Research Progress of Skeleton-Based Action Recognition Technologies

Zexiang Chen*

College of Computer Science and Technology, Guizhou University, 550025, Guiyang, Guizhou, China

Abstract. Skeletal-based action recognition technology, which analyzes the spatiotemporal sequences of human skeletal joints to identify human behaviors, has garnered widespread attention in computer vision in recent years. This review aims to collate and summarize the research advancements in this domain, with a particular focus on the classification and comparison of feature extraction methodologies. The paper commences by elucidating the acquisition and preprocessing of skeletal data, laying the groundwork for subsequent feature extraction. The thematic focus of the research centers on two predominant feature extraction approaches: those based on customary handcrafted features and those predicated on deep learning methodologies. The methodology encompasses a systematic literature review and comparative analysis, complemented by an introduction to the principal benchmark datasets. The paper juxtaposes the strengths and limitations of various feature extraction techniques through these methodologies and explores their potential in practical applications. In conclusion, this review's importance comes from its thorough analysis of the topic of skeletal-based action recognition. In addition to giving a well-organized summary of the status of the field, it also sheds light on how effective various feature extraction methods are.

1 Introduction

Skeletal action recognition, a critical research domain within the realms of computer vision and machine learning, has garnered substantial interest over recent years. Dedicated to analyzing and identifying a spectrum of human actions from sequential data derived from skeletal joints, this technology finds applications across human-computer interaction, intelligent surveillance, virtual reality, and healthcare. Skeletal action recognition has advanced significantly due to the quick growth of deep learning, particularly in managing complex scenes and enhancing recognition precision. Ren et al., in their comprehensive review paper, have for the first time thoroughly spoken about the study on action recognition using 3D skeletal data and deep learning, covering various aspects from theory to practice, providing valuable resources and insights for researchers in this field [1]. Compared to traditional RGB video or depth map sequences, skeletal data for human action recognition offers numerous benefits: it is characterized by high abstraction, low complexity, and

* Corresponding author: 101010662@seu.edu.cn

robustness, and is less susceptible to interference from background, scale, viewpoint, lighting, and other factors. Skeletal data aligns more closely with the actual physical meaning of human actions, providing a superior representation of the human motion process.

This paper categorizes skeletal action recognition methods based on feature extraction techniques, distinguishing between traditional handcrafted feature extraction and deep learning-based approaches. Traditional handcrafted feature extraction methods are primarily discussed in terms of geometric relationships, kinematics, and statistical analysis. These methods rely on expert knowledge to design features that capture intuitive and meaningful attributes of actions but may lack flexibility and efficiency in handling large-scale data and complex motion patterns. In contrast, deep learning-based methods leverage algorithms to automatically learn features that differentiate between various actions, encompassing models based on Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Graph Convolutional Networks (GCN), and Transformers.

The significance of this classification lies in its ability to systematically compare and contrast traditional and modern approaches to feature extraction in skeletal action recognition. Such systematic categorization facilitates a more nuanced understanding of the strengths and weaknesses of each method, which is essential for advancing the field. While traditional handcrafted feature extraction methods offer the advantages of intuitiveness and interpretability, they may be limited in their generalization capabilities and handling of complex data. Deep learning-based methods excel in automated feature extraction and the processing of complexity but require substantial annotated data and computational resources, presenting challenges in model interpretability. This paper introduces several typical public datasets in this domain, which serve as benchmarks for assessing the effectiveness of skeletal action recognition algorithms. These datasets provide a common ground for researchers to compare their methods and results, which is crucial for driving innovation and enhancing the reliability of relevant systems. In assessing model performance, standard metrics are commonly employed, which are benchmarks for evaluating classification models in machine learning. These metrics allow researchers to quantify model performance and make optimizations and improvements.

2 Data preparation

2.1 Basic step

Data preprocessing is a critical step in skeletal action recognition research, ensuring the quality and usability of the data. Utilizing depth cameras such as the Microsoft Kinect v2 or Intel RealSense to capture human movements from multiple perspectives, 3D skeletal data is generated. The collected skeletal data is annotated to provide action category labels for subsequent analysis. When employing multiple sensors, data synchronization is necessary to ensure temporal consistency. Noise in the skeletal data, such as outliers and incoherent joint movements, must be removed. Normalization of skeletal data is conducted, including scale normalization and rotation normalization, to eliminate individual differences. Smoothing of skeletal data is performed to reduce jitter. Techniques for augmenting data, including temporal warping and data synthesis, are used to increase data diversity. Key features of skeletal data are extracted, including joint positions and joint angles.

2.2 Outlier detection

In the domain of skeletal action recognition, the treatment of outliers is a crucial phase in the preprocessing of data. Outliers, defined as information that substantially departs from the

norm, have the potential to adversely affect the effectiveness of algorithms for action recognition. Consequently, the identification and handling of these outliers are of paramount importance. The process of outlier management typically involves the detection and elimination of anomalies within skeletal data, thereby reducing noise and errors and improving the dataset's quality and consistency. By cleansing skeletal data of outliers, researchers can guarantee the precision and dependability of further analysis. Moreover, outlier correction and smoothing techniques are widely applied to mitigate the impact of outliers on the overall data [2]. A variety of methods can be employed to address outliers, including statistical and machine-learning approaches. For instance, the Z-score method and the Interquartile Range (IQR) method are statistical anomaly detection techniques that assess whether data points are outliers based on their distance from the dataset's mean or quartiles [3, 4]. Additionally, model-based detection methods, like decision trees and Support Vector Machines (SVM), can be utilized to identify outliers.

In practical applications, managing outliers involves setting thresholds, identifying them, and deciding on removal or replacement, with caution to avoid bias or error. Outliers should be verified as non-meaningful before deletion, as they can be genuine extreme cases. Addressing outliers is vital for data quality in skeletal action recognition research, influencing the accuracy and robustness of recognition algorithms. Researchers can better preprocess data by cautiously employing various methods to handle outliers.

3 Feature extraction methods

3.1 Manual feature extraction methods

In the early studies of action recognition, researchers primarily relied on manual feature extraction methods due to the limited sample sizes of datasets. These methods typically involved transforming skeletal joint points into descriptors to capture the evolution of human motion in both time and space. These descriptors were capable of directly reflecting the kinematic properties of motion or the statistical distribution characteristics of the data. Subsequently, the extracted features were encoded into feature vectors, which were then subjected to discriminant analysis using classifiers. Based on the type of descriptors employed, These techniques fall into three primary categories: geometric features, kinematic features, and statistical features. Each category reveals the intrinsic characteristics of actions from different perspectives, providing a unique viewpoint for action recognition.

3.1.1 Geometric Feature

Evangelidis et al. developed a local skeletal descriptor for action recognition, encoding joint quadruplets' relative positions to yield a compact, 6D viewpoint-invariant feature known as the skeletal quad, achieved through similarity normalization transformation [5].

3D skeletons were suggested to be represented as points in a Lie Group by Vemulapalli et al., leveraging the mathematical structure to capture spatiotemporal characteristics and provide a discriminative action representation [6]. This approach maps normalized skeletal data to points within the Lie Group, with each point corresponding to a combination of rotation and translation, capturing the skeleton's configuration and motion over time.

3.1.2 Dynamic Feature

In the study by Li et al., a novel method founded on the idea of a "Bag of 3D Points" was proposed [7]. The core idea of this approach is to treat important points of the human skeleton

as points in three-dimensional space and describe and recognize actions through the distribution of these points. This approach is particularly advantageous in handling occlusions. For instance, occlusions can be simulated by ignoring 80 points.

Yang et al. introduced the EigenJoints descriptor, which is designed by combining static posture, motion attributes, and overall dynamic information of the joints [8]. These features, after normalization and Principal Component Analysis (PCA), effectively capture the dynamic variations of actions. The EigenJoints descriptor provides a comprehensive feature set that encompasses the spatial relationships and motion dynamics of the skeletal structure, offering a robust representation for action recognition.

3.1.3 Statistical Feature

A hierarchical approach to human action recognition was presented by Su et al., which analyzes action features at different levels to enhance recognition accuracy [9]. The method extracts motion features from three-view projections to mitigate self-occlusion and employs SVM for coarse classification followed by HMM for fine classification.

Tang et al. introduced a weighted covariance descriptor-based online technique, designed to process continuous data streams and allocate more weight to informative frames like keyframes, diverging from traditional segment-based action recognition methods [10].

3.2 Deep Learning-based

Skeletal-based action recognition has seen a shift in research focus toward deep learning-based techniques. These approaches overcome the limitations of traditional handcrafted feature extraction methods and dramatically improve action recognition's precision and resilience. Deep learning models, particularly CNNs, RNNs, and GCNs, have demonstrated superior performance in handling complex spatiotemporal data. In recent years, researchers have increasingly recognized that deep learning is not only effective in capturing the spatial characteristics of actions but also in modeling the dynamic changes of actions over time. This capability has allowed deep learning methods to excel on datasets such as NTU RGB+D [11] and Kinetics [12], propelling the progress of action recognition technology.

3.2.1 CNN-based

Duan et al. introduced PoseConv3D, a network leveraging 3D heatmap volumes to represent human skeletons, outperforming GCN-based methods in learning spatiotemporal features and showing robustness to noise and better generalization across datasets [13]. PoseConv3D efficiently handles multi-person scenarios and is capable of early fusion integration with other methods. A dual-stream 3D CNN method was formulated by Liu et al for video action recognition, encoding joint coordinates separately along temporal and spatial dimensions for high-dimensional feature extraction, combining spatial and temporal information to enhance accuracy [14]. Li et al. presented the Hierarchical Co-occurrence Network (HCN), a comprehensive framework using CNNs to learn global co-occurrence features in skeleton sequences, with a global spatial aggregation strategy that outperforms local aggregation in learning joint co-occurrence features [15]. The HCN integrates raw skeleton coordinates and motion features through a dual-stream paradigm to improve the model's capacity to represent changes in action.

3.2.2 RNN-based

A technique called Hierarchical Recurrent Neural Networks (HRNN) was developed by Du et al., segmenting the skeleton into five parts for processing in sub-networks, with representations fused layer by layer to form the final decision [16]. Veeriah et al. introduced Differential Recurrent Neural Networks (dRNN) to enhance traditional LSTM networks by focusing on information gain changes due to significant motion patterns between frames, quantified through state derivatives, thus capturing complex action dynamics [17]. Liu et al. invented Spatio-Temporal LSTM (ST-LSTM), extending RNN to the spatial dimension and combining spatiotemporal features, with a tree-structured traversal method inspired by the human skeleton's graphical structure [18].

3.2.3 GCN-based

Spatial-Temporal Graph Convolutional Networks (ST-GCN) were first presented by Yan et al., which automatically pick up temporal and spatial patterns, enhancing expressiveness and generalization compared to previous hand-designed methods [19]. ST-GCN encodes human skeleton dynamics into graph structures, with nodes representing body joints and edges indicating their connectivity, allowing for multi-layer spatial-temporal graph convolutions. Cheng et al. developed the Shift Graph Convolutional Network (Shift-GCN) to address the computational complexity and receptive field limitations of traditional GCNs, utilizing lightweight shift operations for flexible spatial and temporal receptive fields [20]. Li et al. proposed Actional-Structural Graph Convolutional Networks (AS-GCN), utilizing an encoder-decoder architecture to extract latent dependencies unique to a given action and higher-order structural links, combining them into a generalized skeleton graph for learning spatial and temporal features [21].

3.2.4 Transformer-based

Shi et al. introduced the Decoupled Spatial-Temporal Attention Network (DSTA-Net), leveraging attention mechanisms to model joint dependencies without predefined positions or connections [22]. DSTA-Net concentrates on the disentanglement of spatiotemporal attention, the separation of positional encoding, and the implementation of spatial global regularization to enhance the understanding of joint interactions and spatial structures.

Qiu et al. developed the Spatio-Temporal Tuples Transformer (STTFormer) to capture spatiotemporal joint dependencies in skeleton data through "tuples" of segmented sequences and a tuple self-attention module [23]. The method also includes a feature aggregation module for distinguishing similar actions, achieving improved performance on major datasets.

Shi et al. designed the STAR model, representing video frames and skeleton sequences using sparse attention in the spatial dimension and piecewise linear attention in the temporal dimension [24]. The framework integrates comprehensive self-attention along with zigzag and binary spatiotemporal attention mechanisms to effectively capture multifaceted representations of spatial-temporal data.

4 Datasets

4.1 Public Datasets

The NTU RGB+D dataset, introduced by Shahroudy et al. in 2016 [11], consisted of 56,880 video samples that were recorded with the Microsoft Kinect v2 sensor. The skeleton data

consists of 3D positional information from 25 human joint points. The dataset was constructed with 40 subjects performing 60 action categories. Subsequently, the group suggested an expanded dataset, named NTU RGB+D 120 [25]. This extension increased the number of action categories from 60 to 120, amassing a total of 114,480 video samples, and expanded the number of viewpoints from 80 to 155. The extensive action dataset Kinetics-400 served as the basis for the Kinetics-Skeleton dataset [12]. The Kinetics-400 dataset, released in 2017, is compiled from various YouTube videos and contains 306,245 video clips across 400 categories of actions, 400 or more video clips in each category, each approximately 10 seconds long. By extracting human skeleton data from the Kinetics-400 dataset, the dataset was formed. As depicted in Table 1, the classification and characteristics of various skeletal behavior recognition datasets are outlined, providing a comprehensive overview of the sample size, number of joints, action categories, number of subjects, data source, and year of publication for each dataset. This table provides insights into the variety and breadth of datasets available for skeletal action recognition, making it an irreplaceable resource for field researchers and practitioners.

Table 1. Classification and characteristics of skeletal behavior recognition datasets.

Name of the dataset	Sample size	Number of joints	Number of Action Categories	Number of Subjects	Data Source	Year of Database Publication
NTU RGB+D 60	56880	25	60	40	Kinect v2	2016
NTU RGB+D 120	114480	25	120	106	Kinect v2	2020
Kinetics-Skeleton	300000	18	400	400	Youtube	2017
MSR-Action 3D	567	20	20	10	Kinect v1	2010
SBU Kinect Interaction	282	15	8	8	Kinect v1	2012
Florence 3D-Action	215	15	9	10	Kinect v1	2013
Northwestern-UCLA	1494	20	10	10	Kinect v1	2014

4.2 Evaluation Index

Accuracy is frequently used in action recognition as a metric for assessing different approaches, as shown in Equation (1).

$$Accuracy = \frac{N_{corr}}{N_{total}} \tag{1}$$

Where the number of correctly classified samples is denoted by N_corr, and the total number of samples is denoted by N_total. Known as the C-View protocol, Cross-View is the standard used to separate the training and certification sets in the NTU RGB+D 60 dataset. It divides the sets based on cameras, with the test set consisting of 18,960 samples from camera 1 and the training set consisting of 37,920 samples from cameras 2 and 3. Despite having different horizontal angles, the cameras are positioned at the same vertical height, specifically -45°, 0°, and 45°. The C-Set methodology, also known as Cross-Setup, is the benchmark for the NTU RGB+D 120 dataset's training and evaluation sets. It sets up 32 sets based on the height and distance of the cameras from the subjects, training with even-numbered set identifiers and testing with odd-numbered identifiers. The C-Sub protocol holds significant importance in the aforementioned datasets. Persons with the numbers 1, 2,

4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, and 38 are identified by C-Sub as the training set, whereas the test set consists of the remaining individuals.

5 Conclusions

In conclusion, the area of skeleton-based action recognition has witnessed a qualitative leap, with deep learning-based methods emerging as a dominant research direction, offering superior performance and resilience in contrast to conventional manual feature extraction methods. The review has highlighted the importance of feature extraction methodologies, with a detailed examination of both manual and deep learning-based approaches, underscoring their respective strengths and areas for potential improvement. The comprehensive analysis of datasets has provided a clear understanding of the current resources available for research, emphasizing the need for diverse and challenging datasets to further advance the field. The evaluation indices discussed have set a standard for assessing the efficacy of action recognition algorithms, with accuracy and other relevant metrics providing a benchmark for comparing different methods. As the technology continues to evolve, the integration of skeleton-based action recognition with other modalities, such as inertial sensors, presents an exciting avenue for upcoming studies, promising to enhance the robustness and accuracy of recognition systems. This review has synthesized the current state of research and laid the groundwork for future developments, guiding researchers toward innovative solutions and applications.

References

1. B. Ren, M. Liu, R. Ding, H. Liu, A survey on 3D skeleton-based action recognition using learning method. Cyborg and Bionic Systems. **5**, 0100 (2024)
2. H. Aguinis, R. K. Gottfredson, H. Joo, Best-practice recommendations for defining, identifying, and handling outliers. Organizational Research Methods. **16**, 270-301 (2013)
3. P. Venkataanusha, Ch. Anuradha, et al., Detecting outliers in high dimensional data sets using Z-score methodology. International Journal of Innovative Technology and Exploring Engineering. **8**, 48-53 (2019)
4. C. S. K. Dash, A. K. Behera, S., Dehuri, et al., An outliers detection and elimination framework in classification task of data mining. Decision Analytics Journal. **6**, 100164 (2023)
5. G. Evangelidis, G. Singh, R. Horaud, Skeletal quads: Human action recognition using joint quadruples, in Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, (2014)
6. R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3D skeletons as points in a Lie group, in Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, (2014)
7. W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3D points, in Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, USA, (2010)
8. X. Yang, Y. Tian, Effective 3D action recognition using EigenJoints. Journal of Visual Communication and Image Representation. **25**, 2-11 (2014)
9. B. Su, H. Wu, M. Sheng, et al., Accurate hierarchical human actions recognition from Kinect skeleton data. IEEE Access. **7**, 52532-52541 (2019)

10. C. Tang, W. Li, P. Wang, et al., Online human action recognition based on incremental learning of weighted covariance descriptors. Information Sciences. **467**, 219-237 (2018).
11. A. Shahroudy, J. Liu, T. T. Ng, et al., NTU RGB+D: A large-scale dataset for 3D human activity analysis, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1010-1019 (2016)
12. A. Zisserman, J. Carreira, K. Simonyan, et al., The Kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
13. H. Duan, Y. Zhao, K. Chen, et al., Revisiting skeleton-based action recognition, in Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, (2022)
14. H. Liu, J. Tu, M. Liu, Two-stream 3D convolutional neural network for skeleton-based action recognition. arXiv preprint arXiv:1705.08106 (2017)
15. C. Li, Q. Zhong, D. Xie, et al., Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation, in Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, (2018)
16. Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton-based action recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, (2015)
17. V. Veeriah, N. Zhuang, G. J. Qi, Differential recurrent neural networks for action recognition, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), (2015) 4041-4049
18. J. Liu, A. Shahroudy, D. Xu, et al., Spatio-temporal LSTM with trust gates for 3D human action recognition, in Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, October 11-14, 2016. Lecture Notes in Computer Science, 9907, (2016) 816-833
19. S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in Proceedings of the AAAI Conference on Artificial Intelligence, **32**(1), (2018)
20. K. Cheng, Y. Zhang, X. He, et al., Skeleton-based action recognition with shift graph convolutional network, in Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA (2020)
21. M. Li, S. Chen, X. Chen, et al., Actional-structural graph convolutional networks for skeleton-based action recognition, in Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, (2019)
22. L. Shi, Y. Zhang, J. Cheng, et al., Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition, in Proceedings of the Asian Conference on Computer Vision (ACCV), (2020)
23. H. Qiu, B. Hou, B. Ren, et al., Spatio-temporal tuples transformer for skeleton-based action recognition. arXiv preprint arXiv:2201.02849 (2022)
24. F. Shi, C. Lee, L. Qiu, et al., STAR: Sparse transformer-based action recognition. arXiv preprint arXiv:2107.07089 (2021)
25. J. Liu, A. Shahroudy, M. Perez, et al., NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence. **42**(10), 2684-2701 (2019)