UNIVERSAL WISER
PUBLISHER

Remarkable Skeleton Based Human Action Recognition

Sushma Jaiswal[1], **Tarun Jaiswal**[2*]

[1]Guru Ghasidas Central University, Bilaspur (C.G.), India
[2]Department of Computer Applications, NIT Raipur, Raipur, India
E-mail: tjaiswal_1207@yahoo.com

Abstract: Skeleton-based human-action-recognition (SBHAR) has wide applications in cognitive science and automatic surveillance. However, the most challenging and crucial task of the skeleton-based human-action-recognition (SBHAR) is a significant view variation while capturing the data. In this area, a significant amount of satisfactory work has already been done, which include the Red Green Blue (RGB) data method. The performance of the SBHAR is also affected by the various factors such as video frame setting, view variations in motion, different backgrounds and inter-personal differences. In this survey, we explicitly address these challenges and provide a complete overview of advancement in this field. The deep learning method has been used in this field for a long time, but so far, no research has fully demonstrated its usefulness. In this paper, we first highlight the need for action recognition and significance of 3D skeleton data and finally, we survey the largest 3D skeleton dataset, i.e. NTU-RGB+D and its new version NTU-RGB+D 120.

Keywords: skeleton based human action recognition, CNN, SVM-GNN, machine learning, MS-G3D, DGNN

Abbreviations

CNN	Convolution Neural Network
SVM-GNN	Support Vector Machine-Graph Neural Networks
MS-G3D	Multi Scale Spatial-Temporal Graph Convolutional Module
DGNN	Directed Graph Neural Networks
PKU-MMD or PKUMMD	Peking University-Multi modality datasets
NTU-RGB+D	Nanyang Technological University's Red Blue Green and Depth-information
NTU-RGB+D 120	Nanyang Technological University's Red Blue Green and Depth-information 120
SBUKinect	Stony Brook University Kinect
HDM05	Hochschule der Medien (HDM) 05 (developed by Max-Planck-Institut Informatik)
SYSU-3D	Sun Yat-Sen University-3D
UT-Kinect	University of Texas-Kinect
Human3.6M	Human-poses 3.6 Million (developed by/vision.imar.ro)
CMU Mocap	Carnegie Mellon University Motion Capture Database
NTU-60s	Nanyang Technological University-60s
NTU-120	Nanyang Technological University-120
SYSU	Sun Yat-Sen University

1. Introduction

The SBHAR is an extremely challenging and most crucial task in the filed fo machine learning. The recognition of human activities from the various image frames enable its deployment in the several filed such as identification of doubtful activities etc. The main goal of this paper is to describe the various human activity recognition approaches. The most of the authors differentiate the human action recognition approaches in 02 primarily groups namely top-down and bottom-up approaches, and other divided it in another 02 possible ways, these are *unimodal* and *multimodal* action recognition approaches.

Unimodal approach works on a single image, these are intergalactic time, stochastic, rule-based, & shape-based approaches. An intergalactic time approach describes the collection of temporal features [1, 2] or trajectories [3, 4] whereas

DOI: https://doi.org/10.37256/aie.122020562

stochastic approach comprise the action through statistical facsimiles [5-7]. The Rule-based approach follows custom protocol and designates human actions [8, 9]. The shape-based approach signifies the actions when demonstrating the arbitrary view of body parts [10, 11].

A *multimodal approach* integrates and builds the features of numerous frameworks [12], according to the literature, multimodal is divided into 03 main groups, and these groups are affective, behavioral, and social networking groups. The effective approach signifies the action conferring, the expressive skills and the moving position of the subject [13, 14]. The objective of behavioral approach is to diagnose behavioral features, non-verbal multimodal signals, likewise gestures, facial expressions, and auditory cues [15, 16] and last approach, i.e. social networking approach describes the prototypical features and actions of persons in numerous branches of person-to-person interfaces in communal events of gestures, i.e. body movement and dialogue [17, 18].There are many avenues for action recognition, such as-global, local, and depth-based [19].

In *Global Demonstrations*, works on original video or images. Global Demonstrations draw out global features and encrypt them from entire features. In this approach, image is localized and isolated by background subtraction approaches and by which we can create the region of interest or silhouettes. Global demonstration tactics typically suggested the past mechanism and progressively old-fashioned method [19].

In *local demonstration* method, action images are identified as a group of local features. Local demonstration approach concentrates on local blotches and shows interest point. They are beneficial besides noise and partial occlusions as of global demonstration. Local features are then usually collected with the bag-of-visual-words (BoVW). Local demonstration method is the common channel of existing state-of-the-art local demonstration approaches [19].

In *Depth-Based demonstrations* method, an accessible depth maps and the skeletal data energetically is subsidized to the machine learning communal. These 02 features and their imitative features also produce wide attention to resolve-HAR difficulties using depth-based elucidations, substituting conservative RGB-based approaches, or substitute supplements to enrich the RGB-based techniques [19].

In this investigation, we summarize numerous modern mechanisms and present an innovative survey of investigation on SBHAR methods. The action classification methods are summarized into strategies based on remarkable skeleton-based human action recognition. These approaches cover diverse varieties of data.

2. Related work

Remarkable SBHAR has a wide variety of uses, such as intellectual video reconnaissance and environmental home observation [20, 21], video cell and retrieval [22, 23], intellectual human-to-machine interaction [24, 25], and human characteristics recognition [26]. SBHAR insurances numerous research subjects in computer/machine visualisation and human detection/classification and this method is also useful for human posture assessment, tracing of human, and examination and indulgent of time-series information. It is a challenging work in the arena of computer/machine vision and machine/deep learning. Currently, there are several significant difficulties in SBHAR that endure mysteriously. In this section, we provide extensive work related to skeleton human action recognition.

Synergetic Graph Neural Network (Sym-GNN). The authors proposed [27], the framework known as the synergetic model is used to describe the 02 distinct levels graphs for imprisonment relation of joints and parts of the human body. The authors also presented the synergetic graph neural network, which comprises a pillar of action-recognition and motion-based prediction head, that are tied together to each other. The authors also used multi-branch & multi-scale graph Neural Network (NN) to draw out features of spatial and temporal. The multi-scale graph NN, makes joint scale graph that includes actionable-graph, action-based relations, and structural graphs. Physical constraints and part scale graph are incorporated into high-level relations and body-joints to form explicit amounts. Authors also performed an experimental evaluation on four datasets (NTU-RGB+D, Kinetics, Human3.6M, and CMU Mocap), and demonstrated that the proposed methods work better as compared to other prevalent methods.

DenseInd Recurrent Neural Network (RNN). Recurrent connection articulated as Hadamard product, denote to Independently Recurrent Neural Network (IndRNN) [28], wherever the same layer neurons are free from each other and integrated with layers. Through IndRNN the gradient vanishing and exploding glitches are elucidated via normal adoptable recurrent weights. It also works with non-saturated activation function Rectified Linear Unit (Relu). The authors worked on basic stacked, deeper, residual and densely connected IndRNN. IndRNN is 10X faster than the NVIDIA CUDA® Deep Neural Network Library LSTM (CUDNN-LSTM). The authors performed an experimental evaluation and showed that IndRNN outperformed than the RNN and Long-Short Term Memory (LSTM).

Graph Regression Based-GCN (GR-GCN). The authors investigated graph-regression based Graph Convolutional Network (GCN), Known as GR-GCN [29]. This method is designed for SBHAR and indicates the spatio-temporal disparity in the data. Graph regression is trained and learned the fundamental graph form manifold interpretations. The optimized graph is connected to each neighboring joints either strongly or weakly. It also joins prevalent and consecutive frames and nourishes the optimized graph on GCN with the pixels values of the skeleton sequence for feature learning. The experimental evaluation shows the proposed method outperformed than the other state-of-the-art approaches.

Residual frequency attention (rFA). The authors focused on the residual frequency attention (rFA) [30] and organized them (chunk) into discriminative arrangements. The authors also investigated a soft-margin-focal-loss (SMFL) for optimization purposes. The proposed method performed superior to other prevalent methods.

MS-AAGCN. The authors [31] introduced Multi Stream-Attention-enhanced Adaptive Graph Convolutional Neural-network (MS-AAGCN) for SBHAR.The model obtains the input parameter via graph topology. Data-driven tactics enhance the elasticity of the model for graph built up. The considered adaptive graph convolutional layer also increase flexibility via spatial-temporal channel attention phase. The joints and bones data are connected through motion information and demonstrated in a multi-stream framework which enhances the accurateness. Authors judge the performance through 02 datasets and finally summed up, the proposed methods work more rapidly than the other state-of-the-art method with a significant margin.

Survey. The authors [32] used 10 topical Kinect-based processes for cross substance and cross assessment action-recognition with 06 standard datasets. The authors have taken some of the essential features for improvement of the results. The experimental evaluation demonstrated that most of the methods are superior on Cross substance other than cross assessment action recognition. In contrast, skeleton constructed features are most robust for cross assessment action recognition than depth constructed structures.

MS-G3D. The manifold weighbridge aggregation framework unravels the rank of knobs in distinct localities for high operative term modelling [33]. The spatial-temporal-graph is used for scheme integration. The authors also introduced an influential feature extractor, namely MS-G3D. The proposed method overtakes prevalent state-of-the-art approaches.

SGN. The authors considered modest operative semantics-guided-neural-network (SGN) [34] for SBHAR. Through this, authors improved the feature representation via joints networks. The authors used 02 hierarchy of joints, namely joint level and frame level. The first one worked on Associations of joints in a similar edge and second phase for additions of edges by taking the joints in a similar edge as an entire. SGN achieves the state-of-the-art performing results on the NTU-60, NTU-120, & SYSU standard datasets.

GCN-NAS. The authors worked on Neural-Architecture-Search (NAS) [35] and discovered the automated GCN framework for SBHAR. This approach improves the search-space via given multiple dynamic graph segments while discovering the spatial temporal association concerning nodes. The authors also introduced multiple-hop components and improved the representative volume, which comes during first order approximation. The outcome frameworks demonstrate the efficiency of the higher-order approximation and dynamic graphs. Extensive experimental evaluations on 02 vast scale datasets demonstrate the efficiency of the proposed technique and show superior performance over state-of-the-art approaches.

NTU-RGB+D 120. The authors studied [36] large-volume datasets for RGB+D action-recognition, which is under the umbrella of 106 different subjects, 114 thousand audiovisual samples and 08 million casings. The author performed an experimental evaluation on existing 3D activity analysis and examined that Deep Learning (DL) grounded the method for 3D action-recognition gives the benefit. The authors also considered the one-shot 3D action-recognition application on datasets. For this purpose, Action Part Semantic Relevance-aware (APSR) scheme is applied. The authors trust the APSR for depth construction and RGB+D based action-recognition.

View Adaptation Neural Network (VA-NN). The authors [37] introduced 2-way view adaptation NN, namely View Adaptation-Recurrent Neural Network (VA-RNN) and View Adaptation-Convolution Neural Network (VA-CNN) and for this purpose RNN based LSTM & CNN is used. The proposed method controls the viewpoint-transfer to endwise classification network. The method uses the 02 stream structure and generates the final forecasting. The Wide-ranging investigational evaluation on 05 interesting benchmarks determines the efficiency of the suggested view-adaptive net and higher performance over state-of-the-art tactics.

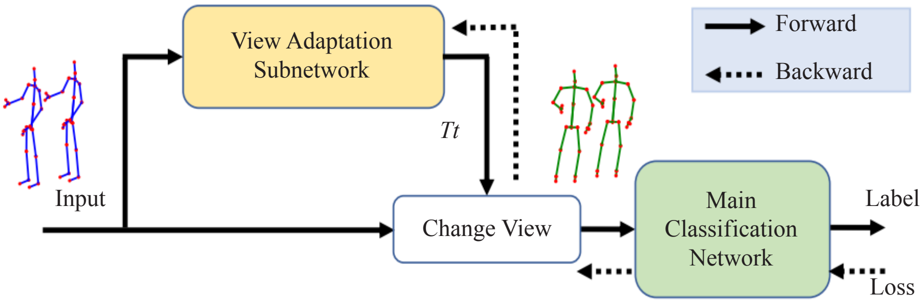

Figure 1. Construction of endwise interpretation adaptive-neural-network, contains 02 parts adaption-subnet and main-classification net [37]

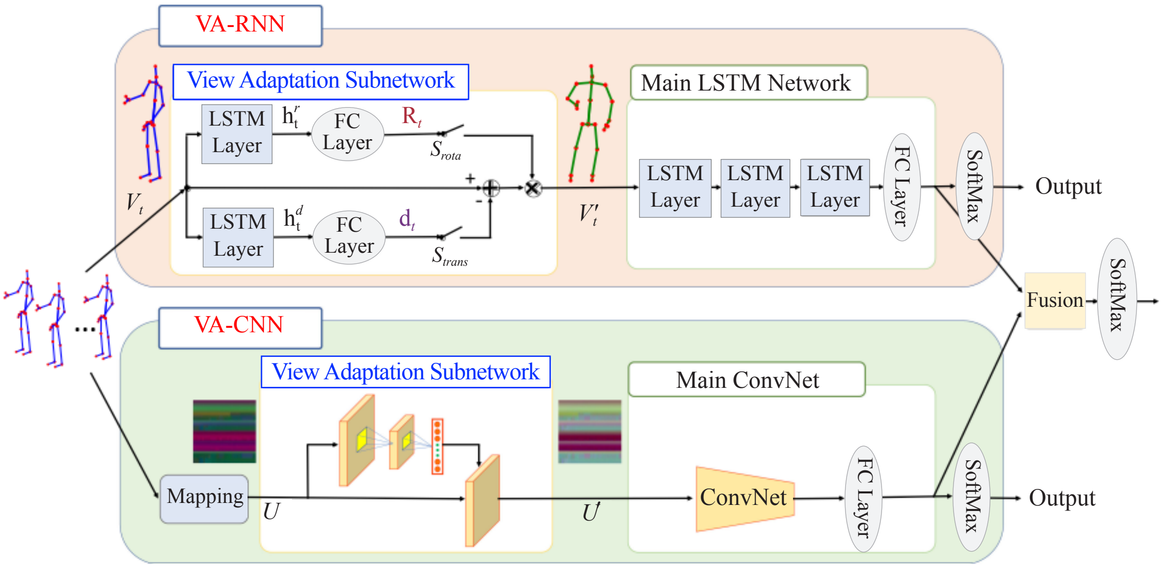

Figure 2. The structural design of the view-adaptive-neural-networks [37]

Bayesian Graph Convolutional LSTM. The authors proposed [38] a vigorous process that produces the motion design for action recognition from skeleton data. There are 03 important elements that affect the motion pattern known as body-joints, temporal enslavements of body, and variation among issues in action accomplishment. Graph convolution extracts the exemplification feature from pose data. The entire process is enhanced with a Bayesian scheme. The authors proved the advantages of this proposed structure by applying distinct benchmark datasets with action recognition under various generalization conditions.

Spatial Residual Layer and Dense Connection Block Enhanced Spatial Temporal Graph Convolutional Network (2s-SDGCN). The authors proposed [39] the 03 facets, 1. Spatial-temporal graph convolution of cross-province finds out the accurate instructive information which decreases the regular iterations by mixed feature (residual spatial layer). 2. By using dense links to earn overall information. 3. Mixed the first + second facet to make a spatial-temporal-graph-convolutional-network (ST-GCN) or in short SDGCN. The authors performed the extensive experimental evaluation and summed up the proposed methods outperformed quantitatively as well as qualitatively than the other state-of-the-art method.

Directed Graph Neural Networks (DGNN). The construction among joints and bones shown by a direct Directed Acyclic Graph (DAG) is established on the kinematics dependency [40]. The data of joints, bones and their associations produce prediction by a new directed graph. The proposed method achieves a substantial upgradation in performance compared to the conventional methods. The authors evaluate the method quantitatively and qualitatively and demonstrate its efficiency.

2s-AGCN. 02 stream adaptive graph convolutional network (2s-AGCN) for action recognition (skeleton) was proposed by the authors [41], and that network continuously learns through the BP algorithm. The proposed method is used together for first and second-order data concurrently, which produces high recognition accurateness. The practical evaluation shows that the performance of the proposed model beats the state-of-the-art methods.

AS-GCN. Encoder decoder framework (A-link inference module) find out the action-based underlying dependencies for structural links [42]. The authors improved the prevalent skeleton graph to signify high order enslavements. By combining two connection, the authors represent actional-structural-graph-convolution-network (AS-GCN) through learn spatial and temporal features. The investigated method AS-GCN achieve significant expansion compared to the other state-of-the-art approaches.

AGC-LSTM. The new method Attention enhanced Graph Convolutional LSTM Network (AGC-LSTM) is introduced for action-recognition [43]. The introduced methods select the discriminative topographies in the altitudinal arrangement and progressive dynamics. This method discovers the co-occurrence connection among altitudinal and progressive provinces. The assessment outcome determines the efficiency of the proposed method and demonstrates that the suggested approach beats the state-of-the-art methods on 02 benchmark datasets.

Motif-Based Spatial Temporal Graph Convolutional Networks (Motif-STGCN). The authors employed Motif-based-graph-convolution [44] to encrypt ranked altitudinal scheme, and a progressive dense segment confined temporal data from distinct varieties of skeleton structures. The proposed methods achieve enhancements over state-of-the-art schemes.

Richly Activated Graph Convolutional Network. The novel method multi-stream graph-convolutional-network (GCN) proposed by the authors [45]. In the proposed method, every part of the net is capable of learning from formerly non-activated linkages. Class activation map (CAM) achieved by the preceding part, and richly-activated-GCN (RA-GCN), and RA-GCN enhance the strength of the framework.

TSRJI. The authors considered Tree Structure Reference Joints Image (TSRJI) [46]. The investigated method get benefit from the integrated reference joints + tree structure. Depth first-order algorithm collects the most imperative spatial relations. Investigational outcomes determine the efficiency of the suggested demonstration for 3D action recognition over 02 datasets.

Skeleton Joint Sequences Based on Motion (SkeleMotion). The investigational method encrypts the temporal dynamics by calculating the degree and angular data of the joints [47]. Distinct levels of temporal work measure the angular data to collect more temporal dynamics in order to be able to produce the filtered boisterous angular data and joints information. Experimental evaluation outcomes demonstrate the efficiency of the proposed method on NTU RGB+D-120 dataset.

Self-Attention Network. There are 03 essential forms of the self-attention-network (SAN) these are SAN-Version 1 (V1), SAN-Version 2 (V2) & SAN-Version 3 (V3) [48]. The importance of the investigations are-03 SAN version, which is used for important correlations to form deep semantics, and accumulated Temporal Segment Network (TSN) thru SAN deviations by which performance can be measured. The most important aspect is that each segment is connected to each other. The suggested method beat state-of-the-art approaches.

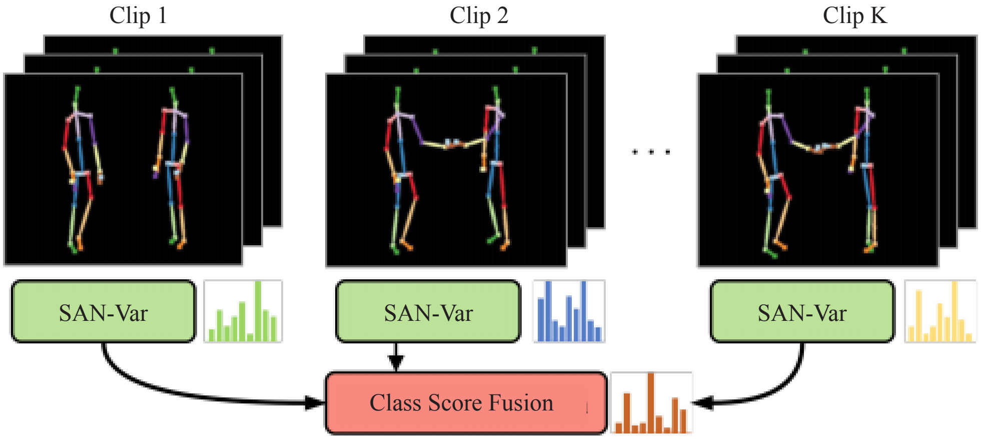

Figure 3. Class-score-fusion structure [48]

Beyond Joints. The authors [49] worked on 03 primary factors such as joints, edges and surfaces. There are 03 incoming vectors used for RNN based network. For the robust representation of action recognition, authors used 02 layers, first viewpoint transformation and second is, temporal dropout layer in RNN network. For action detection, sequence wise action classification is transferred to multi-scale sliding window algorithm. The experimental evaluation is applied to huge 3D action recognition benchmark dataset. The investigated method beat prevalent state-of-the-art approaches.

DPRL. In this paper [50], the authors proposed a deep-progressive-reinforcement learning (DPRL) technique for action-recognition and skeleton-based visualization. There are 02 significant task, quality and relationships among the designated structure. The graph vertices & edges easily represent the connections between bones, so the authors considered graph-based convolutional neural network. The estimated method achieves very competitive performance on 03 extensively benchmark.

SR-TSL. The authors integrated the 02 networks, spatial-reasoning-network (SRN) and a temporal-stack-learning-network (TSLN) for altitudinal reasoning and progressive stack learning (SR-TSL) [51]. The spatial reasoning network can take the altitudinal structural data thru a residual graph neural network. In contrast, temporal stack learning network can produce the progressive dynamics thru an arrangement of multiple skip-clip LSTMs. The assessment consequences demonstrated that the investigated method achieves much improved results than state-of-the-art approaches.

Hierarchical Co-occurrence Network (HCN). The authors discovered endwise convolutional co-occurrence feature learning scheme, to learn ranked methodology via co-occurrence features to integrate distinct altitudes of contextual data [52]. Each level of information is encrypted independently, whereas Spatial and temporal data are collected together. The authors represented the global spatial aggregation framework used for learning specific joints co-occurrence. The extensive experiments indicate that the investigated method rapidly beat other state-of-the-arts methods of human action-recognition and detection standards, namely NTURGB+D, SBUKinect Interaction and PKUMMD.

MAN. For lighten multifaceted variants, the authors [53] introduced a temporal-then-spatial recalibration framework which is associated with Memory-Attention-Networks (MANs). This framework also includes the Temporal-Attention-Recalibration-Module (TARM) and a Spatio-Temporal-Convolution-Module (STCM). The residual learning module includes the reconstructed progressive structures of a skeleton. The STCM behaves like the joint standardization frames which influence the CNN. TARM and STCM are tied together in one network that can be made custom. MANs expressively enhanced the characterization of action-recognition and produced the improved outcomes on 04 different benchmarks, namely NTURGB+D, HDM-05, SYSU3D and UT-Kinect.

ST-GCN. The authors focused on the novel method known as Spatial-Temporal-Graph-Convolutional-Networks (ST-GCN) for SBHAR [54]. The ST-GCN automated learning capability of both spatial & temporal decorations form the data. This construction gives not only superior communicative control but also tougher generality proficiency. On 02 huge datasets, Kinetics and NTU-RGBD, it attains substantial enhancements over conventional approaches.

STGC. The authors focused on spatio-temporal-graph-convolution (STGC) method [55] for accumulating the confined convolutional purifier and increase the learning capability. To encrypt the dynamic graphs, multi-scale local graph convolution filter is made up by matrices of signal mappings. The authors hypothetical demonstrated the immovability of STGC. The authors also used iterative techniques so that it can be arranged into multilayer construction. Extensive experiments on 04 benchmark datasets, showed the enhancement and efficiency of the introduced prototypical.

PB-GCN. The authors demonstrated [56] a part-based graph convolutional network (PB-GCN) scheme, the scheme is distributed in 04 ways, and capable of learning a recognition model. PB-GCN enhance the performance of recognition when compared to the whole skeleton graph. By using neighboring coordinates method performance is significantly improved. The anticipated technique achieved better efficiency on 02 thought-provoking benchmark datasets, i.e. NTURGB+D and HDM-05.

Fine-to-coarse (F2C) CNN framework. 3D skeleton sequence for human action recognition introduced by the authors. In this method, skeleton sequences are broken into different temporal fragments. Simultaneously temporal and spatial features of a skeleton sequence are extracted by fine-to-coarse (F2C) CNN framework [57]. The authors applied the methods over 02 datasets, namely NTURGB+D and SBU Kinect Interaction and reached 79.6 & 84.6 % accurateness. The suggested approach expressively expands the accuracy of the actions in 02-person connections.

VS-CNN. The authors discovered the View-guided Skeleton CNN (VS-CNN) to solve the arbitrary-view action recognition [58]. For this purpose authors, collected a large volume dataset that contains videos, depth and skeleton range. The authors collected action based example pictures in eight immovable views and different frames which insured 360-degree vision angles and 118 humans with 40 actions and 25600 video examples. It provides multiview, arbitrary view & cross-view action examination. The extensive experiment outcomes demonstrate that the VS-CNN achieves higher performance.

Iterative method for joint model. The authors [59] proposed 03 phases iterative method for the joint model. This model is built in an integrated manner. This model first collects same dimensionality subspace in distinct feature passage, second calculates the component feature in a subspace and lastly transfer the feature through learning pipeline. All-encompassing investigational outcomes on 04 famous datasets have established the success of the proposed method.

Visualisation CNN. The authors [60] presented 03 layer work for action recognition that is mainly focused on the invariant view. In the 1st layer, sequence-based view-invariant transform is considered for skeleton joints. In the 2nd layer, transform skeleton is applied on colour pictures as a sequence which is encrypted on the Spatio-temporal data of skeleton joints. In the 3rd layer, CNN based method find out more accurate features from colour pictures. With this action, values are finally generated via decision level synthesis of deep topographies. The extensive experiments on 04 inspiring data sets proved the preeminence of the proposed method.

GCALSTM. The authors [61] worked on 3D action recognition, which concentrates on information joints with the help of comprehensive contextual information, namely Global Context-aware Attention LSTM (GCALSTM) as part of the LSTM network. For iteratively improved responsiveness performance, the authors also considered GCA-LSTM network. The experiments showed that the proposed network harvests state-of-the-art performance on 03 stimulating datasets for 3D human action recognition.

Two-stream RNN. The authors [62] described a 2-way RNN scheme by using altitudinal and progressive dynamics for skeleton action-recognition. In 2-way RNN scheme 02 temporal classes, namely stacked RNN and hierarchical RNN is used, and those construct human body kinematics. Both of the categories used in the spatial scheme transformed into a spatial graph. The authors also used data augmentation method for 3D transformation based on arbitrary view. The experiment conducted on 3D action-recognition benchmark shows that the proposed method conveyed a substantial enhancement for a diversity of actions, i.e. all-purpose movements and collaboration activities.

C-CNN + MTLN. The authors [63] investigated 3D human action recognition with skeleton sequences. Every skeleton sequences divided into three joints, and every joint contains the different frames for spatial-temporal features. Every frame has cylindrical coordinates of the skeleton sequence generated through one tunnel. The authors investigated and used Deep convolution neural networks to be able to learn. The long-lasting progressive information of skeleton-sequence by the use of Multi-Task Learning Network (MTLN). The Investigational result demonstrates the efficiency of the proposed method.

Ensembles Temporal Sliding LSTM Networks (Ensemble TS-LSTM). By using GramSchmidt process, the human skeleton is transferred into the human perceptive coordinates structure, and takeout the vital information of pose and motion [64]. The authors also considered the innovative operative technique of LSTMs rendering to time-step magnitude, with training and testing procedures. The experiment shows the suggested networks overtake several state-of-the-art action-recognition approaches on the 05 different standard datasets.

View Adaptive (VA)-LSTM. The authors [65] projected an adaptive recurrent neural network (RNN) with LSTM scheme, which automate the surveillance viewpoint throughout the proposition of human action. The proposed method independently identify its own values from endpoints. The proposed model attains noteworthy enhancement over the state-of-the-art strategies on 03 yardstick datasets.

Learning action recognition model. The authors [66] considered a deep model which work on human-object interaction and intra-class movements under viewpoints variants. There are two aspects of the research, first depth appearance of body parts which is shared via view-invariant space and Second, endwise learning scheme. The authors estimated the performance of the proposed model against 15 prevalent techniques on 02 enormous benchmarks of human-action-recognition datasets with NTURGB+D and UWA3DII. The Experimental consequences show that the proposed technique delivers a noteworthy improvement over state-of-the-art approaches.

Spatio-Temporal Attention (STA)-LSTM. The authors [67] investigated the endwise spatial-temporal model for action-recognition. The construction of proposed work is based on the Recurrent Neural Networks (RNNs) with Long-Short-Term-Memory (LSTM). The experimental outcomes show the efficiency of the investigated model for SBU & NTU datasets.

LSTM and CNN score fusion. The authors [68] presented an LSTM and CNN score fusion approaches. Both approaches are very advantageous, LSTM provides robust temporal data, whereas CNN is biased to robust spatial data. The multiple-score-fusion approach enhances accuracy. The proposed technique achieves state-of-the-art outcomes on NTURGB+D datasets for 3D human-action investigation. The projected method succeed 87.41% in terms of accurateness and graded 1st position in Large-Scale 3D Human-Activity-Analysis Test in Complexity visualization.

CNN based scheme. The authors [69] presented CNN based scheme for classification and detection procedure. The introduced detection scheme, detects the human-action in a batch processing, whereas online detection is used for the real-time presentation. The authors used a skeleton transformer for recognition construction and via automated procedure selects the skeleton-joints. The proposed method uses the 7-layer network and apply this to NTU RGB+D dataset and achieve 89.3% precision. For uncropped video detection, the authors proposed *window proposal network*. This method is applied to PKU-MMD dataset and achieves 93.7% mAP.

PKU-MMD. The authors [70] proposed a huge volume benchmark standard, namely PKU-MMD. PKU-MMD is used for 3D human action (multi-modality) understanding. The proposed benchmarks cover 1076 video sequences with 51 action groups, and 66 subjects with 03 camera viewpoints execute 20000 actions, 5.4 million frames. The authors conducted experiments on the proposed benchmark in terms of 02 circumstances and evaluate distinct approaches by a different metric. For this purpose, the authors used evaluating protocol, namely 2D-AP. The proposed benchmark gives an advantage for upcoming investigators.

Temporal Conv. The authors [71] considered Temporal-Convolutional-Neural-Networks (TCN) for 3D action-recognition. As associated with LSTM-based Recurrent-Neural-Network models, the proposed method is very useful for the 3D skeleton. By the use of TCN, the Spatio-temporal model is easily understood. The subsequent prototypical, ResTCN, achieve state-of-the-art outcomes on the 3D human action recognition dataset, NTURGBD.

Trust Gate ST-LSTM. The authors [72] investigated the ranking structure based traversal technique to remove smash and occlusion in 3d skeleton information. The authors proposed novel gating instruments within LSTM, which lean the consistency of the entering data by which updated context information is stored in the flip-flop. The proposed method achieved state-of-the-art representation on 04 challenging benchmark datasets for 3D human-action examination.

Part-aware LSTM. The authors [73] constructed a vast dataset for RGB+D-SBHAR, and it contained 56 thousand visualization sequences, 04 million frames composed of 40 different-subjects and 60 different-action groups comprised daily, mutual, and health-related actions. The authors also proposed a recurrent-neural-network scheme model for long period progressive correlation of the features. The experimental outcomes show the benefits of DL methods over state-of-the-art handcrafted features.

Rolling rotations. Rolling rotations worked on 3D rotation among numerous body parts of the skeleton. 3D rotation follows the exceptional orthogonal assemblage, which is a Riemannian-manifold [74]. The human action classification is a challenging task because of non-Euclidean space. The authors opens action-curves with lie-algebra or vector space by articulating the logarithm and rolling maps. Experimental outcomes on 03 action datasets show that the proposed approach achieves outstanding performance when compared to the state-of-the-art approaches [74].

Co-occurrence feature learning. The authors [75] investigated fully connected deep LSTM-network for SBHAR. The authors also considered skeleton as an entering vector according to the time period and explored a new regularization structure. The author also proposed a novel dropout method which handles cells and outgoing replies of the LSTM. Extensive experimental outcomes on 03 SBHAR datasets established the efficiency of the proposed prototypical.

Convolutional-neural-network. In this paper, the authors [76] suggested the ordered architecture of SBHAR with the help of CNN. Initially, the proposed method signifies a sequence of the skeleton by combining the joints directs.

Every vector represented consecutively. The matrix is converted into the images and is also established to grip the problem of variable-length. Afterwards, the final images are served to the CNN-model for recognition and feature-extraction purpose. The modest max-pooling plays an essential role in spatial feature selection and sequential frequency adjustment. Modest max-pooling gains more discriminative joint facts for various action and meanwhile address the variable frequency problem. The result obtained from the experiment shows that the proposed approach captures state-of-the-art representation with great computational-efficiency.

H-RNN. The authors [77] presented hierarchical RNN for SBHAR; in this approach, the authors separates the human-skeleton into five distinct part similar to the human body physical-structure. As the layer is incremented, the result obtained by the subnet is feed to the higher layers that shows an outstanding performance. The outcome of the skeleton order is supplied to the single-layer-perceptron, and the stored output of the perceptron conclude the final result. In the experiment, the authors compared the proposed method with five different Deep RNN and other prevalent methods over the openly accessible dataset. The experimental output shows that the presented model obtain advanced performance with excellent efficiency.

Jointly learning heterogeneous features. The authors presented [78] a novel approach called the Heterogeneous features learning model which provide the different features for recognition with RGB-D activity. The advanced output obtained over the three 3D activity set shows the effectiveness of the novel approach.

LieGroup. In this paper [79], the authors presented the novel skeleton representation, which clearly shows the 3D geometric connection among the different body parts via translation & rotation matrices. Subsequently, 3d stiff body indicates the participant of distinct Euclidean-group SE (3), the presented approach exist in the set of SE (3) ×__ × SE (3), curved multifarious. By using the given representation, action of human can be exhibited as a curve by Lie-group. However, the classification of curves by a Lie-group, not a relatively stress-free task. The authors carried out the classification by using the blend of dynamic-time-wrapping, Fourier-temporal-pyramid depiction and linear SVM. The investigational output over the three-action dataset illustrates that the proposed representation achieve excellent performance in many prevalent skeletal-representations. The proposed method also overtakes the different advanced skeleton-based human action recognition methodology.

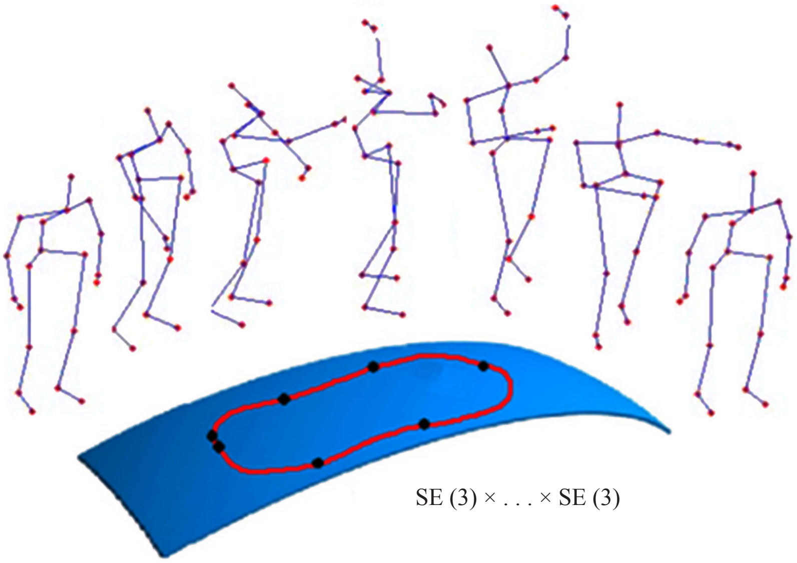

Figure 4. Representation of an action (skeletal sequence) as a curve in the Lie group SE (3) × . . . × SE (3)

Temporal-hierarchy-of-covariance-descriptors. In this paper [80], the authors provided the new method for SBHAR from the 3D-skeleton-sequences. Here, the authors also customized the covariance-matrix for skeleton-joint location over a period as a discriminative-descriptor for a sequence. The authors also use the multiple-covariance-matrices in an ordered manner. The descriptor contains the permanent length, which is free from the size of the designated arrangement. The proposed method outperforms in action-recognition over the multiple-datasets that are captured through the Kinect-type-sensor or via arbitrary view arrangement.

3. Datasets and performance

In this section, we described the two datasets i.e. NTURGB+D [72] and NTURGB+D 120 [36].

NTURGB+D [72] is an extensive state-of-the-art benchmark for SBHAR. It demonstrates a succession of standards and involvement of comprehensive data structure. Newly described outcomes on this dataset have obtained accurateness on this benchmark. NTURGB+D 120 [36] dataset is vast and delivers various variant of environmental circumstances, matters, and camera views/viewpoints etc. It contains 114,480 videos, 120 classes and 106 subjects.

NTURGB+D 120 dataset has several action-classes and numerous visualization samples for corresponding action-class, and extensive intra-class disparity. Kinectv2 delivers supplementary specific depth-maps & 3D-joints, especially in a multi-camera-setup compared to the earlier account of Kinect.

NTU RGB+D 120 dataset has an advantage over various research problems. The datasets worked very well for depth-based, 3D skeleton-based, 3D RGB-based and infrared-based action recognition etc. In addition to the above, the NTU RGB+D 120 dataset also obtain excellent efficiency for Heterogeneous-feature fusion investigation, deep net pre-training, and One-shot 3D action investigation [36].

Figure 5, reports the human action recognition-rates for numerous SBHAR on NTU-RGB+D dataset. The human-recognition in the most recent study provides the gratitude-rates for the cross-subject and cross-view. We can undoubtedly get that the expected representation perform better than the other method. Better performance on MS-G3D point out that the proposed representation serve better than others in terms of modelling complex-actions, and it achieves 91.5 and 96.2, precision in the cross-subject and cross-view.

Figure 6, reports the human action recognition-rates for numerous skeletal-representations on NTU-RGB+D 120 dataset and MS-G3D achieves 86.9 and 88.4 accuracies in cross-subject and cross-setup.

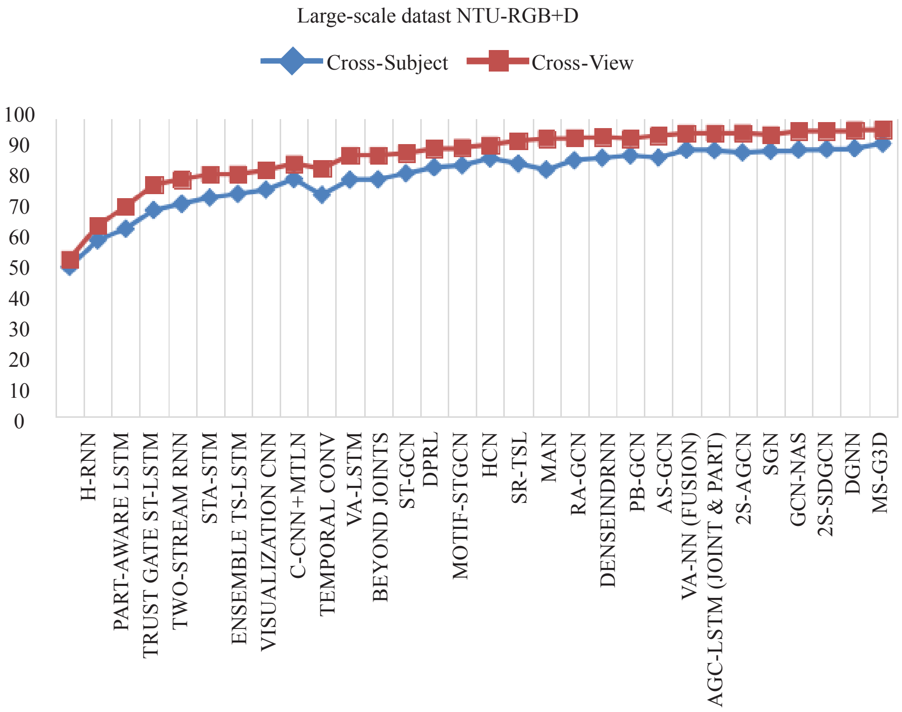

Figure 5. Numerous SBHAR method tested on NTURGB+D

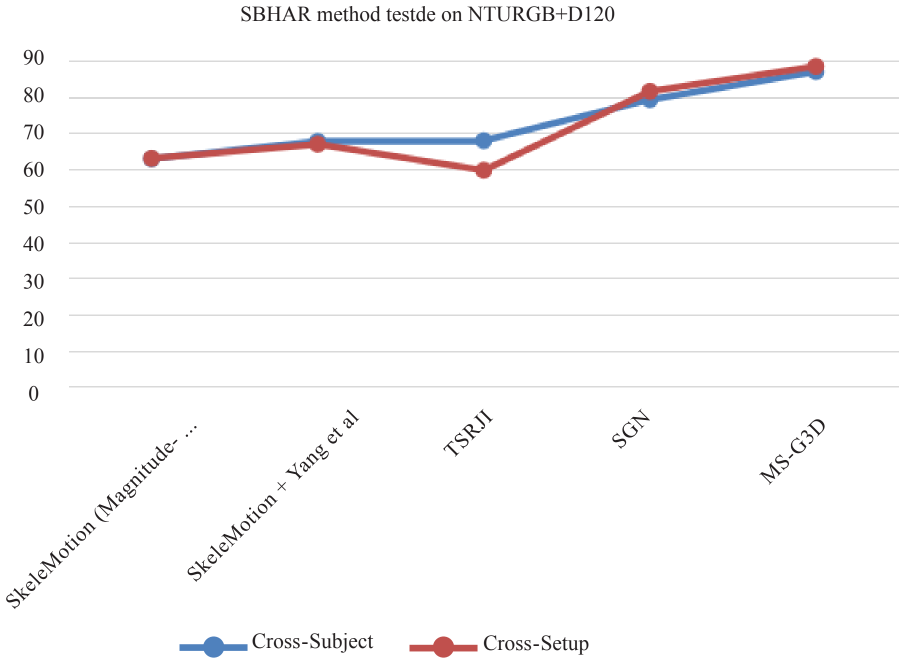

Figure 6. SBHAR method tested on NTURGB+D 120

4. Conclusion

In this research paper, we have discussed the future research direction for those who want to work further in this area, and we have also provided an in-depth overview of SBHAR. However, there has been a lot of good research on SBHAR such as selection and complexity of human body posture, occlusion, and background-clutter, despite this SBHAR in the real scenario is a challenging task. In this paper, we have studied the various SBHAR method and delivered a detailed explanation of current methods that include manual-designed-action features in RGB & Depth-data, DL constructed techniques and action-recognition method. After a detailed literature survey, we introduced the primary operative method, so that researcher is promptly acquainted with the related research area.

References

[1] Shabani A. H., Clausi D., Zelek J. S. Improved spatio-temporal salient feature detection for action recognition. *British Machine Vision Conference.* 2011. p.1-12.

[2] Li R., Zickler T. Discriminative virtual views for cross-view action recognition. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2012. p.2855-2862.

[3] Li B., Camps O. I., Sznaier M. Cross-view activity recognition using hankelets. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2012. p.1362-1369.

[4] Vrigkas M., Karavasilis V., Nikou C., et al. Action recognition by matching clustered trajectories of motion vectors. *International Conference on Computer Vision Theory and Applications*. 2013. p.112-117.

[5] Lan T., Wang Y., Mori G. Discriminative figure-centric models for joint action localisation and recognition. *IEEE International Conference on Computer Vision*. 2011. p.2003-2010.

[6] Iosifidis A., Tefas A., Pitas I. Activity-based person identification using fuzzy representation and discriminant learning. *IEEE Trans. Inform. Forensics Secur*. 2012; 7: 530-542. Available from: doi:10.1109/TIFS.2011.2175921.

[7] Patwary M. J., Wang X. Z., Yan D. Impact of fuzziness measures on the performance of semi-supervised learning. *International Journal of Fuzzy Systems*. 2019; 21(5): 1430-1442.

[8] Morariu V. I., Davis L. S. Multi-agent event recognition in structured scenarios. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2011. p.3289-3296.

[9] Chen C. Y., Grauman K. Efficient activity detection with maxsubgraph search. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2012. p.1274-1281.

[10] Sigal L., Isard M., Haussecker H., et al. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *Int. J. Comput. Vis*. 2012; 98: 15-48. Available from: doi: 10.1007/s11263-011-0493-4.

[11] Tran K. N., Kakadiaris I. A., Shah S. K. Part-based motion descriptor image for human action recognition. *Pattern Recognit*. 2012; 45: 2562-2572. Available from: doi:10.1016/j.patcog.2011.12.028.

[12] Wu Q., Wang Z., Deng F., et al. Realistic human action recognition with multimodal feature selection and fusion. *IEEE Trans. Syst. Man Cybern. Syst*. 2013; 43: 875-885. Available from: doi:10.1109/TSMCA.2012.2226575.

[13] Liu N., Dellandréa E., Tellez B., et al. Associating textual features with visual ones to improve affective image classification. *International Conference on Affective Computing and Intelligent Interaction*. 2011; 6974: 195-204.

[14] Martinez H. P., Yannakakis G. N., Hallam J. Don't classify ratings of affect; rank them! *IEEE Trans. Affective Comput*. 2014; 5: 314-326. Available from: doi: 10.1109/ TAFFC.2014.2352268.

[15] Song Y., Morency L. P., Davis R. Multimodal human behavior analysis: Learning correlation and interaction across modalities. *ACM International Conference on Multimodal Interaction*. 2012. p.27-30.

[16] Vrigkas M., Nikou C., Kakadiaris I. A. Classifying behavioral attributes using conditional random fields. *8th Hellenic Conference on Artificial Intelligence, Lecture Notes in Computer Science*. 2014. p.95-104.

[17] Patron-Perez A., Marszalek M., Reid I., et al. Structured learning of human interactions in TV shows. *IEEE Trans. Pattern Anal. Mach. Intell*. 2012; 34: 2441-2453. Available from: doi:10.1109/TPAMI.2012.24.

[18] Marín Jiménez M. J., Noz Salinas R. M., Yeguas Bolivar E., et al. Human interaction categorization by using audio-visual cues. *Mach. Vis. Appl*. 2014; 25: 71-84. Available from: doi: 10.1007/s00138-013-0521-1.

[19] Shugang Zhang, Zhiqiang Wei, Jie Nie, et al. A review on human activity recognition using vision-based method. *Journal of Healthcare Engineering*. 2017; 1-32. Available from: https://doi.org/10.1155/2017/3090343.

[20] Aggarwal J.K., Ryoo M.S. Human activity analysis: A review. *ACM Comput. Surv*. 2011; 43(3): 16.

[21] Ziaeefard M., Bergevin R. Semantic human activity recognition: A literature review. *Pattern Recognit*. 2015; 48: 2329-2345.

[22] Van Gemert J.C., Jain M., Gati E., et al. APT: Action localisation proposals from dense trajectories. *In Proceedings of the British Machine Vision Conference 2015: BMVC 2015*. 2015. p.1-4.

[23] Zhu H., Vial R., Lu S. Tornado: A spatio-temporal convolutional regression network for video action proposal. *In Proceedings of the CVPR*. 2017. p.5813-5821.

[24] Papadopoulos G.T., Axenopoulos A., Daras P. Real-time skeleton-tracking-based human action recognition using kinect data. *In Proceedings of the International Conference on Multimedia Modeling*. 2014. p.473-483.

[25] Presti L.L., Cascia M.L. 3D skeleton-based human action classification: A survey. *Pattern Recognition*. 2016; 53: 130-147.

[26] Paul S.N., Singh Y.J. Survey on video analysis of human walking motion. *Int. J. Signal Process. Image Process. Pattern Recognition*. 2014; 7(3): 99-122.

[27] Li M., Chen S., Chen X., et al. Symbiotic graph neural networks for 3D skeleton-based human action recognition and motion prediction. *ArXiv: Computer Vision and Pattern Recognition.* 2019. Available from: https://arxiv.org/abs/1910.02212.

[28] Li S., Li W., Cook C., et al. Deep independently recurrent neural network (IndRNN). *ArXiv: Computer Vision and Pattern Recognition.* 2019. Available from: https://arxiv.org/abs/1910.06251.

[29] Gao X., Hu W., Tang J., et al. Optimized skeleton-based action recognition via sparsified graph regression. *Proceedings of the 27th ACM International Conference on Multimedia*. 2019. p.601-610.

[30] Hu G., Cui B., Yu S. Skeleton-based action recognition with synchronous local and non-local spatio-temporal learning and frequency attention. *IEEE International Conference on Multimedia and Expo (ICME)*. 2019. p.1216-1221.

[31] Shi L., Zhang Y., Cheng J., et al. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *ArXiv: Computer Vision and Pattern Recognition*. 2019. Available from: https://arxiv.org/abs/1912.06971.

[32] Wang L., Huynh D.Q., Koniusz P. A comparative review of recent kinect-based action recognition algorithms. *IEEE Transactions on Image Processing*. 2020; 29: 15-28.

[33] Liu Z., Zhang H., Chen Z., et al. Disentangling and unifying graph convolutions for skeleton-based action recognition. *ArXiv: Computer Vision and Pattern Recognition*. 2020. Available from: https://arxiv.org/abs/2003.14111.

[34] Zhang P., Lan C., Zeng W., et al. Semantics-guided neural networks for efficient skeleton-based human action recognition. *ArXiv: Computer Vision and Pattern Recognition*. 2019. Available from: https://arxiv.org/abs/1904.01189.

[35] Peng W., Hong X., Chen H., et al. Learning graph convolutional network for skeleton-based human action Recognition by neural searching. *ArXiv: Computer Vision and Pattern Recognition*. 2019. Available from: https://arxiv.org/abs/1911.04131.

[36] Liu J., Shahroudy A., Perez M., et al. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*. 2019. p. 2684-2701.

[37] Zhang P., Lan C., Xing J., et al. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2019; 41: 1963-1978.

[38] Si C., Chen W., Wang W., et al. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019. p.1227-1236.

[39] Wu C., Wu X., Kittler J. Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition. *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 2019. p.1740-1748.

[40] Shi L., Zhang Y., Cheng J., et al. Skeleton-based action recognition with directed graph neural networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019. p.7904-7913.

[41] Shi L., Zhang Y., Cheng J., et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019. p.12018-12027.

[42] Li M., Chen S., Chen X., et al. Actional-structural graph convolutional networks for skeleton-based action recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019. p.3590-3598.

[43] Si C., Chen W., Wang W., et al. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019. p.1227-1236.

[44] Wen Y., Gao L., Fu H., et al. Graph CNNs with motif and variable temporal block for skeleton-based action recognition. *AAAI.* 2019; 33(1): 8989-8996.

[45] Song Y., Zhang Z., Wang L. Richly activated graph convolutional network for action recognition with incomplete skeletons. *IEEE International Conference on Image Processing (ICIP)*. 2019. p.1-5.

[46] Caetano C., Brémond F., Schwartz W.R. Skeleton image representation for 3D action recognition based on tree structure and reference joints. *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. 2019. p.16-23.

[47] Caetano C., Souza J.S., Brémond F., et al. SkeleMotion: A new representation of skeleton joint sequences based on motion information for 3D action recognition. *16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2019. p.1-8.

[48] Cho S., Maqbool M.H., Liu F., et al. Self-attention network for skeleton-based human action recognition. *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2020. p.624-633.

[49] Wang H., Wang L. Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection. *IEEE Transactions on Image Processing*. 2018; 27: 4382-4394.

[50] Tang Y., Tian Y., Lu J., et al. Deep progressive reinforcement learning for skeleton-based action recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018. p.5323-5332.

[51] Si C., Jing Y., Wang W., et al. Skeleton-based action recognition with spatial reasoning and temporal stack learning. *ECCV.* 2018; 106-121.

[52] Li C., Zhong Q., Xie D., et al. Co-occurrence feature learning from skeleton data for action recognition and detection

with hierarchical aggregation. *ArXiv: Computer Vision and Pattern Recognition.* 2018. Available from: https://arxiv.org/abs/1804.06055.

[53] Xie C., Li C., Zhang B., et al. Memory attention networks for skeleton-based action recognition. *ArXiv: Computer Vision and Pattern Recognition.* 2018. Available from: https://arxiv.org/abs/1804.08254.

[54] Yan S., Xiong Y., Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI Conference on Artificial Intelligence.* 2018. p.7444-7452.

[55] Li C., Cui Z., Zheng W., et al. Spatio-temporal graph convolution for skeleton based action recognition. *ArXiv: Computer Vision and Pattern Recognition.* 2018. Available from: https://arxiv.org/abs/1802.09834.

[56] Thakkar K.C., Narayanan P.J. Part-based graph convolutional network for action recognition. *AArXiv: Computer Vision and Pattern Recognition.* 2018. Available from: https://arxiv.org/abs/1809.04983.

[57] Minh T.L., Inoue N., Shinoda K. A fine-to-coarse convolutional neural network for 3D human action recognition. *ArXiv: Computer Vision and Pattern Recognition.* 2018. Available from: https://arxiv.org/abs/1805.11790.

[58] Ji Y., Xu F., Yang Y., et al. A large-scale varying-view RGB-D action dataset for arbitrary-view human action recognition. *ArXiv: Computer Vision and Pattern Recognition.* 2019. Available from: https://arxiv.org/abs/1904.10681.

[59] Hu J., Zheng W., Lai J., et al. Jointly learning heterogeneous features for RGB-D activity recognition. *CVPR*. 2015; 39(11): 5344-5352.

[60] Liu M., Liu H., Chen C. Enhanced skeleton visualisation for view invariant human action recognition. *Pattern Recognition.* 2017; 68: 346-362.

[61] Liu J., Wang G., Hu P., et al. Global context-aware attention LSTM networks for 3D Action Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017. p.3671-3680.

[62] Wang H., Wang L. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017. p.3633-3642.

[63] Ke Q., Bennamoun M., An S., et al. A new representation of skeleton sequences for 3D action recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017. p.4570-4579.

[64] Lee I., Kim D., Kang S., et al. Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks. *IEEE International Conference on Computer Vision (ICCV)*. 2017. p.1012-1020.

[65] Zhang P., Lan C., Xing J., et al. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017. p.2136-2145.

[66] Rahmani H., Bennamoun M. Learning action recognition model from depth and skeleton videos. *IEEE International Conference on Computer Vision (ICCV)*. 2017. p.5833-5842.

[67] Song S., Lan C., Xing J., et al. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. *AAAI Conference on Artificial Intelligence*. 2017. p.4263-4270.

[68] Li C., Wang P., Wang S., et al. Skeleton-based action recognition using LSTM and CNN. *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. 2017. p.585-590.

[69] Li C., Zhong Q., Xie D., et al. Skeleton-based action recognition with convolutional neural networks. *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. 2017. p.597-600.

[70] Liu C., Hu Y., Li Y., et al. PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding. *ArXiv: Computer Vision and Pattern Recognition.* 2017. Available from: https://arxiv.org/abs/1703.07475.

[71] Kim T.S., Reiter A. Interpretable 3D human action analysis with temporal convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2017. p.1623-1631.

[72] Liu J., Shahroudy A., Xu D., et al. Spatio-temporal LSTM with trust gates for 3D human action recognition. *ArXiv: Computer Vision and Pattern Recognition*. 2016. Available from: https://arxiv.org/abs/1607.07043.

[73] Shahroudy A., Liu J., Ng T., et al. NTU RGB+D: A large scale dataset for 3D human activity analysis. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. p.1010-1019.

[74] Vemulapalli R., Chellappa R. Rolling rotations for recognising human actions from 3D skeletal data. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. p.4471-4479.

[75] Zhu W., Lan C., Xing J., et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. *AAAI Conference on Artificial Intelligence*. 2016. p.3697-3703.

[76] Du Y., Fu Y., Wang L. Skeleton based action recognition with convolutional neural network. *3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. 2015. p.579-583.

[77] Du Y., Wang W., Wang L. Hierarchical recurrent neural network for skeleton based action recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015. p.1110-1118.

[78] Hu J., Zheng W., Lai J., et al. Jointly learning heterogeneous features for RGB-D activity recognition. *CVPR*. 2015; 39(11): 5344-5352.

[79] Vemulapalli R., Arrate F., Chellappa R. Human action recognition by representing 3D skeletons as points in a lie

group. *IEEE Conference on Computer Vision and Pattern Recognition*. 2014. p.588-595.

[80] Hussein M.E., Torki M., Gowayyed M.A., et al. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. *IJCAI*. 2013; 2466-2472.