataset NTU RGB+D. The paper is concluded in Sect. 4 with a discussion of some future approaches.

2 Skeleton-Based HAR Techniques

Skeleton-based HAR techniques can be categorized as in Fig. 2. In recent years, deep learning-based HAR become dominant with remarkable performance. popular deep learning techniques are Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Graph Convolution Networks (GCN).

2.1 Convolutional Neural Network (CNN)

The CNN-based approach retrieves the features from the texture color images created by first encoding the skeleton data. These approaches still consider skeleton data as picture characteristics. The skeleton-based action recognition of CNN-based techniques can also benefit from the experience obtained from the earlier CNN application. The ability of CNN to recognize patterns in images has been demonstrated. Nevertheless, it must still model temporal dependency for video action recognition. The skeleton sequence is represented as an image rather than processing individual frames. Both geographical and temporal information are considered in the resulting image. Deep CNNs can successfully learn the skeleton sequence's spatial and long-term temporal structure from the generated image. A normalization procedure produces three gray images for every skeleton sequence, corresponding to the sequence's 3D coordinates. CNN-based methods are shown in Table 2.

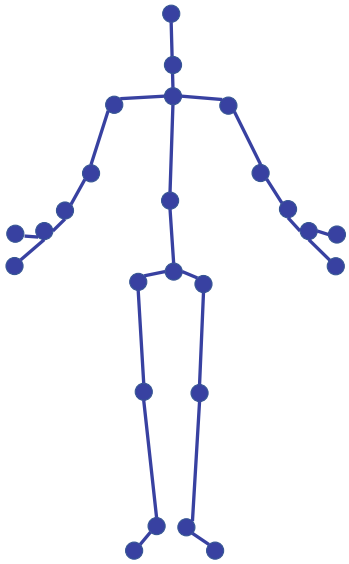

Fig. 1. Skeleton model with 25 joints.

2.2 Recurrent Neural Network (RNN)

RNN was first widely employed in the Natural Language Processing discipline. When processing sequential data, the RNN offers a significant benefit. Later, it was progressively used in many sequential data-related applications, including machine translation, sentiment analysis, picture captioning, time-series prediction, etc. RNNs may accommodate sequences of any length in theory. An excessively lengthy sequence during training can cause issues with gradient explosion and vanishing gradient during optimization. However, a maximum length is typically provided in practice because the extended feedforward neural network will consume too much memory. The series is terminated when it gets longer than the allowed length. RNN's shortcoming is addressed by Long Short-Term Memory (LSTM) networks. LSTM can handle the trajectories of skeleton joints as a time series and investigate the temporal structure of the skeleton sequences for action recognition. LSTMs help RNNs overcome recurrent issues like disappearing gradients by enabling training on lengthy sequences. Even though LSTM networks are meant to investigate long-term dependencies, they need help to learn the contents of a sequence with numerous timestamps. Furthermore, LSTM networks have trouble extracting high-level characteristics. Typical RNN-based methods are summarized in Table 2.

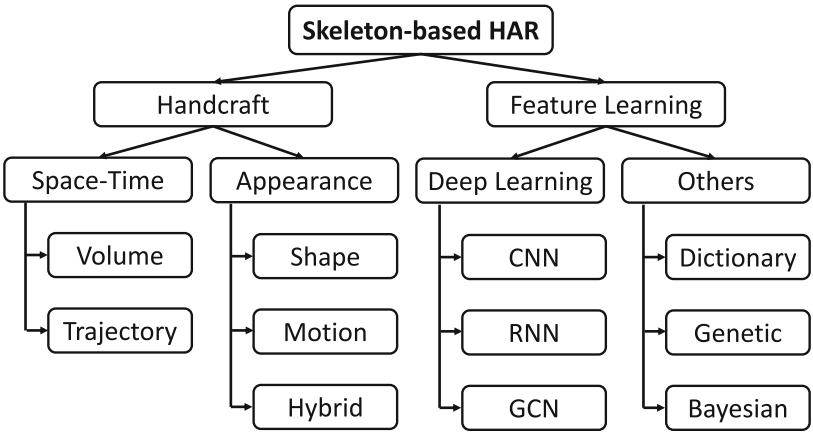

Fig. 2. Skeleton-based HAR techniques.

2.3 Graph Convolution Networks (GCN)

In contrast to the conventional CNN and RNN network model, GCN may handle data with a generalized topology structure and thoroughly mine its properties. This has much in common with the convolution kernel's weighted average and parameter-sharing mechanisms. Skeleton architecture data are transformed into standard grid structure data before applying CNN structure to the data. This significantly reduces CNN's effectiveness in skeleton-based HAR.

GCN extends the use of the convolution operator for images to the graph domain. Yan *et al.* first proposed GCN for HAR in Spatial-Temporal GCN (ST-GCN) [31]. The skeletal data are represented as graphs, with joints as graph nodes. The spatial graph is constructed using the innate joint connections seen in the human body. The temporal graph is made up of connections between joints in consecutive frames. Tang *et al.* provide a system in [28] that consists of two networks: the Graph-based CNN (GCNN) and the frame distillation network (FDNet). The FDNet aims to extract critical frames from input sequences. GCNN considers the human body a graph structure and uses this information for action recognition. The framework can identify the most illuminating frames in every sequence. A progressive approach is used to model how frame selection works. Action detection is performed using the chosen frames. In [11], the Actional-Structural GCN (AS-GCN) concept is presented, utilizing an inference block. The model creates a graph by combining structural and actional links. The structural linkages represent higher-order dependencies, and the actional links embed action-specific dependencies. In [1], Channel-wise Topology Refinement GCN (CRT-GCN) is proposed. A parameterized adjacency matrix known as the shared topology provides generic correlations between vertices and topological priors for all channels. The channel-specific correlations capture the nuanced interactions between vertices inside each channel, which are dynamically inferred for every sample. The multi-stream EfficientGCN-B4 model is proposed by Song

et al. [26]. This model uses early fusion to merge the inputs, including joint locations and motion velocities. In [30], the Language Knowledge-Assisted GCN (LA-GCN) is proposed. The model directs the creation of new representations to highlight important node information at the data level and to replicate category prior knowledge in human brain.

3 Performance Evaluation

Table 1. Benchmark datasets with data modalities: Skeleton (S), Depth (D).

No.	Dataset	Modality	#Class	#Subject	#Sample
1	MSR-Action3D [15]	D,S	20	10	557
2	HMDB51 [8]	RGB	51	–	6,766
3	UCF101 [27]	RGB	101	–	13,320
4	UTKinect [29]	RGB,D,S	10	10	200
5	MSRAction-Pair [18]	RGB,D,S	12	10	360
6	NTU RGB+D [20]	RGB,D,S	60	40	56,880
7	Kinetics-400 [4]	RGB	400	–	306,245
8	UAV-Human [14]	RGB,D,S	155	119	67,428

There are many skeleton datasets in literature built for developing and evaluating HAR techniques as in Table 1. In this paper, the most popular large-scale dataset, NTU RGB+D, is selected as the benchmark dataset for evaluating deep learning methods. The NTU RGB+D dataset is introduced in [20], utilizing data modalities gathered from Kinect v2 sensors. The skeleton model has twenty-five joints as in Fig. 1. The dataset contains 56,880 sequences that are divided into 60 activity classes. There are forty people involved in the actions. The three Kinect sensors are positioned at different angles but at the same height. Seventeen camera sets with varying heights and distances are used to gather the dataset. The authors of the dataset suggest two benchmark protocols: (1) Cross-subject: half of the participants are trained, while the other half are tested. The training set consists of 40,320 sequences, whereas the testing set contains 16,560 sequences. (2) Cross-view: 18,960 sequences (from camera one) make up the testing set, while 37,920 sequences (from cameras two and three) make up the training set. A comparative study on the performance of deep learning techniques on NTU RGB+D is shown in Table 2. It can be seen that state-of-the-art methods for HAR are graph-based deep learning models. This comes from the graph nature of skeleton data in human action presentation. Action recognition in this case is a graph classification task.

Table 2. Performance evaluation by accuracy (%) on NTU RGB+D.

No.	Method	Year	Cross-subject	Cross-view
1	Bi-directional RNN [2]	2015	59.1	64.0
2	Part-based LSTM [20]	2016	60.7	67.3
3	ST-LSTM [16]	2016	69.2	77.7
4	STA-LSTM [24]	2017	73.4	81.2
5	VA-LSTM [33]	2017	79.2	87.7
6	Res-TCN [6]	2017	74.3	83.1
7	Clip CNN [5]	2017	79.6	84.8
8	Synthesized CNN [17]	2017	80.0	87.2
9	Motion CNN [10]	2017	83.2	89.3
10	Multi-scale CNN [9]	2017	85.0	92.3
11	ST-GCN [31]	2018	81.5	88.3
12	GCNN [28]	2018	83.5	89.8
13	IndRNN [13]	2018	81.8	88.0
14	SRN+TSL [23]	2018	84.8	92.4
15	ARRN-LSTM [34]	2019	80.7	88.8
16	Dense IndRNN [12]	2019	86.7	94.0
17	AS-GCN [11]	2019	86.8	94.2
18	AGCN [21]	2019	87.3	93.7
19	3 s RA-GCN [25]	2020	87.3	93.6
20	AAGCN [22]	2020	88.0	95.1
21	CTR-GCN [1]	2021	92.4	96.8
22	EfficientGCN-B4 [26]	2022	92.1	96.1
23	LA-GCN [30]	2023	93.5	97.2

4 Conclusion

Understanding human actions is essential in computer vision, robotics, security surveillance, and home monitoring applications. This paper provides a summary of the most recent research in this area. It suggests a classification based on deep learning techniques for skeleton-based action recognition. The analysis of the state-of-the-art methodologies and the methods employed for their validation are presented. Action recognition technology applied to real-world scenes may encounter several practical issues, including occlusion, non-unique characters and activities in the video, low picture resolution, and challenging access to vast numbers of labeled data sets. Due to these issues, many action recognition systems are still in the lab. The performance of HAR can be improved using multi-modal HAR or combining it with action-related object information.

Acknowledgments. This research is funded by International School, Vietnam National University, Hanoi, Vietnam.

References

1. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13359–13368 (2021)
2. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1110–1118 (2015)
3. Johansson, G.: Visual perception of biological motion and a model for its analysis. Percept. Psychophys. **14**(2), 201–211 (1973)
4. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The Kinetics human action video dataset, pp. 1–22. arXiv preprint arXiv:1705.06950 (2017)
5. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3D action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3288–3297 (2017)
6. Kim, T.S., Reiter, A.: Interpretable 3D human action analysis with temporal convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1623–1631. IEEE (2017)
7. Kong, Y., Fu, Y.: Human action recognition and prediction: a survey. Int. J. Comput. Vision **130**(5), 1366–1401 (2022)
8. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: 2011 International Conference on Computer Vision, pp. 2556–2563. IEEE (2011)
9. Li, B., Dai, Y., Cheng, X., Chen, H., Lin, Y., He, M.: Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN. In: 2017 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 601–604. IEEE (2017)
10. Li, C., Zhong, Q., Xie, D., Pu, S.: Skeleton-based action recognition with convolutional neural networks. In: 2017 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 597–600. IEEE (2017)
11. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3595–3603 (2019)
12. Li, S., Li, W., Cook, C., Gao, Y.: Deep independently recurrent neural network (IndRNN), pp. 1–18. arXiv preprint arXiv:1910.06251 (2019)
13. Li, S., Li, W., Cook, C., Zhu, C., Gao, Y.: Independently recurrent neural network (IndRNN): Building a longer and deeper RNN. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5457–5466 (2018)
14. 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/CVF Conference on Computer Vision and Pattern Recognition, pp. 16266–16275 (2021)

15. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 9–14. IEEE (2010)
16. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal LSTM with trust gates for 3D human action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 816–833. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_50
17. Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. Pattern Recogn. **68**, 346–362 (2017)
18. Oreifej, O., Liu, Z.: HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 716–723 (2013)
19. Ren, B., Liu, M., Ding, R., Liu, H.: A survey on 3D skeleton-based action recognition using learning method, pp. 1–8. arXiv preprint arXiv:2002.05907 (2020)
20. 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)
21. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12026–12035 (2019)
22. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. IEEE Trans. Image Process. **32** 9532–9545 (2020)
23. Si, C., Jing, Y., Wang, W., Wang, L., Tan, T.: Skeleton-based action recognition with spatial reasoning and temporal stack learning. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11205, pp. 106–121. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_7
24. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end Spatio-temporal attention model for human action recognition from skeleton data. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, pp. 1–7 (2017)
25. Song, Y.F., Zhang, Z., Shan, C., Wang, L.: Richly activated graph convolutional network for robust skeleton-based action recognition. IEEE Trans. Circuits Syst. Video Technol. **31**(5), 1915–1925 (2020)
26. Song, Y.F., Zhang, Z., Shan, C., Wang, L.: Constructing stronger and faster baselines for skeleton-based action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **45**, 1474–1488 (2022)
27. Soomro, K., Zamir, A.R., Shah, M.: UCF101: a dataset of 101 human actions classes from videos in the wild, pp. 1–6. arXiv preprint arXiv:1212.0402 (2012)
28. Tang, Y., Tian, Y., Lu, J., Li, P., Zhou, J.: Deep progressive reinforcement learning for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5323–5332 (2018)
29. Xia, L., Chen, C.C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3D joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 20–27. IEEE (2012)
30. Xu, H., Gao, Y., Hui, Z., Li, J., Gao, X.: Language knowledge-assisted representation learning for skeleton-based action recognition. arXiv preprint arXiv:2305.12398 (2023)
31. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI Conference on Artificial Intelligence, pp. 1–20 (2018)

32. Yue, R., Tian, Z., Du, S.: Action recognition based on RGB and skeleton data sets: a survey. Neurocomputing **512**, 287–306 (2022)
33. Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2117–2126 (2017)
34. Zheng, W., Li, L., Zhang, Z., Huang, Y., Wang, L.: Relational network for skeleton-based action recognition. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 826–831. IEEE (2019)