Cyborg and Bionic Systems
A SCIENCE PARTNER JOURNAL

REVIEW ARTICLE

A Survey on 3D Skeleton-Based Action Recognition Using Learning Method

Bin Ren[1,2], Mengyuan Liu[3*], Runwei Ding[4], and Hong Liu[3]

[1]University of Pisa, Pisa, Italy. [2]University of Trento, Trento, Italy. [3]National Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School, Shenzhen, China. [4]Peng Cheng Laboratory, Shenzhen, China.

*Address correspondence to: nkliuyifang@gmail.com

Citation: Ren B, Liu M, Ding R, Liu H. A Survey on 3D Skeleton-Based Action Recognition Using Learning Method. *Cyborg Bionic Syst.* 2024;5:Article 0100. https://doi.org/10.34133/cbsystems.0100

Submitted 9 October 2023
Accepted 25 January 2024
Published 16 May 2024

Three-dimensional skeleton-based action recognition (3D SAR) has gained important attention within the computer vision community, owing to the inherent advantages offered by skeleton data. As a result, a plethora of impressive works, including those based on conventional handcrafted features and learned feature extraction methods, have been conducted over the years. However, prior surveys on action recognition have primarily focused on video or red-green-blue (RGB) data-dominated approaches, with limited coverage of reviews related to skeleton data. Furthermore, despite the extensive application of deep learning methods in this field, there has been a notable absence of research that provides an introductory or comprehensive review from the perspective of deep learning architectures. To address these limitations, this survey first underscores the importance of action recognition and emphasizes the significance of 3-dimensional (3D) skeleton data as a valuable modality. Subsequently, we provide a comprehensive introduction to mainstream action recognition techniques based on 4 fundamental deep architectures, i.e., recurrent neural networks, convolutional neural networks, graph convolutional network, and Transformers. All methods with the corresponding architectures are then presented in a data-driven manner with detailed discussion. Finally, we offer insights into the current largest 3D skeleton dataset, NTU-RGB+D, and its new edition, NTU-RGB+D 120, along with an overview of several top-performing algorithms on these datasets. To the best of our knowledge, this research represents the first comprehensive discussion of deep learning-based action recognition using 3D skeleton data.

Introduction

Action analysis, a pivotal and vigorously researched topic in the field of computer vision, has been under investigation for several decades [1–4]. The ability to recognize actions is of paramount importance, as it enables us to understand how humans interact with their surroundings and express their emotions [5,6]. This recognition can be applied across a wide range of domains, including intelligent surveillance systems, human–computer interaction, virtual reality, and robotics [7–9]. In recent years, the field of skeleton-based action recognition has made significant strides, surpassing conventional hand-crafted methods. This progress has been chiefly driven by substantial advancements in deep learning methodologies [10–19].

Traditionally, action recognition has relied on various data modalities, such as red-green-blue (RGB) image sequences [20–24], the depth image sequences [25,26], videos, or a fusion of these modalities (e.g., RGB combined with the optical flow) [27–31]. These approaches have yielded impressive results through various techniques. Compared to skeleton data, which offers a detailed topological representation of the human body through joints and bones, these alternative modalities often prove computationally intensive and less robust when confronted with complex backgrounds and variable conditions. This includes challenges posed by variations in body scales, viewpoints, and motion speeds [32,33].

Furthermore, the availability of sensors like the Microsoft Kinect [34] and advanced human pose estimation algorithms [35–38] has facilitated the acquisition of accurate 3-dimensional (3D) skeleton data [39]. Figure 1 provides a visual representation of human skeleton data. In this case, 25 body joints are captured for a given human body. Skeleton sequences possess several advantages over other modalities, characterized by 3 notable features: (a) Intraframe spatial information, where strong correlations exist between joints and their adjacent nodes, enabling the extraction of rich structural information. (b) Interframe temporal information, which captures strong and clear temporal correlations between frames of each body joint, enhancing the potential for action recognition. (c) A co-occurrence relationship between spatial and temporal domains when considering joints and bones, offering a holistic perspective. These unique attributes have catalyzed substantial research endeavors in human action recognition and detection. The escalating integration of skeleton data is anticipated to pervade diverse applications in the field.

The recognition of human actions using skeleton sequences predominantly hinges on a temporal dimension, transforming it into both a spatial and temporal information modeling challenge. As a result, traditional approaches in skeleton-based methods focus on extracting motion patterns from these sequences, prompting extensive research into handcrafted features. [31,40–44]. These features often entail capturing the

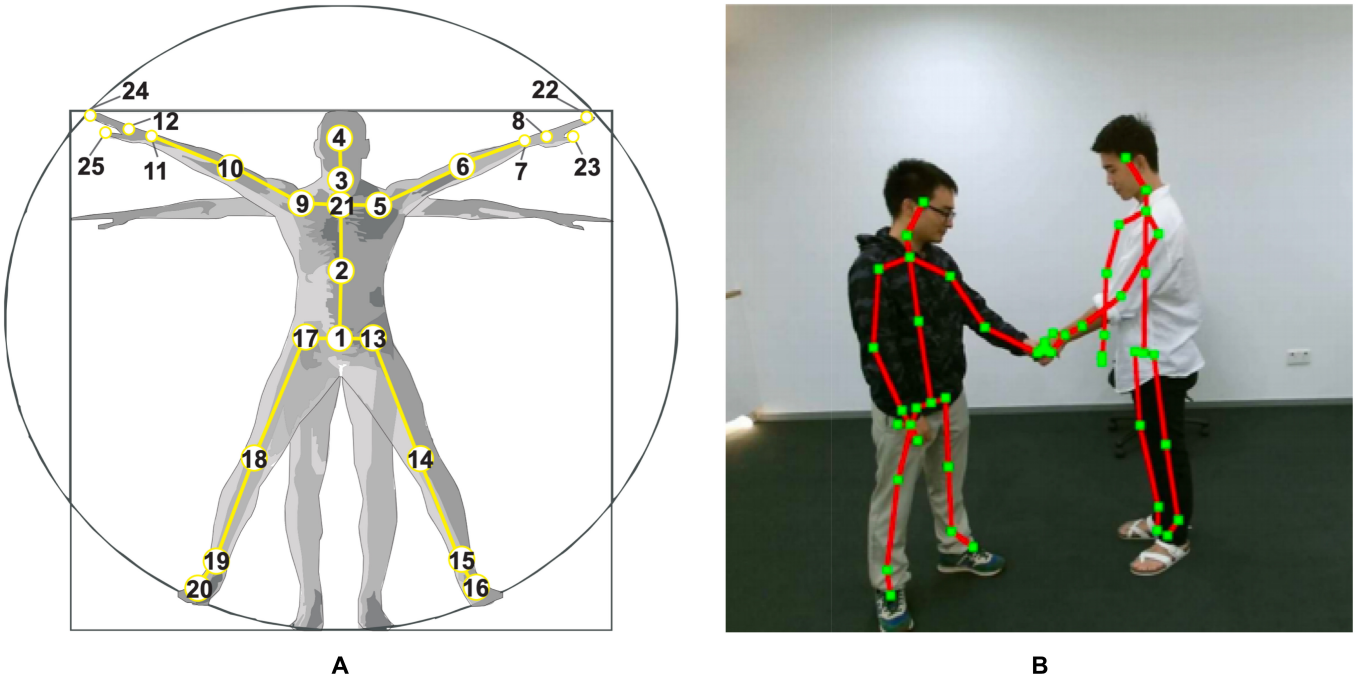

Fig. 1. Examples of skeleton data in NTU RGB+D / NTU RGB+D 120 datasets [131,132]. (A) Configuration of 25 body joints in the dataset. (B) RGB+joints representation of the human body.

relative 3D rotations and translations among different joints or body parts [13,45]. However, it has become evident that handcrafted features perform well only on specific datasets [46], highlighting the challenge that features tailored for one dataset may not be transferable to others. This issue hampers the generalization and broader application of action recognition algorithms.

With the remarkable development and outstanding performance of deep learning methods in various computer vision tasks, such as image classification [47,48] and object detection [49,50], the application of deep learning to skeleton data for action recognition has gained prominence. Nowadays, deep learning techniques utilizing recurrent neural networks (RNNs) [51], convolutional neural networks (CNNs) [52], graph convolutional networks (GCNs), and Transformer-based methods have emerged in this field [53,54]. Figure 2 provides an overview of the general pipeline for 3D skeleton-based action recognition (3D SAR) using deep learning, starting from raw RGB sequences or videos and culminating in action category prediction. RNN-based methods leverage skeleton sequences as natural time series data, treating joint coordinates as sequential vectors, aligning well with the RNN's capacity for processing time series information. To enhance the learning of temporal context within skeleton sequences, variants like long short-term memory (LSTM) and gated recurrent unit (GRU) have been employed. Meanwhile, CNNs complement RNN-based techniques, as they excel at capturing spatial cues in the input data, which RNNs may lack. Additionally, a relatively recent approach, the GCNs has gained attention for its ability to model skeleton data in a natural topological graph structure, with joints and bones as vertices and edges, respectively, offering advantages over alternative formats like images or sequences. Transformer-based methods [55–60] capture the spatial-temporal relation of the input 3D skeleton data mainly based on its core multihead self-attention (MSA) mechanism.

All these 3 kinds of deep learning-based architectures have already gained unprecedented performance, but most review works just focus on traditional techniques or deep learning-based methods just with the RGB image or RGB-D data method. Ronald Poppe et al. [61] firstly addressed the basic challenges and characteristics of this domain and then gave a detailed illumination of basic action classification methods about direct classification and temporal state-space models. Daniel and Remi et al. [62] showed an overall overview of the action representation only in both spatial and temporal domains. Though the methods mentioned above provide some inspiration that may be used for input data preprocessing, neither skeleton sequence nor deep learning strategies were taken into account. Recently, Wu et al. [63] and Herath et al. [64] offered a summary of deep learning-based video classification and captioning tasks, in which the fundamental structure of CNN, as well as RNNs, was introduced, and the latter made a clarification about common deep architectures and quantitative analysis for action recognition. To our best knowledge, [65] is the first work recently giving an in-depth study in 3D SAR, which concludes this issue from the action representation to the classification methods. In the meantime, it also offers some commonly used datasets such as UCF, MHAD, MSR daily activity 3D, etc. [66–69], while it does not cover the emerging GCN based methods. Finally, [46] proposed a new review based on Kinect-dataset-based action recognition algorithms, which organized a thorough comparison of those Kinect-dataset-based techniques with various types of input data including RGB, Depth, RGB+Depth, and skeleton sequences. Reference [70] presented an overview of the action recognition across all the data modalities but without presenting the Transformer-based methods. In addition, all these works mentioned above also ignore the differences and motivations among CNN-based, RNN-based, GCN-based, and Transformer-based methods, especially when taking the 3D skeleton sequences into account.

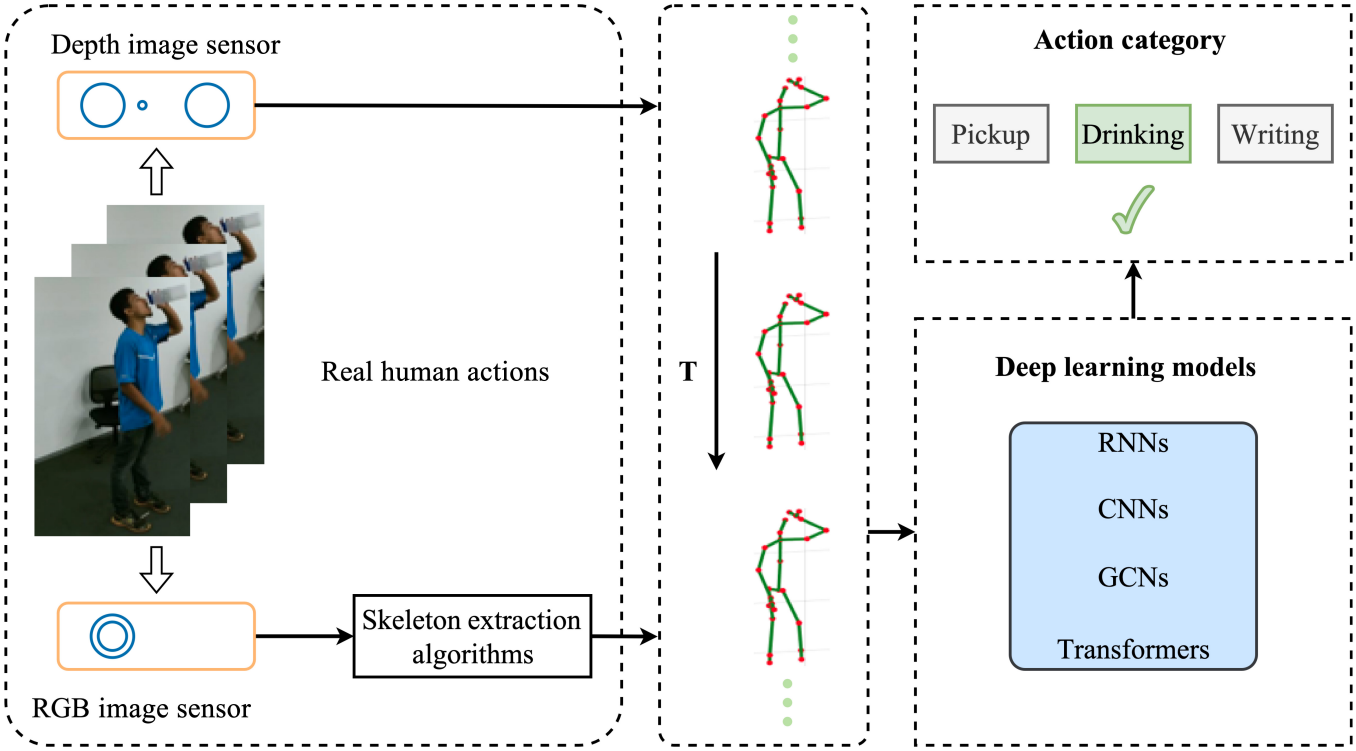

Fig. 2. The general pipeline of skeleton-based action recognition using deep learning methods. Firstly, the skeleton data was obtained in 2 ways, directly from depth sensors or from pose estimation algorithms. The skeleton will be sent into RNNs, CNNs, GCNs, or Transformer-based neural networks. Finally, we get the accurate action category.

To address these issues comprehensively, this survey aims to provide a detailed summary of 3D SAR employing 4 fundamental deep learning architectures: RNNs, CNNs, GCNs, and Transformers. Additionally, we delve into the motivations behind the choice of these models and offer insights into potential future directions for research in this field.

In summary, our study encompasses 4 key contributions:

• A comprehensive introduction about the superiority of 3D skeleton sequence data and characteristics of 3 kinds of fundamental deep architectures are presented in a detailed and clear manner, and a general pipeline in 3D SAR using deep learning methods is illustrated.

• Within each type of deep architecture, numerous contemporary methods leveraging skeleton data are introduced, focusing on data-driven approaches. These encompass spatial-temporal modeling, innovative skeleton data representation, and methods for co-occurrence feature learning.

• The discussion begins by addressing the latest challenging datasets, notably the NTU-RGB+D 120, along with an exploration of several top-ranked methods. Subsequently, it delves into envisaged future directions in this domain.

• The initial study comprehensively examines 4 foundational deep architectures, encompassing RNN-based, CNN-based, GCN-based, and Transformer-based methods within the domain of 3D SAR.

3D SAR with Deep Learning

While existing surveys have offered comprehensive comparisons of action recognition techniques based on RGB or skeleton data, they often lack a detailed examination from the perspective of neural networks. To bridge this gap, we provide a concise introduction to the fundamental properties of each architecture (Preliminaries: Basic properties of RNNs, CNNs, GCNs, and Transformers). Then our survey provides an exhaustive discussion and comparison of RNN-based (RNN-based methods), CNN-based (CNN-based methods), GCN-based (GCN-based methods), and Transformer-based (Transformer-based methods) methods for 3D SAR. We will explore these methods in-depth, highlighting their strengths and weaknesses, and introduce several latest related works as case studies, focusing on specific limitations or classic spatial-temporal modeling challenges associated with these neural network models.

Preliminaries: Basic properties of RNNs, CNNs, GCNs, and Transformers

Before delving into the specifics of each method, we provide a brief overview of the fundamental architecture, outlining their respective advantages, disadvantages, and coarse selection criteria under the 3D SAR setting.

RNNs

RNNs are ideal for capturing temporal dependencies in sequences of joint movements over time and are suited for modeling action sequences due to their ability to retain temporal information. However, RNNs are also vulnerable to long-term dependencies, potentially missing complex relationships in lengthy sequences, and are computationally inefficient due to sequential processing, leading to longer training times for large-scale datasets.

CNNs

CNNs are not only effective in capturing spatial patterns from the joint coordinates, recognizing spatial features within individual

frames of the 3D skeleton data but also great for local spatial relationships among joints. However, CNNs are limited to capturing temporal evolution in sequences, potentially missing out on the temporal dynamics crucial for action recognition.

GCNs

GCNs are designed to manage graph-structured data such as skeletal joint connections in action recognition, enabling the learning of relationships between joints and their connectivity while integrating spatial and temporal information. However, GCNs can be sensitive to noisy or irregular connections among joints, potentially impacting recognition accuracy, particularly in complex actions.

Transformers

Transformers is not only efficient at capturing long-range dependencies without the vanishing/exploding gradient issue but also versatile in handling multiple modalities and learning global relationships. However, it is also computationally intensive due to attention mechanisms, potentially requiring substantial computational resources. What is more, compared to RNNs, it is also limited to sequential locality

RNN-based methods

Recursive connections within the RNN structure are established by feeding the output of the previous time step as the input to the current time step, as demonstrated in prior work [71]. This approach is known to be effective for processing sequential data. In a similar vein, models like the standard RNN, LSTM, and GRU were introduced to address limitations such as gradient-related issues and the modeling of long-term temporal dependencies that were present in the standard RNN.

From the first aspect, spatial-temporal modeling can be seen as the principle in action recognition tasks. Due to the weakness of the spatial modeling ability of RNN-based architecture, the performance of some related methods generally could not gain a competitive result [72–74]. Recently, Hong et al. [75] proposed a novel 2-stream RNN architecture to model both temporal dynamics and spatial configurations for skeleton data. Figure 3 shows the framework of their work. An exchange of the skeleton axes was applied for the data level preprocessing for the spatial dominant learning. Unlike [75], Jun et al. [76] stepped into the traversal method of a given skeleton sequence to acquire the hidden relationship of both domains. Compared with the general method which arranges joints in a simple chain so that ignores the kinetic dependency relations between adjacent joints, the mentioned tree-structure-based traversal would not add false connections between body joints when their relation is not strong enough. Then, using an LSTM with a trusted gate, the input is treated discriminately, through which, if the tree-structured input unit is reliable, the memory cell will be updated by importing input latent spatial information. Inspirited by the property of CNN, which is extremely suitable for spatial modeling. Li et al. [77] incorporated an attention RNN with a CNN model to enhance the complexity of spatial-temporal modeling. Initially, they introduced a temporal attention module within a residual learning module, allowing for the recalibration of temporal attention across frames within a skeleton sequence. Subsequently, they applied a spatial-temporal convolutional module to this first module, treating the calibrated joint sequences as images. Furthermore, in the work by Lin et al. [78], an attention recurrent relation LSTM network was employed. This network combines a recurrent relation network for spatial features with a multilayer LSTM to capture temporal features within skeleton sequences.

The second aspect involves the network structure, serving as a solution to address the limitations of standard RNNs. While RNNs are inherently suitable for sequence data, they often suffer from well-known problems like gradient exploding and vanishing. Although LSTM and GRU have alleviated these issues to some extent, the use of hyperbolic tangent and sigmoid activation functions can still result in gradient decay across layers. In response, new types of RNN architectures have been proposed [79–81]. Shuai et al. [81] introduced an independently recurrent neural network (IndRNN) designed to address gradient exploding and vanishing problems, making it feasible and more robust to construct longer and deeper RNNs for high-level semantic feature learning. This modification for RNNs is not limited to skeleton-based action recognition but can also find applications in other domains, such as language modeling. In the IndRNN structure, neurons in one layer operate independently of each other, enabling the processing of much longer sequences.

Finally, the third aspect is the data-driven pipeline. In the consideration that not all joints are informative for an action analysis, [82] add global context-aware attention to LSTM networks, which selectively focus on the informative joints in a skeleton sequence. Figure 4 illustrates the visualization of the proposed method, from the figure we can conclude that the more informative joints are addressed with a red circle color area, indicating those joints are more important for this special action. In addition, because the skeletons provided by datasets or depth sensors are not perfect, which would affect the result of an action recognition task, [83] transform skeletons into another coordinate system for the robustness to scale, rotation and translation first and then extract salient motion features from the transformed data instead of sending the raw skeleton data to LSTM. Figure 4B shows the feature representation process.

Numerous valuable works have utilized RNN-based methods to address challenges related to large viewpoint changes and the relationships among joints within a single skeleton frame. However, it is essential to acknowledge that in specific modeling aspects, RNN-based methods may exhibit limitations compared to CNN-based approaches. In the following sections, we delve into an intriguing question: How do CNN-based methods perform temporal modeling, and how can they strike the right balance between spatial and temporal information in action recognition?

CNN-based methods

While CNNs offer efficient and effective high-level semantic cue learning, they are primarily tailored for regular image tasks. However, action recognition from skeleton sequences presents a distinct challenge due to its inherent time-dependent nature. Achieving the right balance and maximizing the utilization of both spatial and temporal information within a CNN-based architecture remains a challenging endeavor.

Typically, from the spatial-temporal modeling aspect, most of the CNN-based methods explored the representation of 3D skeleton sequences. Specifically, to accommodate the input requirements of CNNs, 3D-skeleton sequence data undergoes the transformation from a vector sequence to a pseudo-image format. However, achieving a suitable representation that effectively combines both spatial and temporal information can be challenging. Consequently, many researchers opt to encode skeleton joints

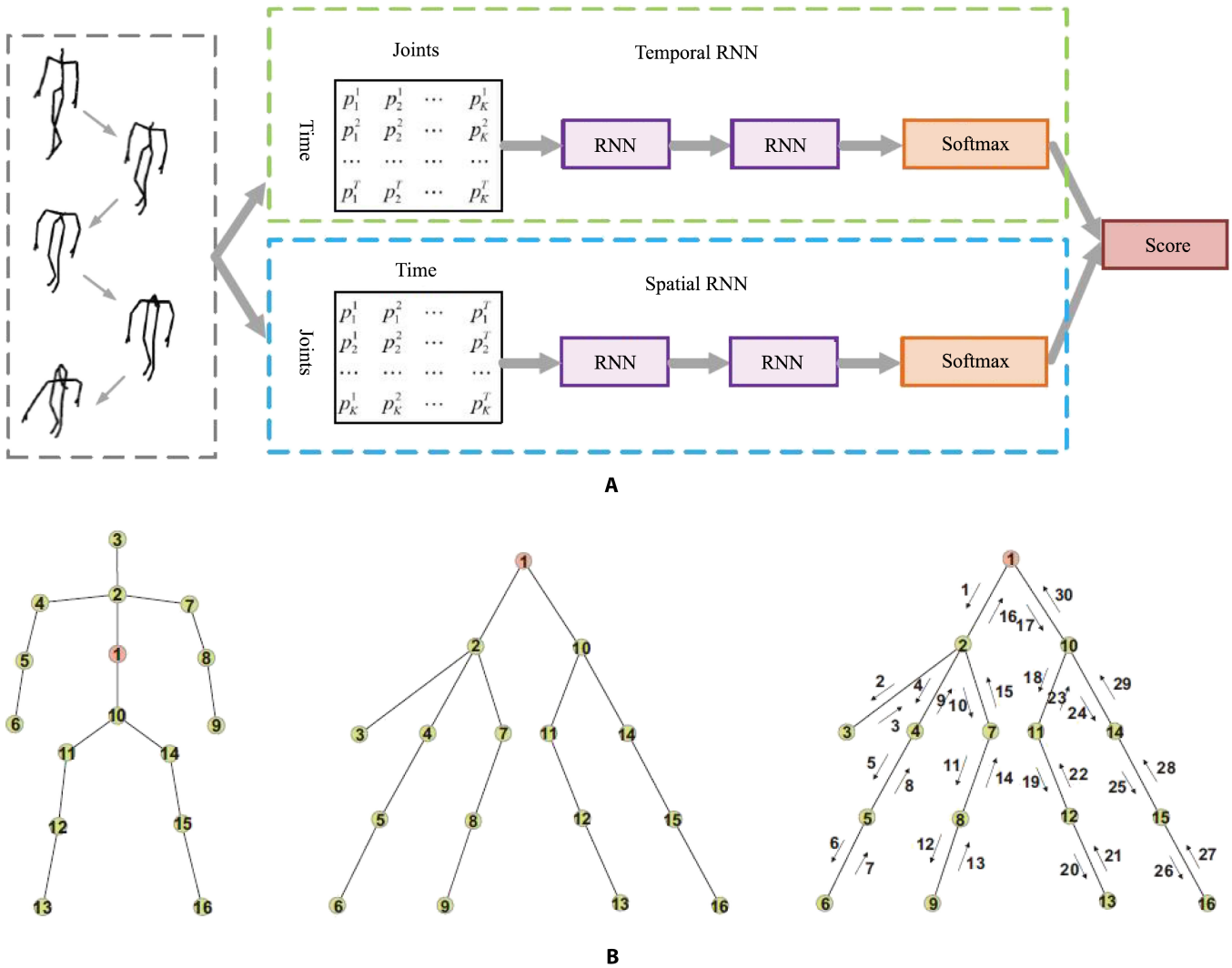

Fig. 3. Examples of mentioned methods for dealing with spatial modeling problems. (A) Two-stream framework that enhances the spatial information by adding a new stream [75]. (B) Data-driven technique that addresses the spatial modeling ability by giving a transform toward original skeleton sequence data [76].

into multiple 2D pseudo-images, which are subsequently fed into CNNs to facilitate the learning of informative features [84,85]. Wang et al. [86] proposed the joint trajectory maps, which represent spatial configuration and dynamics of joint trajectories into 3 texture images through color encoding. However, this kind of method is a little complicated and also loses importance during the mapping procedure. To tackle this shortcoming, Li et al. [87] used a translation-scale invariant image mapping strategy which firstly divided human skeleton joints in each frame into 5 main parts according to the human physical structure, then those parts were mapped to 2D form. This method makes the skeleton image consist of both temporal information and spatial information. However, though the performance was improved, there is no reason to take skeleton joints as isolated points, cause in the real world, imitate connection exists among our body, for example, when waiting for the hands, not only the joints directly within the hand should be taken into account, but also other parts such as shoulders and legs are considerable. Li et al. [88] proposed the shape-motion representation from geometric algebra, which addressed the importance of both joints and bones and fully utilized the information provided by the skeleton sequence. Similarly, [13] also use the enhanced skeleton visualization to represent the skeleton data, and Carlos et al. [89] also proposed a new representation named SkeleMotion based on motion information that encodes the temporal dynamics by explicitly computing the magnitude and orientation values of the skeleton joints. Figure 5A shows the shape-motion representation proposed by [88], while Fig. 5B illustrates the SkeleMotion representation. What is more, similarly to SkeleMotion, [90] uses the framework of SkeleMotion but is based on tree structure and reference joints for a skeleton image representation.

Commonly, CNN-based methods represent a skeleton sequence as an image by encoding temporal dynamics and skeleton joints as rows and columns, respectively. However, this simplistic approach may limit the model's ability to capture co-occurrence features, as it considers only neighboring joints within the convolutional kernel and may overlook latent correlations involving all joints. Consequently, CNNs might fail to learn the corresponding and useful features. In response to this limitation, Chao et al. [91] introduced an end-to-end framework designed to learn co-occurrence features through a hierarchical methodology. This approach gradually aggregates different levels of contextual information, beginning with the independent encoding of point-level information, which

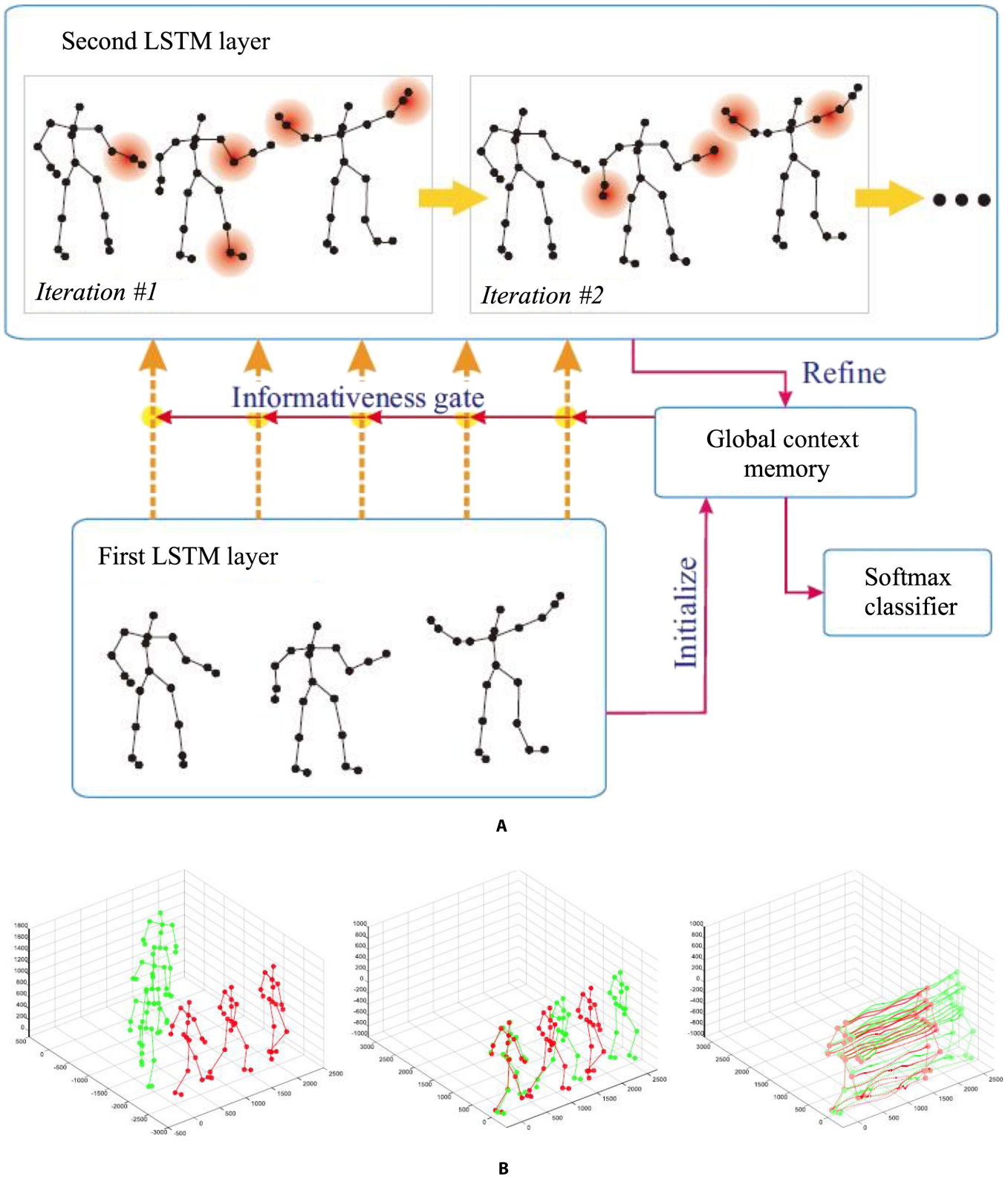

Fig. 4. Data-driven based method. (A) Different importance among different joints for a given skeleton action [82]. (B) Feature representation processes, from left to right are original input skeleton frames, transformed input frames, and extracted salient motion features, respectively [83].

is then assembled into semantic representations within both temporal and spatial domains.

Besides explorations in the representation of 3D skeleton sequences, there also exist some other problems in CNN-based techniques. For example, to find a balance between the model size and the corresponding inference efficiency, DD-Net [14] was proposed to model double feature and double motion via CNN for efficient solutions. Kim et al. [92] proposed to use

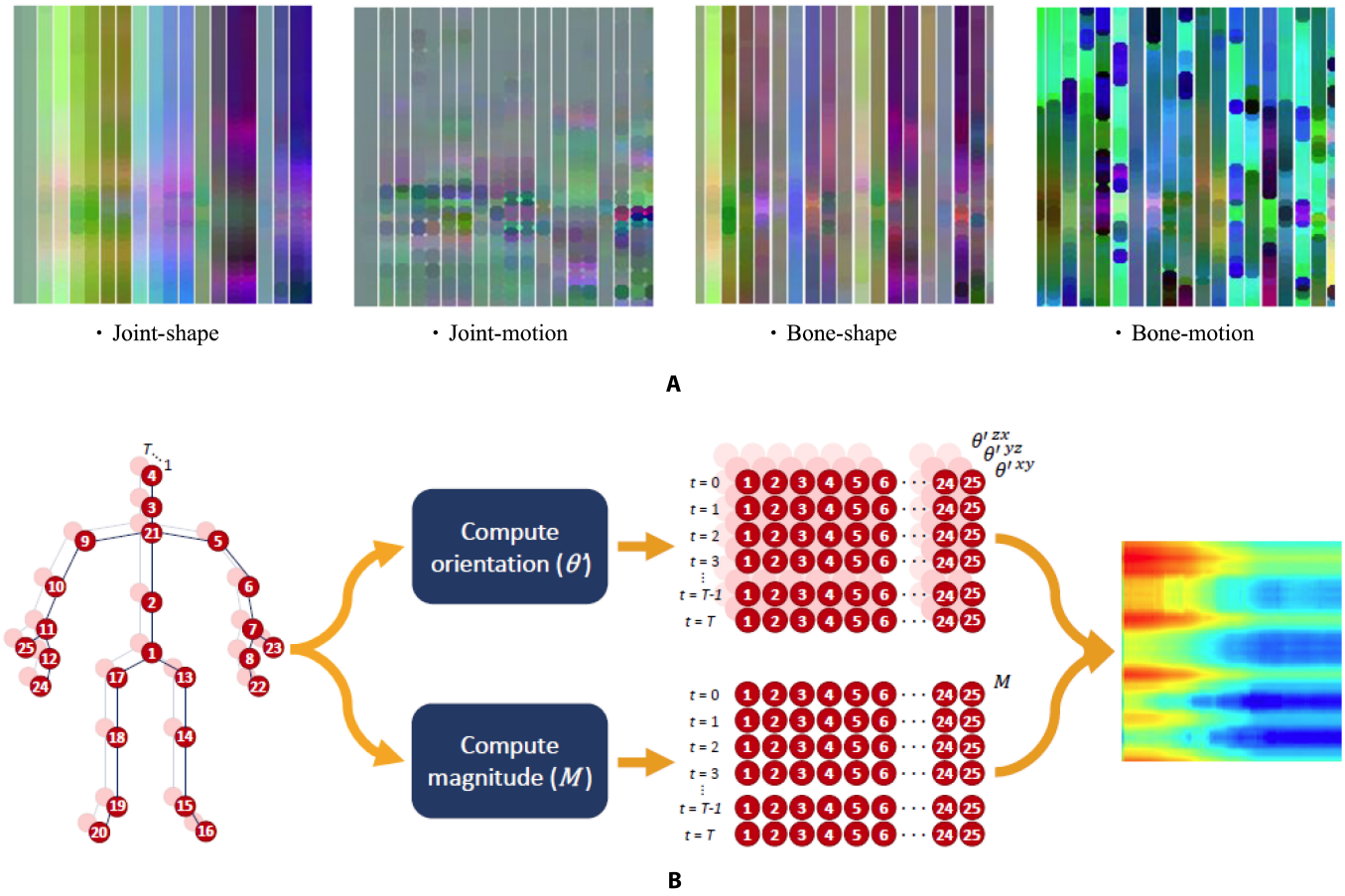

Fig. 5. Examples of the proposed representation skeleton image. (A) Skeleton sequence shape-motion representations [88] generated from "pick up with one hand" on Northwestern-UCLA dataset [133]. (B) SkeleMotion representation workflow [90].

the temporal CNN for modeling the interpretable spatiotemporal cues [93]. As a result, the point-level feature of each joint is learned. In addition, 2-stream and 3-stream CNN-based heavy models are also proposed for improving the representation learning ability for spatial-temporal modeling [94]. So the skeleton-based action recognition using CNN is still an open problem waiting for researchers to dig in.

GCN-based methods

Drawing inspiration from the inherent topological graph structure of human 3D-skeleton data, distinct from the sequential vector or pseudo-image treatments in RNN-based or CNN-based methods. Recently, the graph convolution network has been adopted in this task frequently due to the effective representation of the graph structure data. Generally, 2 kinds of graph-related neural networks can be found, i.e., the graph neural networks and RNNs, and graph and convolutional neural networks (GCNs). In this survey, we mainly pay attention to the latter. This focus yielded compelling results, as evidenced by the performance of the GCN-based method on the rank board. Furthermore, merely encoding the skeleton sequence into a vector or 2D grid fails to fully capture the interdependence among correlated joints from the skeleton's perspective. Conversely, GCNs present adaptability to diverse structures, such as the skeleton graph. Nonetheless, the principal challenge within GCN-based approaches persists in the handling of skeleton data, particularly in structuring the original data into a coherent graph format. Yan et al. [53] first presented a novel model, the spatial-temporal graph convolutional networks (ST-GCNs), for skeleton-based action recognition. Specifically, the approach first involved the creation of a spatial-temporal graph, wherein the joints functioned as graph vertices, establishing inherent connections within the human body structure and across temporal sequences as the graph edges. Following this step, the ST-GCN's higher-level feature maps on the graph underwent classification using a standard Softmax classifier, assigning them to their respective action categories. This work has notably directed attention toward employing GCNs for skeleton-based action recognition, resulting in a surge of recent related research [95–100].

Built upon GCNs, 2 main common aspects are explored, i.e., more representative manner for the construction of the skeleton data and more effective designs of the GCN-based model [101,102].

From the first aspect, [101] proposed the action-structural graph convolutional networks (AS-GCNs) could not only recognize a person's action but also use a multitask learning strategy to output a prediction of the subject's next possible pose. The constructed graph in this work can capture richer dependencies among joints by 2 modules called Actional Links and Structural Links. Figure 6 shows the feature learning and its generalized skeleton graph of AS-GCN. Multitask learning strategy used in this work may be a promising direction because the target task would be improved by the other task as a complementary. To capture and enhance richer feature representations,

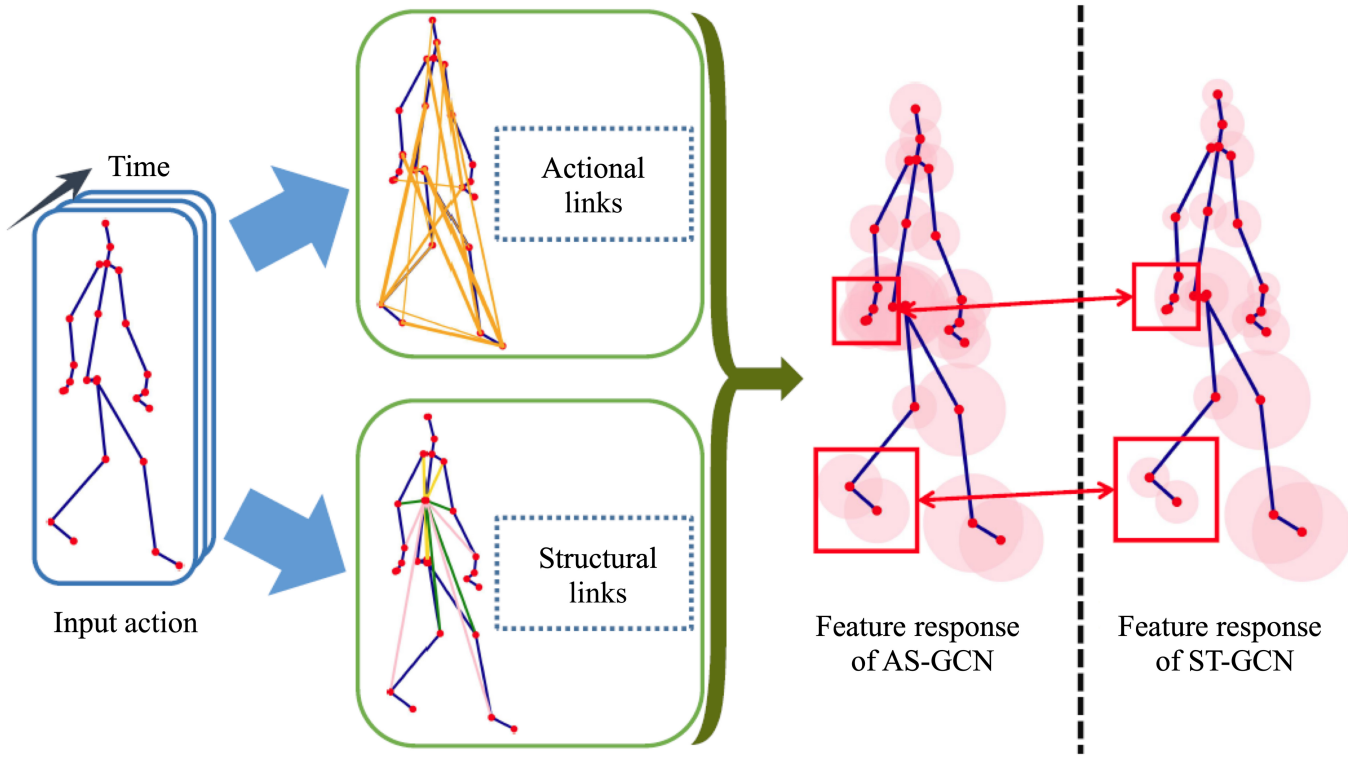

Fig. 6. Feature learning with generalized skeleton graphs [101].

Shi et al. [95] introduced the 2s-AGCN, which incorporates an adaptive topology graph. This approach allows for automatic updates leveraging the neural network's backpropagation algorithm, effectively enhancing the characterization of joint connection strengths. Liu et al. [103] proposes MS-G3D which constructs a unified spatial-temporal graph. This big spatial-temporal graph is composed of several subgraphs, and each subgraph represents the spatial relationships of joints on a certain frame. This form of the adjacent matrix can effectively model the relationship between different joints in different frames. Similarly, there are also a lot of following-up methods proposed for constructing more representative graphs [104–106].

From the second aspect, traditional GCNs operate as straight feed-forward networks, limiting low-level layers' access to semantic information from higher-level layers. To address this, Yang et al. [107] introduce the feedback graph convolutional network (FGCN) aimed at incrementally acquiring global spatial-temporal features. Departing from direct manipulation of the complete skeleton sequence, FGCN adopts a multistage temporal sampling strategy to sparsely extract a sequence of input clips from the skeleton data. Furthermore, Bian et al. [108] introduces a structural knowledge distillation scheme aimed at mitigating accuracy loss resulting from low-quality data, thereby enhancing the model's resilience to incomplete skeleton sequences. Fang et al. [109] presents the spatial-temporal slow-fast graph convolutional network (STSF-GCN), which conceptualizes skeleton data akin to a unified spatial-temporal topology, reminiscent of MS-G3D.

From the preceding introduction and discussion, it is evident that the predominant concern revolves around data-driven approaches, seeking to uncover latent insights within 3D skeleton sequence data. In the realm of GCN-based action recognition, the central query persists: "How do we extract this latent information?" This question remains an ongoing challenge. Particularly noteworthy is the inherent temporal-spatial correlation within the skeleton data itself. The optimal utilization of these 2 aspects warrants further exploration. There remains substantial potential for enhancing their effectiveness, calling for deeper investigation and innovative strategies to maximize their utilization.

Transformer-based methods

Transformers [110] demonstrated their overwhelming power on a broad range of language tasks (e.g., text classification, machine translation, or question answering [110,111]), and the vision community follows it closely and extends it for vision tasks, such as image classification [48,112,113], object detection [49,50], segmentation [114], image restoration [115,116], and point cloud registration [117–119]. The emergence of transformer algorithms marks a pivotal shift in point-centric research. These transformer-based methods are gradually challenging the dominance of GCN methods, showcasing promising advancements in computational efficiency and accuracy. Upon analysis, we firmly believe that transformer-based approaches retain robust potential and are poised to become the mainstream technique in the future.

The core module in Transformer, MSAs [48,110] aggregate sequential tokens with normalized attention as: $\mathbf{z}_j = \sum_i \text{Softmax}\left(\frac{\boldsymbol{QK}}{\sqrt{d}}\right)_i \boldsymbol{V}_{i,j}$ where $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ are query, key and value matrices, respectively. d is the dimension of query and key, and $\mathbf{z}_j$ is the j-th output token. This step usually represents the context relation computation and update of the overall 3D skeleton features. Building upon the MSA from the Transformer for solving

the 3D-SAR problem, there are lots of transformer architecture-based solutions are proposed.

In particular, Cho et al. [120] proposed a novel model called Self-Attention Network (SAN) that completely utilizes the self-attention mechanism to model spatial-temporal correlations. Shi et al. [121] proposed a decoupled spatial-temporal attention network (DSTA-Net) that contains spatial-temporal attention decoupling, decoupled position encoding, and global spatial regularization. DSTA-Net decouples the skeleton data into 4 streams, namely, spatial-temporal stream, spatial stream, slow-temporal stream, and fast-temporal stream; each data stream focuses on expressing a particular aspect of the action. Plizzari et al. [122] proposed a novel Spatial-Temporal Transformer network (ST-TR) in which the spatial self-attention module and temporal self-attention module are used to capture the correlation between different nodes in a frame and the dynamic relationship between the same node in the whole frames. To handle action sequences of varying lengths proficiently, Ibh et al. [123] proposed TemPose, which leaves out the padded temporal and interaction tokens in the attention map. At the same time, Tempose codes the position of the player and the position of the badminton ball to predict the action class together.

The Transformer-based approach effectively mitigates the issue of solely concentrating on local information and excels in capturing extensive dependencies over long sequences. When applied to tasks involving skeleton-based human behavior recognition, the Transformer architecture demonstrates adeptness in capturing temporal relationships. However, its efficacy in modeling spatial relationships remains constrained due to limitations in capturing and encoding the intricate high-dimensional semantic information inherent in skeleton data [124,125]. Simultaneously, numerous approaches have emerged that amalgamate the Transformer with GCNs or CNNs, thereby forming hybrid architectures. These models are designed with the aspiration of harnessing the strengths inherent in each fundamental architecture. By combining the Transformer's capabilities with the specialized strengths of RNNs, CNNs, or GCNs, these hybrid models aim to achieve a more comprehensive and powerful framework for diverse tasks [56,126–128].

Latest Datasets and Performance

Skeleton sequence datasets such as MSRAAction3D [129], 3D Action Pairs [130], and MSR Daily Activity3D [68] have be analyzed in lots of previous surveys [46,64,65]. In this survey, we mainly address the following 2 recent datasets, NTU-RGB+D [131] and NTU-RGB+D 120 [132].

The NTU-RBG+D dataset, introduced in 2016, stands as a significant resource, comprising 56,880 video samples gathered through Microsoft Kinect-v2. This dataset holds a prominent position as one of the largest collections available for skeleton-based action recognition. It furnishes the 3D spatial coordinates of 25 joints for each human depicted in an action, as illustrated in Fig. 1A. For assessing the proposed methods, 2 evaluation protocols are suggested: Cross-Subject and Cross-View. The Cross-Subject setting involves 40,320 samples, with 16,560 allocated for training and evaluation, employing a split of 40 subjects into training and evaluation groups. In the case of Cross-View, comprising 37,920 and 18,960 samples, the evaluation uses camera 1 while training is conducted using cameras 2 and 3. Recently, an extended version of the original NTU-RGB+D dataset known as NTU-RGB+D 120 has been introduced. This extended dataset comprises 120 action classes and encompasses a total of 114,480 skeleton sequences, significantly expanding the scope. Additionally, the viewpoints have increased to 155.

In Tables 1 and 2, we present the performance of recent skeleton-based techniques relevant to NTU-RGB + D and NTU-RGB + D 120 datasets, respectively. Note that in NTU-RGB+D, "CS" stands for Cross-Subject, and "CV" stands for Cross-View. For NTU-RGB + D120, there are 2 settings, i.e., Cross-Subject (C-Subject), and Cross-Setup (C-Setup).

Based on the observation of the performance of these 2 datasets, we find that it is evident that existing algorithms have achieved impressive performances in the original NTU-RGB+D dataset. However, the newer NTU-RGB+D 120 poses a significant challenge, indicating that further advancements are needed to effectively address this more complex dataset. It is worth noting that the GCN-based methods achieved the leading results compared to the other 2 architectures. In addition to the very fundamental architectures (i.e., RNNs, CNNs, and GCN), the most recent Transformer [110] based methods also show their promising performance on both datasets. It is also easy to find that a hybrid Transformer and other architectures also further boost the overall performance of the 3D SAR.

Discussion

Considering the performance and attributes of the aforementioned deep architectures, several critical points warrant further discussion concerning the criteria for architecture selection. In terms of accuracy and robustness, GCNs demonstrate potential excellence by adeptly capturing spatial and temporal relationships among joints. RNNs exhibit proficiency in capturing temporal dynamics, while CNNs excel in identifying spatial features. When evaluating computational efficiency, CNNs boast faster processing capabilities owing to their parallel processing nature, contrasting with RNNs' slower sequential processing. Additionally, RNNs tend to excel in recognizing fine-grained actions, where temporal dependencies play a crucial role, while CNNs may better suit the recognition of gross motor actions based on spatial configurations. Considering factors like dataset size and hardware resources, the choice becomes more adaptable, contingent on the final model's scale. The size of the dataset and available computational resources for training become pivotal considerations, as different architectures might entail varying requirements. In summary, when recognizing actions reliant on temporal sequences, RNNs prove suitable for capturing the nuanced temporal dynamics within joint movements. In contrast, CNNs excel in identifying static spatial features and local patterns among joint positions. However, for comprehensive action recognition, leveraging both spatial and temporal relationships among joints, GCNs offer a beneficial approach when dealing with 3D skeletal data.

A possible in-practical solution can be also proposed to integrate not only one architecture but also a combination of them. This may make the final model absorb the advantages of each fundamental architecture. Furthermore, beyond the choice of deep architectures, the trajectory of 3D skeleton action recognition (SAR) navigation is a crucial consideration. Building upon our earlier discussions, we deduce that long-term action recognition, optimizing 3D-skeleton sequence representations, and achieving real-time operation remain significant open challenges. Moreover, annotating action labels for given 3D skeleton data remains

Table 1. The performance of the latest state-of-the-art 3D skeleton-based methods on NTU-RGB+D dataset

NTU-RGB+D dataset					
Rank	Paper	Year	Accuracy (C-View)	Accuracy (C-Subject)	Method
1	Wang et al. [55]	2023	98.7	94.8	Two-stream Transformer
2	Duan et al. [134]	2022	97.5	93.2	Dynamic group GCN
3	Liu et al. [135]	2023	96.8	92.8	Temporal decoupling GCN
4	Zhou et al. [56]	2022	96.5	92.9	Transformer
5	Chen et al. [136]	2021	96.8	92.4	Topology refinement GCN
6	Zeng et al. [137]	2021	96.7	91.6	Skeletal GCN
7	Liu et al. [103]	2020	96.2	91.5	Disentangling and unifying GCN
8	Ye et al. [138]	2020	96.0	91.5	Dynamic GCN
9	Shi et al. [139]	2019	96.1	89.9	Directed graph neural networks
10	Shi et al. [95]	2018	95.1	88.5	Two-stream adaptive GCN
11	Zhang et al. [140]	2018	95.0	89.2	LSTM-based RNN
12	Si et al. [141]	2019	95.0	89.2	AGC-LSTM(Joints&Part)
13	Hu et al. [142]	2018	94.9	89.1	Nonlocal S-T + frequency attention
14	Li et al. [101]	2019	94.2	86.8	GCN
15	Liang et al. [143]	2019	93.7	88.6	3S-CNN + multitask ensemble learning
16	Song et al. [144]	2019	93.5	85.9	Richly activated GCN
17	Zhang et al. [145]	2019	93.4	86.6	Semantics-guided GCN
18	Xie et al. [77]	2018	93.2	82.7	RNN+CNN+Attention

Table 2. The performance of the latest state-of-the-art 3D skeleton-based methods on NTU-RGB+D 120 dataset

NTU-RGB+D 120 dataset					
Rank	Paper	Year	Accuracy (C-Subject)	Accuracy (C-Setup)	Method
1	Wang et al. [55]	2023	92.0	93.8	Two-stream Transformer
2	Xu et al. [146]	2023	90.7	91.8	Language knowledge-assisted
3	Zhou et al. [56]	2022	89.9	91.3	Transformer
4	Duan et al. [134]	2022	89.6	91.3	Dynamic group GCN
5	Chen et al. [136]	2021	88.9	90.6	Topology refinement GCN
6	Chen et al. [147]	2021	88.2	89.3	Spatial-temporal GCN
7	Liu et al. [103]	2020	86.9	88.4	Disentangling and unifying GCN
8	Cheng et al. [148]	2020	85.9	87.6	Shift GCN
9	Caetano et al. [90]	2019	67.9	62.8	Tree structure + CNN
10	Caetano et al. [89]	2019	67.7	66.9	SkeleMotion
11	Liu et al. [149]	2018	64.6	66.9	Body pose evolution map
12	Ke et al. [150]	2018	62.2	61.8	Multitask CNN with RotClips
13	Liu et al. [151]	2017	61.2	63.3	Two-stream attention LSTM
14	Liu et al. [12]	2017	60.3	63.2	Skeleton visualization (single stream)
15	Jun et al. [152]	2019	59.9	62.4	Online+Dilated CNN
16	Ke et al. [153]	2017	58.4	57.9	Multitask learning CNN
17	Jun et al. [82]	2017	58.3	59.2	Global context-aware attention LSTM
18	Jun et al. [76]	2016	55.7	57.9	Spatiotemporal LSTM

exceptionally labor-intensive. Exploring avenues such as unsupervised or weakly-supervised strategies, along with zero-shot learning, may pave the way forward.

Conclusion

This paper presents an exploration of action recognition using 3D skeleton sequence data, employing 4 distinct neural network architectures. It underscores the concept of action recognition, highlights the advantages of skeleton data, and delves into the characteristics of various deep architectures. Unlike prior reviews, our study pioneers a data-driven approach, providing comprehensive insights into deep learning methodologies, encompassing the latest algorithms spanning RNN-based, CNN-based, GCN-based, and Transformer-based techniques. Specifically, our focus on RNN and CNN-based methods centers on addressing spatial-temporal information by leveraging skeleton data representations and intricately designed network architectures. In the case of GCN-based approaches, our emphasis lies in harnessing joint and bone correlations to their fullest extent. Furthermore, the burgeoning Transformer architecture has garnered significant attention, often employed in conjunction with other architectures for action recognition tasks. Our analysis reveals that a fundamental challenge across diverse learning structures lies in effectively extracting pertinent information from 3D skeleton data. The topology graph emerges as the most intuitive representation of human skeleton joints, a notion substantiated by the performance metrics observed in datasets like NTU-RGB+D. However, this does not negate the suitability of CNN or RNN-based methods for this task. On the contrary, the introduction of innovative strategies, such as multi-task learning, shows promise for substantial improvements, particularly in cross-view or cross-subject evaluation protocols. Nevertheless, achieving further accuracy enhancements on datasets like NTU-RGB+D presents increasing difficulty due to the already high-performance levels attained. Hence, redirecting focus toward more challenging datasets, such as the enhanced NTU-RGB+D 120 dataset, or exploring other fine-grained human action datasets becomes imperative. Finally, we delve into an exhaustive discussion on the selection of foundational deep architectures and explore potential future pathways in 3D skeleton-based action recognition.

Acknowledgments

Funding: This work was supported by the National Natural Science Foundation of China (No. 62203476) and the Natural Science Foundation of Shenzhen (No. JCYJ20230807120801002).
Author contributions: B. Ren has conducted the survey. B. Ren and M. Liu have written the draft. R. Ding and H. Liu have polished the draft.
Competing interests: The authors declare they have no competing interests.

Data Availability

Data are available upon reasonable request.

References

1. Wang Y, Kang H, Wu D, Yang W, Zhang L. Global and local spatio-temporal encoder for 3D human pose estimation. *IEEE Trans Multimedia*. 2023;1–11.
2. Tu Z, Liu Y, Zhang Y, Mu Q, Yuan J. Joint optimization of dark enhancement and action recognition in videos. *IEEE Trans Image Process*. 2023;32:3507–3520.
3. Zhang Y, Xu X, Zhao Y, Wen Y, Tang Z, Liu M. Facial prior guided micro-expression generation. *IEEE Trans Image Process*. 2024;33:525–540.
4. Wang X, Zhang W, Wang C, Gao Y, Liu M. Dynamic dense graph convolutional network for skeleton-based human motion prediction. *IEEE Trans Image Process*. 2024;33:1–15.
5. Liu H, Tian L, Liu M, Tang H. Sdm-bsm: A fusing depth scheme for human action recognition. Paper presented at: IEEE International Conference on Image Processing (ICIP); 2015 Sep 27–30; Quebec City, QC, Canada.
6. Liu M, He Q, Liu H. Fusing shape and motion matrices for view invariant action recognition using 3D skeletons. Paper presented at: IEEE International Conference on Image Processing (ICIP); 2017 Sep 17–20; Beijing, China.
7. Zhang FL, Cheng MM, Jia J, Hu SM. Imageadmixture: Putting together dissimilar objects from groups. *IEEE Trans Vis Comput Graph*. 2012;18(11):1849–1857.
8. Zhang FL, Wu X, Li RL, Wang J, Zheng ZH, Hu SM. Detecting and removing visual distractors for video aesthetic enhancement. *IEEE Trans Multimedia*. 2018;20(8):1987–1999.
9. Chen C, Liu M, Meng X, Xiao W, Ju Q. Refinedetlite: A lightweight one-stage object detection framework for cpu-only devices. Paper presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020 Jun 14–19; Seattle, WA.
10. Ren Z, Meng J, Yuan J, Zhang Z. Robust hand gesture recognition with kinect sensor. *IEEE Trans Image Process*. 2013;15(5):1110–1120.
11. Liu M, Meng F, Chen C, Wu S. Novel motion patterns matter for practical skeleton-based action recognition. In: AAAI Conference on Artificial Intelligence (AAAI); 2023 Feb 7, p.1701–1709.
12. Ren B, Tang H, Meng F, Ding R, Torr PH, Sebe N. Cloth interactive transformer for virtual try-on. *ACM Trans Multimed Comput Commun Appl*. 2023;20(4):1–20.
13. Liu M, Liu H, Chen C. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognit*. 2017;68:346–362.
14. Yang F, Wu Y, Sakti S, Nakamura S. Make skeleton-based action recognition model smaller, faster and better. Paper presented at: Proceedings of the ACM multimedia asia; 2019 Dec 15–18; Beijing, China.
15. Liu H, Tu J, Liu M. Two-stream 3d convolutional neural network for skeleton-based action recognition. arXiv. 2017. https://doi.org/10.48550/arXiv.1705.08106
16. Tang H, Ding L, Wu S, Ren B, Sebe N, Rota P. Deep unsupervised key frame extraction for efficient video classification. *ACM Trans Multimed Comput Commun Appl*. 2023;19(3):1–17.
17. Theodoridis T, Hu H. Action classification of 3d human models using dynamic anns for mobile robot surveillance. Paper presented at: 2007 IEEE International Conference on Robotics and Biomimetics (ROBIO); 2004 Dec 15–18, Sanya China.
18. Zhao M, Liu M, Ren B, Dai S, Sebe N. Modiff: Action-conditioned 3d motion generation with denoising diffusion probabilistic models. arXiv. 2023. https://doi.org/10.48550/arXiv.2301.03949

19. Wang Y, Tian Y, Zhu J, She H, Jiang Y, Jiang Z, Yokoi H. A hand gesture recognition strategy based on virtual dimension increase of EMG. *Cyborg Bionic Syst*. 2023;5:Article 0066.
20. Lin J, Gan C, Han S. Temporal shift module for efficient video understanding. arXiv. 2019. https://doi.org/10.48550/arXiv.1811.08383
21. Feichtenhofer C, Fan H, Malik J, He K. Slowfast networks for video recognition. Paper presented at: Proceedings of the IEEE/CVF international conference on computer vision. 2019; Oct–Nov 27–02; Seoul South Korea.
22. Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M. A closer look at spatiotemporal convolutions for action recognition. Paper presented at: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT.
23. Liu H, Ren B, Liu M, Ding R. Grouped temporal enhancement module for human action recognition. In: 2020 IEEE International Conference on Image Processing (ICIP); 2020 Oct 25–28; Abu Dhabi, UAE.
24. Thatipelli A, Narayan S, Khan S, Anwer RM, Khan FS, Ghanem B. Spatio-temporal relation modeling for few-shot action recognition. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA.
25. Xu C, Govindarajan LN, Zhang Y, Cheng L. Lie-x: Depth image based articulated object pose estimation, tracking, and action recognition on lie groups. *Int J Comput Vis*. 2017;123:454–478.
26. Baek S, Shi Z, Kawade M, Kim TK. Kinematic-layout-aware random forests for depth-based action recognition. arXiv. 2016. https://doi.org/10.48550/arXiv.1607.06972
27. Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Paper presented at: NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems; 2014 Dec 8; p. 568–576.
28. Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016 Jun 27–30; Las Vegas, NV.
29. Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, van Gool L *Temporal segment networks: Towards good practices for deep action recognition*. In: European conference on computer vision. (Springer. 2016). p. 20–36.
30. Gu Y, Sheng W, Ou Y, Liu M, Zhang S. Human action recognition with contextual constraints using a RGB-D sensor. Paper presented at: 2013 IEEE International Conference on Robotics and Biomimetics (ROBIO). 2013 Dec 12–14; Shenzhen, China.
31. Hu J-F, Zheng WS, Lai J, Zhang J. Jointly learning heterogeneous features for RGB-D activity recognition. *IEEE Trans Pattern Anal Mach Intell*. 2015;(11):5344–5352.
32. Johansson G. Visual perception of biological motion and a model for its analysis. *Percept psychophys*. 1973;14:201–211.
33. Liu C, Zhao M, Ren B, Liu M, Sebe N. Spatio-Temporal Graph Diffusion for Text-Driven Human Motion Generation. In: British Machine Vision Conference. 2023.
34. Zhang Z. Microsoft kinect sensor and its effect. *IEEE Multimedia*. 2012;19(2):4–10.
35. Chu X, Yang W, Ouyang W, Ma C, Yuille AL, Wang X. Multi-context attention for human pose estimation. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017 Jul 21–26; Honolulu, HI.
36. Yang W, Ouyang W, Li H, Wang X. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016 Jun 27–30; Las Vegas, NV.
37. Cao Z, Hidalgo G, Simon T, Wei SE, Sheikh Y. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. arXiv 2019. https://doi.org/10.48550/arXiv.1812.08008
38. Zhao Q, Zheng C, Liu M, Chen C. A Single 2D Pose with Context is Worth Hundreds for 3D Human Pose Estimation. In: Thirty-seventh Conference on Neural Information Processing Systems. 2023.
39. Si C, Chen W, Wang W, Wang L, Tan T. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. Paper presented at: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019 Jun 15–20; Long Beach, CA.
40. Vemulapalli R, Arrate F, Chellappa R. Human action recognition by representing 3d skeletons as points in a lie group. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2014 Jun 23–28; Columbus, OH.
41. Hussein ME, Torki M, Gowayyed MA, El-Saban M. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In: Twenty-third international joint conference on artificial intelligence. 2013.
42. Zhou Q, Yu S, Wu X, Gao Q, Li C, Xu Y. Hmms-based human action recognition for an intelligent household surveillance robot. Paper presented at: 2009 IEEE International Conference on Robotics and Biomimetics (ROBIO); 2009 Dec 19–23; Guilin, China.
43. Wang T, Liu H, Ding R, Li W, You Y, Li X. Interweaved Graph and Attention Network for 3D Human Pose Estimation. Paper presented at: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 Jun 4–10; Rhodes Island, Greece.
44. You Y, Liu H, Wang T, Li W, Ding R, Li X. Co-Evolution of Pose and Mesh for 3D Human Body Estimation from Video. Paper presented at: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 1–6; Paris, France.
45. Vemulapalli R, Chellapa R. Rolling rotations for recognizing human actions from 3d skeletal data. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016 Jun 37–30; Las Vegas, NV.
46. Wang L, Huynh DQ, Koniusz P. A comparative review of recent kinect-based action recognition algorithms. *IEEE Trans Image Process*. 2019;29:15–28.
47. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Commun ACM*. 2012;60(6):84–90.
48. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv. 2020. https://doi.org/10.48550/arXiv.2010.11929
49. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. arXiv. 2020. https://doi.org/10.48550/arXiv.2005.12872

50. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: Deformable Transformers for end-to-end object detection. arXiv. 2021. https://doi.org/10.48550/arXiv.2010.04159
51. Lev G, Sadeh G, Klein B, Wolf L. Rnn fisher vectors for action recognition and image annotation. Paper presented at: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016; 2016 Oct 8–16; Amsterdam Netherlands.
52. Cheron G, Laptev I, and Schmid C. P-cnn: Pose-based cnn features for action recognition. In: *Proceedings of the IEEE international conference on computer vision*. 2015:3218–26.
53. Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. Paper presented at: Proceedings of the AAAI conference on artificial intelligence. 2018.
54. Si C, Jing Y, Wang W, Wang L, Tan T. Skeleton-based action recognition with spatial reasoning and temporal stack learning. Paper presented at: Proceedings of the European conference on computer vision (ECCV). 2018
55. Wang L, Koniusz P. 3Mformer: Multi-order Multi-mode Transformer for Skeletal Action Recognition. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada.
56. Zhou Y, Cheng ZQ, Li C, Fan Y, Geng Y, Xie X, Keuper M. Hypergraph transformer for skeleton-based action recognition. arXiv. 2023. https://doi.org/10.48550/arXiv.2211.09590
57. Plizzari C, Cannici M, Matteucci M. Skeleton-based action recognition via spatial and temporal transformer networks. *Comput Vis Image Underst*. 2021;208–209:Article 103219.
58. Zhu X, Huang PY, Liang J, Melo CM de, Hauptmann AG. Stmt: A spatial-temporal mesh transformer for mocap-based action recognition. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada.
59. Bai D, Liu T, Han X, Yi H. Application research on optimization algorithm of sEMG gesture recognition based on light CNN+ LSTM model. *Cyborg Bionic Syst*. 2021;2021:Article 9794610.
60. You Y, Liu H, Li X, Li W, Wang T, Ding R. Gator: Graph-Aware Transformer with Motion-Disentangled Regression for Human Mesh Recovery from a 2D Pose. Paper presented at: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 Jun 4–10; Rhodes Island, Greece.
61. Poppe R. A survey on vision-based human action recognition. *Image Vis Comput*. 2010;28(6):976–990.
62. Weinland D, Ronfard R, Boyer E. A survey of vision-based methods for action representation segmentation and recognition. *Comput Vis Image Underst*. 2011;115(2):224–241.
63. Wu Z, Yao T, Fu Y, Jiang YG. Deep learning for video classification and captioning. Paper presented at: Frontiers of multimedia research; 2017 Dec 19; p. 3–29.
64. Herath S, Harandi M, Porikli F. Going deeper into action recognition: A survey. *Image Vis Comput*. 2017;60:4–21.
65. Lo Presti L, La Cascia M. 3D skeleton-based human action classification: A survey. *Pattern Recognit*. 53:130–147.
66. Ellis C, Masood SZ, Tappen MF, Laviola JJ Jr, Sukthankar R. Exploring the trade-off between accuracy and observational latency in;action recognition. *Int J Comput Vis*. 2013;101: 420–436.
67. Ofli F, Chaudhry R, Kurillo G, Vidal R, Bajcsy R. Berkeley MHAD: A comprehensive Multimodal Human Action Database. In: Applications of eComputer Vision. 2013.
68. Wang J, Liu Z, Wu Y, Yuan J. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In: Computer Vision and Pattern Recognition; 2012 Jun 16–21; Providence, RI.
69. Liu M, Meng F, Liang Y. Generalized pose decoupled network for unsupervised 3d skeleton sequence-based action representation learning. *Cyborg Bionic Syst*. 2022;2022:0002.
70. Sun Z, Ke Q, Rahmani H, Bennamoun M, Wang G, Liu J. Human action recognition from various data modalities: A review. *IEEE Trans Pattern Anal Mach Intell*. 2022;45(3): 3200–3225.
71. Zhang P, Xue J, Lan C, Zeng W, Gao Z, Zheng N. Adding attentiveness to the neurons in recurrent neural networks. Paper presented at: proceedings of the European conference on computer vision (ECCV). 2018. p. 135–151.
72. Wu D Shao L. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2014 Jun 23–28; Columbus OH.
73. Zhao R, Ali H, Van der Smagt P. Two-stream RNN/CNN for action recognition in 3D videos. Paper presented at: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2017 Sep 24–28; Vancouver BC, Canada.
74. Li W, Wen L, Chang MC, Nam Lim S, Lyu S. Adaptive RNN tree for large-scale humean action recognition. Paper presented at: Proceedings of the IEEE international conference on computer vision; 2017 Oct 22–29; Venice Italy.
75. Wang H, Wang L. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017 Jul 21–26; Honolulu, HI.
76. Liu J, Shahroudy A, Xu D, Wang G, Wang G. Spatio-temporal lstm with trust gates for 3d human action recognition. Paper presented at: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. Springer. 2016. p. 816–33.
77. Li C, Xie C, Zhang B, Han J, Zhen X, Chen J. Memory attention networks for skeleton-based action recognition. *IEEE Trans Neural Netw Lear Syst*. 2021;33:4800–4814.
78. Li L, Zheng W, Zhang Z, Huang Y, Wang L. Skeleton-based relational modeling for action recognition. arXiv. 2018. https://doi.org/10.48550/arXiv.1805.02556
79. Bradbury J, Merity S, Xiong C, Socher R. Quasi-Recurrent Neural Networks. In: International Conference on Learning Representations. 2016.
80. Lei T, Zhang Y, Artzi Y. Training rnns as fast as cnns. 2018.
81. Li S, Li W, Cook C, Zhu C, Gao Y. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018 Jun 18–23; Salt Lake City, UT.
82. Liu J, Wang G, Hu P, Duan LY, Kot AC. Global context-aware attention lstm networks for 3d action recognition. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017 Jul 21–26; Honolulu, HI.
83. Lee I, Kim D, Kang S, Lee S. Ensemble deep learning for skeleton-based action recognition using temporal sliding

lstm networks. Paper presented at: Proceedings of the IEEE international conference on computer vision; 2017 Oct 22–29; Venice Italy.
84. Ding Z, Wang P, Ogunbona PO, Li W. Investigation of different skeleton features for cnn-based 3d action recognition. Paper presented at: 2017 IEEE International conference on multimedia & expo workshops (ICMEW); 2017 Jul 10–14; Hong Kong, China.
85. Xu Y, Cheng J, Wang L, Xia H, Liu F, Tao D. Ensemble one-dimensional convolution neural networks for skeleton-based action recognition. *IEEE Signal Process Lett*. 2018;25: 1044–1048.
86. Wang P, Li W, Li C, Hou Y. Action Recognition Based on Joint Trajectory Maps with Convolutional Neural Networks. In: Acm on Multimedia Conference. 2016.
87. Bo L, Dai Y, Cheng X, Chen H, He M. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In: IEEE International Conference on Multimedia & Expo Workshops. 2017.
88. Li Y, Xia R, Liu X, Huang Q. Learning shape-motion representations from geometric algebra spatio-temporal model for skeleton-based action recognition. Paper presented at: 2019 IEEE international conference on multimedia and Expo (ICME); 2019 Jul 8–12; Shanghai, China.
89. Caetano C, Sena J, Br´emond F, Dos Santos JA, and Schwartz WR. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. Paper presented at: 2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS); 2019 Sep 18–21; Taipae, Taiwan.
90. Caetano C, Br´emond F, Schwartz WR. Skeleton image representation for 3d action recognition based on tree structure and reference joints. Paper presented at: 2019 32nd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI). 2019:16–23.
91. Chao L, Zhong Q, Di X, Pu S. Co-occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation. 2018.
92. Soo Kim T, Reiter A. Interpretable 3d human action analysis with temporal convolutional networks. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2017 Jul 21–26; Honolulu, HI.
93. Lea C, Flynn MD, Vidal R, Reiter A, Hager GD. Temporal convolutional networks for action segmentation and detection. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI.
94. Ruiz AH, Porzi L, Bul`o SR, and Moreno-Noguer F. 3D CNNs on Distance Matrices for Human Action Recognition. Paper presented at: MM '17: Proceedings of the 25th ACM international conference on Multimedia; 2024 Oct–Nov 28–01; Melbourne, VIC, Australia.
95. Shi L, Zhang Y, Cheng J, Lu H. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019 June 15–20; Long Beach, CA.
96. 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; 2020 Jun 13–19; Seattle, WA.
97. Cheng K, Zhang Y, Cao C, Shi L, Cheng J, Lu H. Decoupling gcn with dropgraph module for skeleton-based action recognition. Paper presented at: Computer Vision–ECCV 2020: 16th European Conference 2020, Proceedings, Part XXIV 16. 2020 Aug 23–28. Glasgow, UK.
98. Chi Hg, Ha MH, Chi S, Lee SW, Huang Q, Ramani K. Infogcn: Representation learning for human skeleton-based action recognition. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA.
99. Duan H, Zhao Y, Chen K, Lin D, Dai B. Revisiting skeleton-based action recognition. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA.
100. Zhou H, Liu Q, Wang Y. Learning discriminative representations for skeleton based action recognition. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. p. 10608–10617.
101. 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. 2019. p. 3595–3603.
102. Lei S, Yifan Z, Jian C, Hanqing L. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In: IEEE Conference on Computer Vision & Pattern Recognition. 2019.
103. Liu Z, Zhang H, Chen Z, Wang Z, Ouyang W. Disentangling and unifying graph convolutions for skeleton-based action recognition. Paper presented at: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020 Jun 13–19; Seattle, WA.
104. Wang X, Dai Y, Gao L, Song J. Skeleton-based action recognition via adaptive crossform learning. In: Proceedings of the 30th ACM International Conference on Multimedia. 2022. p. 1670–1678.
105. Hao X, Li J, Guo Y, Jiang T, Yu M. Hypergraph neural network for skeleton-based action recognition. *IEEE Trans Image Process*. 2021;30:2263–2275.
106. Lee J, Lee M, Lee D, Lee S. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. Paperr presented at: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 1–6; Paris, France.
107. Yang H, Yan D, Zhang L, Sun Y, Li D, Maybank SJ. Feedback graph convolutional network for skeleton-based action recognition. *IEEE Trans Image Process*. 2021;31:164–175.
108. Bian C, Feng W, Wan L, Wang S. Structural knowledge distillation for efficient skeleton-based action recognition. *IEEE Trans Image Process*. 2021;30:2963–2976.
109. Fang Z, Zhang X, Cao T, Zheng Y, Sun M. Spatial-temporal slowfast graph convolutional network r skeleton-based action recognition. *IET Comput Vis*. 2022;16:205–217.
110. Vaswani A, Shazeer N, Parmar N, Uskorelt J, Jones L, Gomez AN, Kaise L, Polosukhin I. Attention is all you need. Paper presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. p. 6000–6010.
111. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: A survey. *ACM Comput Surveys*. 2022;54(10):1–41.

112. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jegou H. Training data-efficient image transformers & distillation through attention. Paper presented at: International Conference on Machine Learning (ICML). 2021.
113. Ren B, Liu Y, Song Y, Bi W, Cucchiara Rita, Sebe N, Wang W. Masked Jigsaw Puzzle: A Versatile Position Embedding for Vision Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. p. 20382–20391.
114. Ye L, Rochan M, Liu Z, Wang Y. Cross-modal self-attention network for referring image segmentation. Paper presented at: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. p. 10502–10511.
115. Chen H, Wang Y, Guo T, Xu C, Deng Y, Liu Z, Ma S, Xu C, Xu C, Gao W. Pre-trained image processing transformer. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 June 20–25; Nashville, TN.
116. Li Y, Fan Y, Xiang X, Demandoix D, Ranjan R, Timofte R, Gool Van L. Efficient and explicit modelling of image hierarchies for image restoration. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, Canada.
117. Mei G, Poiesi F, Saltori C, Zhang J, Ricci E, Sebe N. Overlap-guided gaussian mixture models for point cloud registration. Paper presented at: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2023 Jan 2–7; Waikoloa, HI.
118. Huang X, Mei G, Zhang J. Cross-source point cloud registration: Challenges, progress and prospects. *Neurocomputing*. 2023;548:126383.
119. Wang W, Mei G, Ren B, Huang X, Poiesi F, Gool Van L, Sebe N, Lepri B. Zero-shot point cloud registration. arXiv. 2023. https://doi.org/10.48550/arXiv.2312.03032
120. Cho S, Maqbool M, Liu F, Foroosh H. Self-attention network for skeleton-based human action recognition. arXiv. 2019. https://doi.org/10.48550/arXiv.1912.08435
121. Shi L, Zhang Y, Cheng J, Lu H. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In: Proceedings of the Asian Conference on Computer Vision. 2020.
122. Plizzari C, Cannici M, Matteucci M. Spatial temporal transformer network for skeleton-based action recognition. In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, Proceedings, Part III. Springer. 2021 Jan 10–15; p. 694–701.
123. Ibh M, Grasshof S, Witzner D, Madeleine P. TemPose: A New Skeleton-Based Transformer Model Designed for Fine-Grained Motion Recognition in Badminton. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023 Jun 17–24; Vancouver, BC, Canada.
124. Zhu W, Ma X, Liu Z, Liu L, Wu W, Wang Y. Motionbert: A unified perspective on learning human motion representations. Paper presented at: Proceedings of the IEEE/CVF International Conference on Computer Vision 2023 Oct 1–6; Paris, France.
125. Xiang W, Li C, Zhou Y, Wang B, Zhang L. Generative Action Description Prompts for Skeleton-based Action Recognition. Paper presented at: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 1–6; Paris, France.
126. Yuan L, He Z, Wang Q, Xu L, Ma X. Spatial transformer network with transfer learning for small-scale fine-grained skeleton-based tai chi action recognition. Paper presented at: *IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society*. IEEE. 2022. p. 1–6.
127. Zhang J, Jia Y, Xie W, Tu Z. Zoom transformer for skeleton-based group activity recognition. *IEEE Trans Circuits Syst Video Technol*. 2022;32(12):8646–8659.
128. Gao Z, Wang P, Lv P, Jiang Z, Liu Q, Wang P, Xu M, Li W. Focal and global spatial-temporal transformer for skeleton-based action recognition. Paper presented at: Proceedings of the Asian Conference on Computer Vision. 2022. p. 382–398.
129. Li W, Zhang Z, Liu Z. Action recognition based on a bag of 3D points. Paper presented at: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops; 2010 Jun 13–18; San Francisco, CA.
130. Oreifej O, Liu Z. HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences. Paper presented at: IEEE Conference on Computer Vision & Pattern Recognition; 2013 Jun 23–28; Portland, OR.
131. Shahroudy A, Liu J, Ng T-T, Wang G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; p. 1010–1019.
132. Liu J, Shahroudy A, Perez M, Wang G, Duan LY, Kot AC. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans Pattern Anal Mach Intell*. 2019;42(10):2684–2701.
133. Wang J, Nie X, Xia Y, Wu Y, and Zhu SC. Cross-view action modeling, learning and recognition In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 2649–56.
134. Duan H, Wang J, Chen K, Lin D. DG-STGCN: dynamic spatial-temporal modeling for skeleton-based action recognition. arXiv. 2022. https://doi.org/10.48550/arXiv.2210.05895
135. Liu J, Wang X, Wang C, Gao Y, Liu M. Temporal Decoupling Graph Convolutional Network for Skeleton-based Gesture Recognition. *IEEE Trans Multimedia*. 2023;26:811–823.
136. Chen Y, Zhang Z, Yuan C, Li B, Deng Y, Hu W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. Paper presented at: Proceedings of the IEEE/CVF international conference on computer vision. 2021 Oct 10–17; Montreal, QC, Canada.
137. Zeng A, Sun X, Yang L, Zhao N, Liu M, Xu Q. Learning skeletal graph neural networks for hard 3d pose estimation. Paper presented at: Proceedings of the IEEE/CVF international conference on computer vision. 2021 Oct 10–17; Montreal, QC, Canada.
138. Ye F, Pu S, Zhong Q, Li C, Xie D, Tang H. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. Paper presented at: Proceedings of the 28th ACM international conference on multimedia; 2020 Oct 12–16; WA, Seattle.
139. Shi L, Zhang Y, Cheng J, Lu H. Skeleton-Based Action Recognition with Directed Graph Neural Networks. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA.
140. Zhang P, Lan C, Xing J, Zeng W, Xue J, Zheng N. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE Trans Pattern Anal Mach Intell*. 2019;41(8):1963–1978.

141. Si C, Chen W, Wang W, Wang L, Tan T. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA.
142. Hu G, Cui B, Yu S. Skeleton-based action recognition with synchronous local and non-local spatio-temporal learning and frequency attention. Paper presented at: 2019 IEEE International Conference on Multimedia and Expo (ICME), 2019 Jul 8–12; Shanghai, China.
143. Liang D, Fan G, Lin G, Chen W, Pan X, Zhu H. Three-stream convolutional neural network with multi-task and ensemble learning for 3d action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops; 2019 Jun 16–17; Long Beach, CA.
144. Song YF, Zhang Z, Wang L. Richly activated graph convolutional network for action recognition with incomplete skeletons. arXiv. 2019. https://doi.org/10.48550/arXiv.1905.06774
145. Zhang P, Lan C, Zeng W, Xue J, Zheng N. Semantics-guided neural networks for efficient skeleton-based human action recognition. arXiv. 2020. https://doi.org/10.48550/arXiv.1904.01189
146. Xu H, Gao Y, Hui Z, Li J, Gao X. Language knowledge-assisted representation learning for skeleton-based action recognition. arXiv. 2023. https://doi.org/10.48550/arXiv.2305.12398
147. Chen T, Zhou D, Wang J, Wang S, Guan Y, He X, Ding E. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. Paper presented at: Proceedings of the 29th ACM international conference on multimedia; 2021 Oct 20–24; Virtual Event, China.
148. Cheng K, Zhang Y, He X, Chen W, Cheng J, Lu H. Skeleton-based action recognition with shift graph convolutional network. Paper presented at: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020 Jun 13–19; Seattle, WA.
149. Liu M, Yuan J. Recognizing human actions as the evolution of pose estimation maps. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT.
150. Ke Q, Bennamoun M, An S, Sohel F, Boussaid F. Learning clip representations for skeleton-based 3D action recognition. *IEEE Trans Image Process*. 2018;27(6):2842–2855.
151. Liu J, Wang G, Duan LY, Abdiyeva K, Kot AC. Skeleton-based human action recognition with global context-aware attention LSTM networks. *IEEE Trans Image Process*. 2017;27(4):1586–1599.
152. Liu J, Shahroudy A, Wang G, Duan LY, Chichung AK. Skeleton-based online action prediction using scale selection network. *IEEE Trans Pattern Anal Mach Intell*. 2019;42(6):1453–1467.
153. Ke Q, Bennamoun M, An S, Sohel F, Boussaid F. A new representation of skeleton sequences for 3d action recognition. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 3288–3297.