Pattern Recognition Letters 145 (2021) 216–224

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Action recognition using kinematics posture feature on 3D skeleton joint locations

Md Atiqur Rahman Ahad[a,d,*], Masud Ahmed[b], Anindya Das Antar[c], Yasushi Makihara[a], Yasushi Yagi[a]

[a] Osaka university, Japan
[b] University of Maryland, Baltimore, USA
[c] University of Michigan, Ann Arbor, USA
[d] University of Dhaka, Bangladesh

ARTICLE INFO

Article history:
Received 9 September 2020
Revised 31 January 2021
Accepted 26 February 2021
Available online 3 March 2021

MSC:
41A05
41A10
65D05
65D17

Keywords:
Action recognition
Skeleton data
Kinematics posture feature (KPF)
Position-based statistical feature (PSF)
Joint angle
Joint position
Deep neural network
Ensemble architecture
Convrnn
Benchmark datasets
Linear joint position feature (LJPF)
Angular joint position feature (AJPF)

ABSTRACT

Action recognition is a very widely explored research area in computer vision and related fields. We propose Kinematics Posture Feature (KPF) extraction from 3D joint positions based on skeleton data for improving the performance of action recognition. In this approach, we consider the skeleton 3D joints as kinematics sensors. We propose Linear Joint Position Feature (LJPF) and Angular Joint Position Feature (AJPF) based on 3D linear joint positions and angles between bone segments. We then combine these two kinematics features for each video frame for each action to create the KPF feature sets. These feature sets encode the variation of motion in the temporal domain as if each body joint represents kinematics position and orientation sensors. In the next stage, we process the extracted KPF feature descriptor by using a low pass filter, and segment them by using sliding windows with optimized length. This concept resembles the approach of processing kinematics sensor data. From the segmented windows, we compute the Position-based Statistical Feature (PSF). These features consist of temporal domain statistical features (e.g., mean, standard deviation, variance, etc.). These statistical features encode the variation of postures (i.e., joint positions and angles) across the video frames. For performing classification, we explore Support Vector Machine (Linear), RNN, CNNRNN, and ConvRNN model. The proposed PSF feature sets demonstrate prominent performance in both statistical machine learning- and deep learning-based models. For evaluation, we explore five benchmark datasets namely UTKinect-Action3D, Kinect Activity Recognition Dataset (KARD), MSR 3D Action Pairs, Florence 3D, and Office Activity Dataset (OAD). To prevent overfitting, we consider the leave-one-subject-out framework as the experimental setup and perform 10-fold cross-validation. Our approach outperforms several existing methods in these benchmark datasets and achieves very promising classification performance.

1. Introduction

Human Action Recognition (HAR) is one of the most prominent and challenging research areas of computer vision and ubiquitous computing in recent years [1,2]. HAR plays a fundamental role in numerous relevant and heterogeneous application fields from the most commercial to the most assistive ones. The most significant application domains include assistive living, health care, video-surveillance, augmented reality, patient monitoring, intelligent surgery, and so on [1,2]. We can also take the help of Active and Assisted Living (AAL) tools to reduce the social-cost and mitigate the challenges of the aging population in the modern and developed society. For example, automated vision-based assistive systems can keep track of how often a person drinks water or take medicines to coach the users for behavior modifications to maintain a healthy lifestyle.

Despite the research endeavor by numerous researchers and promising advances over the past decade, there are still some vital challenges for the accurate recognition of human actions. Because of the articulated nature of human motion, there lies a major issue to model the human actions that are ambiguous, dynamic, and interactive with other objects. This difficulty poses a limitation on the performance of video-based action recognition as indicated in the previous studies [3]. Modeling specific temporal structures of human actions is another challenging task, along with privacy

* Corresponding author.
E-mail addresses: atiqahad@du.ac.bd (M.A.R. Ahad), mahmed10@umbc.edu (M. Ahmed), adantar@umich.edu (A. Das Antar), makihara@am.sanken.osaka-u.ac.jp (Y. Makihara), yagi@sanken.osaka-u.ac.jp (Y. Yagi).

https://doi.org/10.1016/j.patrec.2021.02.013
0167-8655/

issues. Most of the previous works on activity classification have focused on RGB videos [4–6] RFID sensors and radio-based solutions [7], a combination of wearable and ambient sensor data [8,9]. The use of RGB videos performs relatively poor even when there is no clutter [2]. Because of the placement of RFID tags, the use of this method is generally too intrusive and uncomfortable for the users. RGB videos and images also pose security and privacy issues. Recently, cost-effective depth cameras have fostered the progress of promising approaches to develop reliable and cost-effective solutions by providing 3D depth and skeleton data of the scene that largely eases the task of action recognition.

In this paper, we have exploited the skeleton data in our research work, as depth data perform poorly in low resolution and demonstrate poor results in capturing the appearance information. As we have the availability of skeletal joint data, the primary hypothesis is to consider the change of joint position across video frames to encode complex human actions. In this regard, our proposed idea considers each available body joint as kinematics sensors (for example, acceleration sensors or position sensors). The hypothesis is that, if we consider a position or acceleration sensor on each human body joint location (torso, head, leg, neck, elbow, shoulder, etc.) and collect time-series data from each sensor, these data can help to identify complex human activities. In this case, body skeleton data (joint position and orientation) can also serve the same purpose if we consider those joints as kinematics sensors (similar to wearable sensors) and extract robust features after the data processing stage.

Thus, we have mainly proposed a novel idea to extract a Kinematics Posture Feature (KPF) set (linear joint position and the angle between bone segments) from 3D joint positions in the temporal domain based on skeleton data. These KPF feature sets combine Linear Joint Position Feature (LJPF) and Angular Joint Position Feature (AJPF), where LJPF encodes normalized linear joint position information and AJPF encodes the information about the angles between bone segments considering each joint as kinematics sensors. These Kinematics Posture Feature (KPF) sets across video frames can encode the motion information of human body joints with time while performing different actions. After processing and segmenting the data from this feature set, we have computed some statistical features (e.g., mean, standard deviation, variance, etc.) separately. Based on these steps, we have produced Position-based Statistical Feature (PSF) from the Kinematics Posture Feature (KPF) set mentioned earlier. In the next stage, we have classified the actions using the Support Vector Machine (SVM) with linear kernel, Convolutional Recurrent Neural Network (ConvRNN), CNNRNN, and RNN using the feature set (PSF) to predict the change of motion with time in real-time application.

The background and development of RGBD and skeleton data-based action recognition sector have been discussed in this Section 1. The rest of the paper is organized as follows: Section 2 provides a brief review of related research work. The proposed method of extracting Kinematics Posture Feature (KPF) set combining Linear Joint Position Feature (LJPF) and Angular Joint Position Feature (AJPF), which can encode motion information across video frames from skeleton data has been described in detail in Section 3. Here, we discussed the processing part of the extracted features to eliminate noise components and sliding window-based segmentation methods for feature extraction. We also proposed Position-based Statistical Feature (PSF) sets consisting of statistical features in the temporal domain. We extracted PSF feature sets from segmented KPF features across segmented windows consisting of video frames. Section 4 describes the basic information of five benchmark datasets used in this work for the evaluation of our method. Section 5 discusses the applicability of our method in action recognition scenario, showing the results and comparison with previous methods on these datasets. Finally, we concluded the paper in Section 6, highlighting some future challenges to improve this method.

2. Related works

In the last several years, a number of solutions are proposed for skeleton-based human action recognition. Multi-camera motion capture (MoCap) systems can be used to produce more accurate 3D joint positions. However, the MoCap system is based on various on-body markers and is very expensive. The low-cost Kinect sensor can capture depth images and track skeletal joints. Each joint can depict position and orientation in the 3D space. These are explored for action recognition, as affirmed by many propositions in the literature.

Different representations are adopted based on the set of joints, such as the simple joint coordinates, normalized according to the body reference measure [10,11] or joint distances [12], histograms of 3D joints [13], EigenJoints in [14] where PCA is applied to static and dynamic posture features to create a motion model, 3D representation of skeleton joints positions using Gaussian Mixture Models [15], Dynamic Bayesian Mixture Model of 3D skeleton features [16], or spatiotemporal interest points and descriptors derived from the depth image [17]. There are also some other common approaches called action lets [3,18–20] where a hierarchical representation is adopted assuming that an activity is composed of a set of sub-activities. Skeleton and RGB are jointly explored where, the temporal evolution is retained by using RGB [5,21].

Besides, the interaction of humans with objects have been analyzed for better scene understanding. Research work [20] adopted a Markov Random Field, where the edges represent the relationships among object affordances, whereas, the nodes represent objects and sub-activities along with their relation with sub-activities. On the other hand, the authors in [20] proposed a graph-based representation. The Histograms of 3D Joint Locations (HOJ3D) [13] is another approach of joint representation. For the HOJ3D, the 3D space is split into several bins, and the 3D skeleton joint locations are associated with the bins by employing Gaussian weight function. Afterward, a Hidden Markov Model (HMM) is modeled to engrave the temporal evolution of posture visual words. The method by [12] employs joint spherical coordinates to represent the skeleton, and a framework composed of a multiclass SVM and a discrete HMM to recognize activities.

Deep learning-based approaches especially neural-networks are also famous for classifying skeletal data. As Long short-term networks can process the changes over time, this method has been analyzed in research work [22]. They have shown a graph-based deep learning approach named as Time based Graph Long Short-Term Memory (TGLSTM) network for gait and action recognition, which can learn dynamically while changing over time.

However, we analyze that to classify complex action, the processed kinematics features that are extracted from skeletal joints can be a possible option, which has not been explored in previous research works. As most of the actions consist of several sub-actions, the kinematics of body joint positions across frames can play an important role to distinguish among different actions. Based on this idea, the work presented in this paper relies on kinematics features, extracted from the joint skeleton data to classify the actions based on the motion information of joint positions and angles. The proposed features overcome the limitations of complex expensive algorithms based on RGB and depth data, with good performance and execution time. Though there is a recent trend of exploiting deep learning-based methods for skeleton-based activity recognition, we feel that this domain can be richer by introducing smarter features as well combined with a deep learning-based approach.

3. Methodology

Our proposed feature sets are based on position- and angle-based kinematics that confers some advantages: firstly, there are no strict privacy issues for the user due to the use of skeleton data. Secondly, the angle information obtained from skeletons is intrinsically normalized. Therefore, angle-based features are independent of a users physical build. We tried to capture the motion of performed actions across frames by calculating the joint linear position to head. Besides, we also analyzed the angle between bone segments to distinguish among the action classes.

3.1. Position-based kinematics feature

The skeleton data consists of a set of P joints $J = [J_1, J_2, J_3, \ldots, J_P]$ where, P = the number of joints, that depends on the methods used for skeleton tracking. In general, most of the datasets contain 15, 20, or 30 joints. Each joint can be represented by $J_i = (p_i, o_i)$; where, p_i = position vector of each joint, and o_i = 3D orientation vector of each joint with respect to the world coordinate. We worked with skeleton joints and computed a vector of features for each activity considering the joints as kinematics sensors (for example, wearable acceleration or position sensor). The posture features were extracted from each video frame using the skeleton joints to evaluate the feature vectors, which represent human postures. This feature extraction process consists of three steps. Firstly, we extracted the posture features using the distance of joint locations with respect to the head. The rate of change of joint positions across video frames contains spatial and temporal data along with motion information. We summarized the entire process in the following parts.

We computed position-based kinematics feature vector (linear joint position) for each skeleton frame, where each joint is represented by a three-dimensional vector J_i in the coordinate space of Kinect. For the i^{th} joint J_i, a feature vector $d_{i(head)}$ has been calculated. The $d_{i(head)}$ is the distance vector between any joint J_i and head joint J_{head}. Moreover, we normalized the distance vector with respect to the distance between neck joint J_{neck} and torso joint J_{torso}. This normalization procedure makes this feature invariant to the physical build of the person. The distance vectors with respect to head for each joint are demonstrated in Fig. 1 and $d_{i(head)}$ is computed as follows:

$$d_{i(head)} = \frac{J_i - J_{head}}{||J_{neck} - J_{torso}||}, \quad i = 1, 2, \ldots, P-1 \tag{1}$$

where, J_{neck} = position coordinate of the neck joint, J_{head} = position coordinate of the head joint, and J_{torso} = position coordinate of the torso joint, and P = number of joints.

In the second stage, for each n^{th} frame, we calculated a posture feature vector namely, Linear Joint Position Feature (LJPF), $Pos_{n(head)}$ for each skeleton frame that can be represented by the following equation, respectively:

$$Pos_{n(head)} = [d_{\{n,1\}}, d_{\{n,2\}}, ., d_{\{n,P-1\}}], \quad n = 1, 2, ., N \tag{2}$$

where, P = number of joints, $d_{n,1}$ = distance component of joint 1 of the n^{th} frame with respect to head, $d_{n,2}$ = distance component of joint 2 of the n^{th} frame with respect to head, $Pos_{n(head)}$ = Linear Joint Position Feature (LJPF) encoding normalized distance vectors with respect to head, and N = number of frames. In this case, a set of N feature vectors was computed for each case having an activity constituted by N frames.

3.2. Angle-based kinematics feature

In this case, we compiled the relative positions of the different body parts by encoding each frame of a video sequence as a set of angles, which can be derived from the human skeleton data. We named this feature as the angle between the bone segments. In this approach, we computed the relevant angles between two bone segments, whose Spatio-temporal evolution characterizes an activity. The rate of change of angular displacement is useful to distinguish between the action patterns.

Among all possible angles between available joint positions, we considered only a subset of the possible angles to remove the non-informative angles: e.g., the angles between head and neck for all frames are almost constant over time and may not provide useful information for the discrimination of activities. The entire process

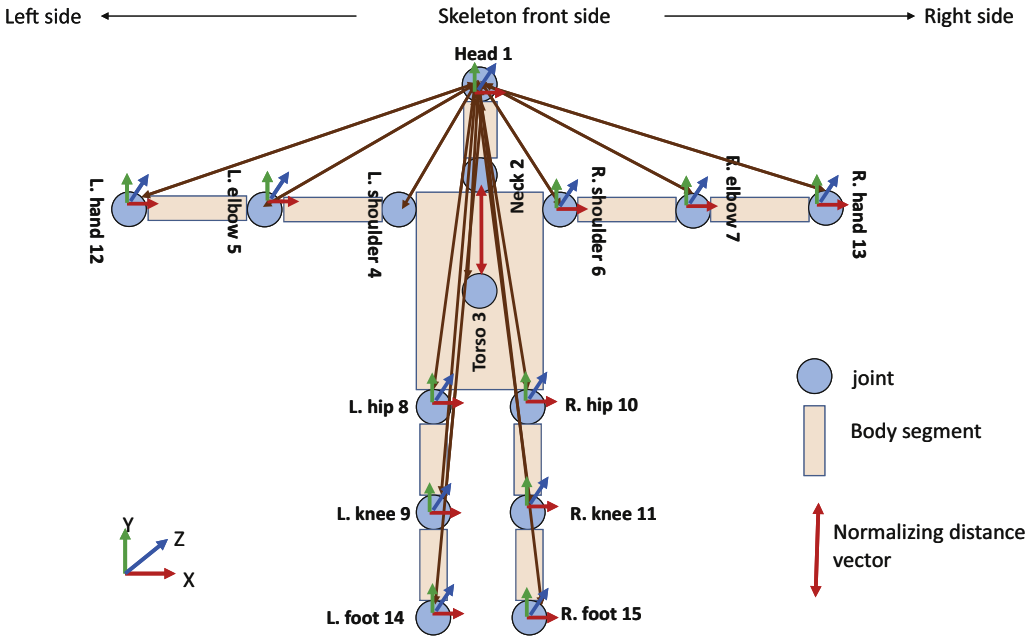

Fig. 1. Distance vector calculation for each joint with respect to head and normalized by the distance between torso and neck (each body joint has been considered as a kinematics sensor).

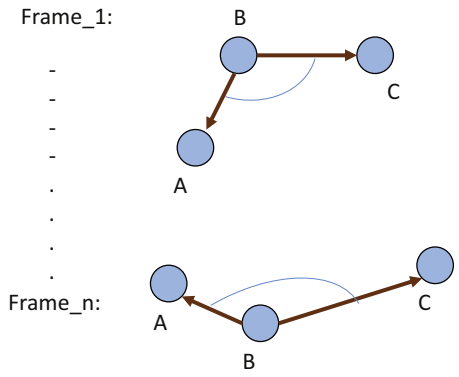

Fig. 2. Set of three joints for calculating the angle between two bone segments.

is illustrated below. To compute the angle between two bone segments for each frame, we considered 3 joint positions at the initial stage (Fig. 2).

We considered the following set of joints for calculating the angle between two bone segments after empirical investigation as shown in Fig. 3. Mentioned joint sets for calculating the angle-based kinematics feature, $\angle\theta(radian)$: [(Torso, L. hip, L. knee), (Torso, R. hip, R. knee), (L. hip, L. knee, L. foot), (R. hip, R. knee, R. foot), (L. shoulder, L. elbow, L. hand), and (R. shoulder, R. elbow, R. hand)], where R. denotes Right and L. means Left side.

We considered these angle values for each set of joints for each frame. This angular kinematics feature has lots of importance, for example, the change of angular information regarding the angle between hip, knee, and left foot and right foot will help to distinguish between sitting and standing actions. Besides, the angular information regarding shoulder, elbow, and hand joints can help to recognize the actions and gestures mostly performed by hands. This angle-based kinematics feature is termed as Angular Joint Position Feature (AJPF).

3.3. Kinematics posture feature (KPF)

We already described the extraction of two feature sets namely Linear Joint Position Feature (LJPF) and Angular Joint Position Feature (AJPF). For each frame, Linear Joint Position Feature (LJPF) encodes the normalized distance vectors of all considered body joints with respect to the head. Similarly, for each frame, Angular Joint Position Feature (AJPF) encodes the angles between different bone segments (e.g., shoulder, knee, hip, etc.). These two features (position and angular information of all body joints) work as a blueprint of the corresponding body-posture for that particular video frame. We considered an analogy that, by this approach, we can utilize all of the body joints as individual kinematics position and orientation sensors. We can track the change of postures across video frames by tracking the change in joint position and angles through these two features (LJPF and AJPF). Thus, for each video frame, we combined these two features to generate Kinematics Posture Feature (KPF) set. This feature encodes the change in joint position and angles across video frames.

While combining LJPF and AJPF features to create the KPF feature, we considered normalized joint positions with respect to the head and angle between bone segments for the body joints in a frame. These positions and angular information are combined (jointly considered as two feature vectors for each video frame) to create a Kinematics Posture Feature (KPF) set for each frame. We hypothesized that the change of joint position and angular information across video frames from this feature set can encode the variation in posture. Thus, it can be helpful to encode complex motion information for an accurate action recognition procedure.

3.4. Processing, segmentation, and generation of position-based statistical feature (PSF)

The shape of the position vector and angle vector are noisy and unsmooth. Because of any unsmooth amplitude, even after the normalization procedure with respect to the body joint, the Kinemat-

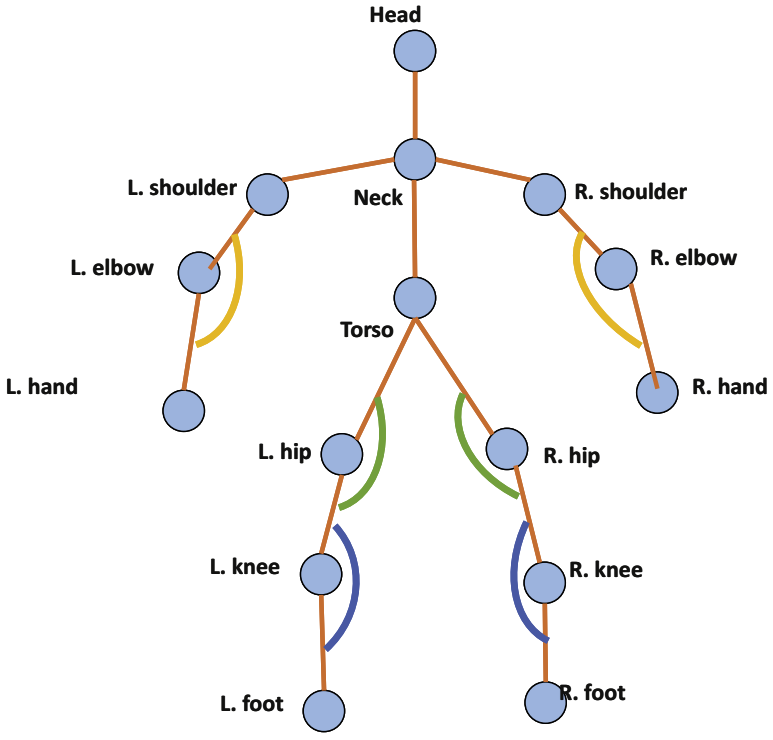

Fig. 3. Considered angles between bone segments.

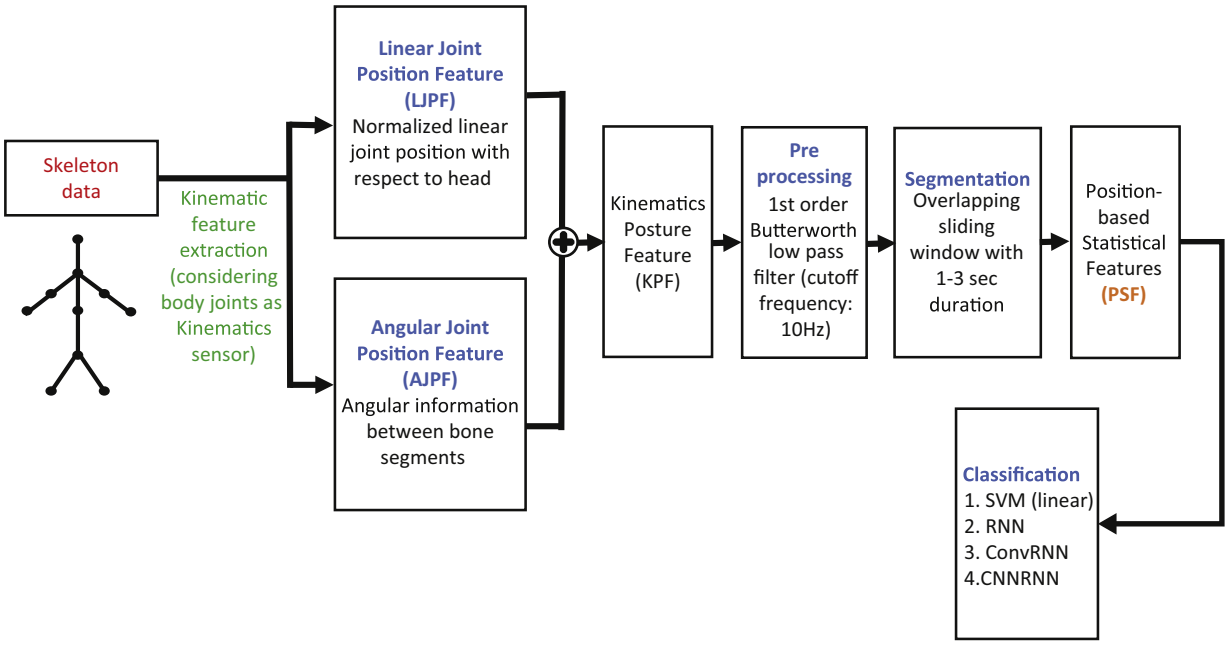

Fig. 4. Basic workflow diagram of the proposed method.

ics Posture Feature (KPF) can contain amplified noise. The main reason for this noise augmentation is the propagation of errors of the sum or difference between two time-variant amplitudes. For polishing the resultant data of the Kinematics Posture Feature (KPF) set (normalized linear joint positions and the angles between bone segments), we used the 1st order Butterworth low-pass filter with a corner frequency of 10Hz after empirical investigation. As human actions are usually restricted in the range of 0∼10Hz [23], this filtering technique increases the signal to noise ratio with minimal waveform delay.

After filtering, we extracted some useful statistical features by adopting a sliding window technique. We applied the 50% overlapping window with a variation of 1∼3sec. From the window, we extracted some statistical features namely mean, standard deviation, variance, and median absolute deviation (MAD) of the filtered signals of filtered Kinematics Posture Feature (KPF) set data. We named the statistical features extracted from filtered Kinematics Posture Feature (KPF) descriptor, as "Position-based Statistical Feature (PSF)". After separating data from the Position-based Statistical Feature (PSF) for validation and test, the rest of the data from this feature set are finally fed to the classification model for classifying the complex human actions. The system flow diagram of the entire process is demonstrated in Fig. 4.

3.5. Classification models

For classifying actions from skeleton data, we used Support Vector Machine (SVM) with Linear Kernel, ConvRNN, CNNRNN, and RNN. For the SVM linear model, we exploited the statistical features, as discussed above. However, in the deep learning model, we fed the raw segmented data into the model. We adopted the sliding window technique for the segmentation. Optimized window lengths found by cross-validation for each dataset are 0.7s, 2.5s, 3s, 0.5s, 2s for UTKinect, KARD, MSR Pair, Florence 3D, and OAD respectively. The purpose of using the RNN model is to utilize the temporal relations in input data frames. Besides, in a single data frame, skeleton joints have spatial relationships with each other. Therefore, we also implemented the ConvRNN model and CNNRNN along with LSTM. The designs of these deep learning models are depicted in Fig. 5.

For the first LSTM model, we put two LSTM cells with 100 units each. For the ConvRNN model, we designed the model with a convolution filter and LSTM cell. Before entering the LSTM cell, data have to go through a 1D convolution filter layer. In the convolution layer, there are 64 1D convolution filters with kernel size 3, and it is followed by the Relu activation function. In the CNNRNN model, we implemented maxpool operation after two convolution layers. The kernel size of the maxpool is 3 with stride 1.

We utilized the already separated validation set from Position-based Statistical Feature (PSF) sets for tuning the hyperparameters of these models. The test sets for each dataset was prepared in a similar fashion, following the existing research works to compare the performance of their proposed methods. The performance results also reported by following the same performance measures, as described in the previous works for each dataset.

4. Dataset description

In this section, we summarize five benchmark datasets for the evaluation of our proposed method.

UTKinect-Action3D Dataset: In this dataset [13], they collected 10 different classes of human action, by 10 subjects (including a left-handed person) in the indoor environment from a Kinect sensor. The classes are: walk, sit down, stand up, pick up, carry, throw, push, pull, wave, and clap hands. The dataset is challenging due to some factors. The first reason is the divergence among different understandings of the same action, for example, in order to perform the "pick up" action, some participants do this using one hand, while others prefer to utilize both hands. Another challenge was the notable variation of the action clip's duration.

Kinect Activity Recognition Dataset (KARD): KARD [10] dataset has 18 activities. Of them, 10 are gestures type (e.g., horizontal arm wave, high arm wave, high throw, draw x, draw tick, two hand wave, forward kick, side kick, hand clap, and bend), and other 8 activities (e.g., catch cap, toss paper, walk, phone call, drink, take an umbrella, sit down, and stand up). Each activity is performed 3 times by 10 subjects (9 males and 1 female).

MSR 3D Action Pairs Dataset: The Microsoft Research (MSR) 3D Action Pairs Dataset [24] contains 3D actions that are selected

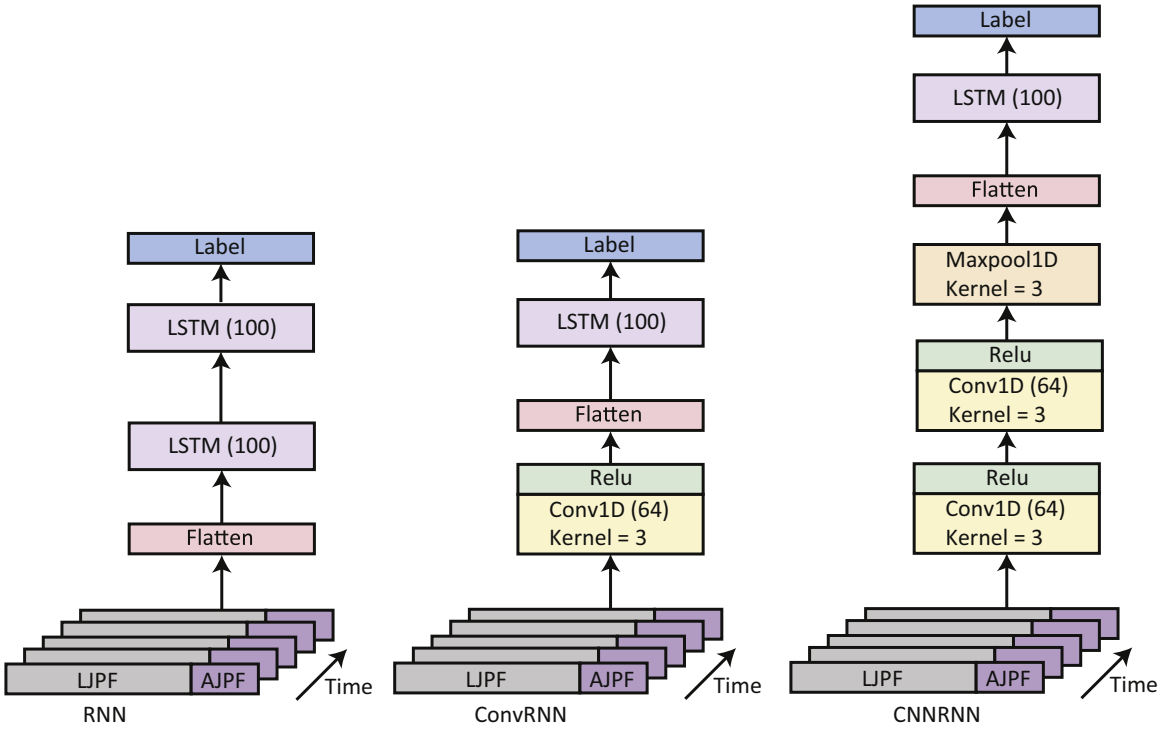

Fig. 5. Designs of the deep learning models.

in pairs, so that the 2 actions per pair have similarities in terms of motion information or similar trajectories, and similar objects. The relation between motion and shape is different in this dataset. 10 subjects acted 12 actions for 3 times. The actions are: picking up a box, put down the box, lift box, place box, push a chair, pull a chair, wear a hat, take off a hat, put on a backpack, put off a backpack, stick a poster, and remove a poster.

Florence 3D Dataset: This dataset [25] has 9 activities including wave, drink from a bottle, answer phone, clap, tight lace, sit down, stand up, read watch, and bow. 10 subjects were asked to perform each action for 2/3 times.

Office Activity Dataset (OAD): The OAD dataset has 14 activities by 10 subjects (equal genders and one left-handed person) in different office environments [26]. The classes are: drinking, getting up, pour a drink, scrolling book pages, grabbing an object from the ground, sitting, stacking items, talking on the phone, throwing something in the bin, take objects from a shelf, waving hand, wearing coat, writing on a paper, and working on computer.

5. Result and analysis

5.1. Experimental setup

We validated our proposed feature set (PSF) in five benchmark datasets by showing an ablation experiment. For proper evaluation and to prevent overfitting, we followed the "New Person" method [18]. For overcoming subject bias, we selected one subject randomly for testing purposes while we used other subjects for model training. We exploited the SVM classifier with the linear kernel for classifying the actions. Besides, we have also tried other classification methods like Random Forest, K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Naive Bayes, and Decision Tree (DT). The performance of SVM with linear kernel was better than the rest of the classifiers. This is why we have reported the results using SVM with linear kernel for the rest of the analysis. Besides, we also used deep learning-based approaches, ConvRNN, CNNRNN, and RNN to prove the generalization and robustness of our proposed feature sets. We have compared our results achieved by the "New person" setting with other existing works on the datasets.

5.2. Ablation experiment with the kinematics features

Here, we performed an ablation experiment considering subsets of feature vectors to validate the importance of our proposed LJPF and AJPF features. We performed this experiment for all of the datasets by considering the ConvRNN model. Later, we found that this model demonstrated the best performance in most of the datasets. The goal behind this experiment is to identify the importance of our proposed features for classifying different actions by utilizing skeleton data. Thus, we considered three different cases with three sets of features. We analyzed the results considering the construction of PSF features based on – (i) the combination of LJPF and AJPF features, which we named KPF features, (ii) only joint-based LJPF feature, and (iii) only angle-based AJPF feature. In Table 1, our analysis demonstrates the importance of the Kinematics Posture Feature (KPF), which is constructed based on the combination of LJPF and AJPF features.

5.3. Experimental results

Results on UTKinect-Action3D Dataset: We achieved the highest accuracy of 94.73% by using the Position-based Statistical Feature (PSF) in the "New Person" setting using ConvRNN. For the classification of this dataset, we segmented each action sequence by the 1sec sliding window technique with 50% overlapping. Our method achieved an accuracy of 94.73% by utilizing our PSF method with ConvRNN in this dataset. The comparison with existing works is presented in Table 2.

Results on the Kinect Activity Recognition Dataset (KARD): Similar to the previous dataset, the sliding window method also performed well on this dataset. For extracting PSF feature, we used a 3s sliding window with 50% overlapping. For this dataset, existing research works published their results based on precision and

Table 1
Comparison of our proposed kinematics features for classifying human actions in five benchmark datasets.

Dataset	PSF Feature Component Extracted From			Model	Comparison Parameter
	KPF (LJPF + AJPF)	LJPF	AJPF		
KARD	98.11	96.27	68.32	ConvRNN	Accuracy (%)
OAD	98.44	96.87	94.09	ConvRNN	Precision (%)
Florence	96.00	87.50	66.67	ConvRNN	Accuracy (%)
UTKinect	94.73	91.43	77.14	ConvRNN	Accuracy (%)
MSR Pair	95.45	93.94	68.18	ConvRNN	Accuracy(%)

Table 2
Accuracy comparison with existing works on UTKinect dataset.

Method	Accuracy (%)
HOJ3D [13]	90.90
Spatiotemporal features and joints fusion [27]	91.90
APJ3D and Random Forest [28]	92.00
UPCV [29]	90.95
STFC: Spatiotemporal feature chain [30]	91.50
Skeleton contexts [31]	91.90
cHCRF (Coupled hidden conditional random fields) [32]	92.00
LARP+mfPCA (manifold functional PCA) [33]	95.10
Pachinko allocation model [34]	94.80
3d-based Deep CNN [35]	96.00
Latent SVM [36]	91.50
Covariance Descriptor [37]	97.02
Decision fusion strategy with LOOCV [38]	84.00
CFM+DNM+LSTM [39]	94.36
SM+MM [40]	92.93
Modified spherical harmonics [41]	93.00
N-posture selection [12]	93.10
RA-GCN [42]	89.23
STA-LSTM [43]	79.26
ST-TR [44]	86.30
HCN [45]	90.10
HDM-BG [46]	87.50
IndRNN [47]	82.40
SVM-Linear + PSF	**93.91**
RNN + PSF	**85.96**
CNNRNN + PSF	**89.47**
ConvRNN + PSF	**94.73**

Table 3
Precision and recall comparison with existing works on KARD.

Method	Precision (%)	Recall (%)	Accuracy (%)
K-means + SVM + HMM [10]	84.80	84.50	90.83
N-posture selection (N = 7) [12]	94.00	93.70	-
N-posture selection (N = 15) [12]	95.10	95.00	-
Fusion of features + multiclass SVM [48]	-	-	99.31
Depth data + SVM [49]	-	-	96.64
Skeleton data + ConvNet [50]	-	-	98.50
SVM-Linear + PSF	**97.43**	**97.61**	**97.51**
RNN + PSF	**96.83**	**97.22**	**96.27**
CNNRNN + PSF	**96.83**	**97.22**	**96.27**
ConvRNN + PSF	**98.22**	**98.25**	**98.11**

recall. By utilizing ConvRNN along with our proposed PSF feature, we achieved the highest accuracy of 98.11%, precision of 98.22%, and recall of 98.11%. In Table 3, we have shown a comparison between our results and other existing results.

Results on MSR 3D Action Pairs Dataset: This dataset is challenging because of the similar motion information among action classes. For this reason, we segmented an action with a higher window duration, so that we can get as much information as needed in a single window. For the segmentation process, we chose a sliding window with a duration of 3s for extracting the PSF feature set. In Table 4, we showed the comparison between our results and existing works on this dataset. We achieved the highest 95.45% accuracy using ConvRNN and our proposed PSF feature set.

Table 4
Accuracy comparison with existing works on MSR Pair dataset.

Method	Accuracy (%)
Skeleton + LOP [19]	63.33
DMM + HOG [51]	66.11
Skeleton + LOP + Pyramid [19]	82.22
HON4D [52]	93.33
Lie Group + SVM [53]	93.33
Lie Group + CNN [54]	93.68
Spatiotemporal feature + SVM [55]	94.50
SVM-Linear + PSF	**74.78**
RNN + PSF	**86.36**
CNNRNN + PSF	**92.42**
ConvRNN + PSF	**95.45**

Table 5
Accuracy comparison with existing works on Florence dataset.

Method	Accuracy (%)
Multi-part bag-of-poses [25]	82.00
Trajectory Riemannian manifold + kNN [56]	87.04
LARP [57]	90.90
LARP+mfPCA (manifold functional PCA) [33]	89.70
N-posture selection [12]	84.70
Pachinko allocation model [34]	90.23
Latent maxmargin multitask learning [58]	93.42
Covariance Descriptor [37]	91.00
CFM+DNM+LSTM [39]	94.36
RA-GCN [42]	81.36
STA-LSTM [43]	80.29
ST-TR [44]	85.20
HCN [45]	89.30
HDM-BG [46]	91.30
IndRNN [47]	85.60
Lie Group + CNN [54]	93.00
SVM + PSF	**71.58**
RNN + PSF	**91.66**
ConvRNN + PSF	**96.00**

Table 6
Accuracy comparison with existing works on OAD.

Method	Precision (%)	Recall (%)
Joint orientations [26]	80.85	80.86
Skeleton approach [21]	80.60	80.50
RGB (20 sectors) [21]	85.80	85.90
Score-level fusion of RGB & skeleton [21]	90.60	90.40
SVM + PSF	**94.92**	**92.14**
CNNRNN + PSF	**92.37**	**91.25**
ConvRNN + PSF	**98.44**	**97.25**
RNN + PSF	**95.34**	**88.69**

Results on Florence 3D Dataset: Due to the very short duration of each action in this dataset, we segmented the dataset using the 1sec sliding window. Because of the large inter-class correlation, high inter-class variability, and short duration of actions, this dataset is a challenging dataset. Despite the presence of transitory actions, our proposed PSF feature set with ConvRNN obtained the highest 96% accuracy. The comparison of our method with other works is shown in Table 5.

Results on Office Activity Dataset (OAD): To process this dataset, we used a sliding window (duration 2sec and 50% overlapping) for extracting PSF feature set. We achieved the highest precision of 95.34% and recall of 92.14% by using the Position-based Statistical Feature (PSF) in the "New Person" setting using RNN architecture and SVM model respectively. In Table 6, we compared our method with the existing works on this dataset. From the Table, we can infer that the activities in this dataset can be differentiated more accurately by the PSF.

6. Conclusions

Skeleton-based action recognition has flourished recently, especially due to the advent of low-cost Kinect sensors and OpenPose. In this work, we classified the daily actions by extracting Kinematics Posture Feature (KPF) from skeleton data, which combines Linear Joint Position Feature (LJPF) and Angular Joint Position Feature (AJPF) to encode motion while performing different actions. This method separates the actions based on a feature set, that focuses on normalized joint positions with respect to the head, and angles between the bone segments. The change of joint position and angular information across video frames from these KPF feature sets can also encode the variation in posture. In our proposed method, we considered the body joints as kinematics sensors (e.g., position and orientation sensors). Thus, we processed and segmented the extracted Kinematics Posture Features (KPF) by exploiting filtering techniques and sliding window-based approach with optimized window length. Afterward, we computed Position-based Statistical Feature (PSF) to train SVM (linear), RNN, CNNRNN, and ConvRNN models. For evaluation, we applied our method on five benchmark datasets. We obtained excellent or comparable results for these publicly available benchmark datasets. Our proposed Position-based Statistical Feature (PSF) takes into account the primary posture and the key directional snippets of action, which outperformed more complex existing algorithms. It is indeed rare in the literature to find a method that is suitable for 5 benchmark datasets. However, we would like to explore 2-persons interactions and multi-view cases in the future.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M.A.R. Ahad, Vision and Sensor-based Human Activity Recognition: Challenges Ahead, in: Advancements in Instrumentation and Control in Applied System Applications, IGI Global, 2020, pp. 17–35.
[2] M.A.R. Ahad, A.D. Antar, O. Shahid, Vision-based action understanding for assistive healthcare: A short review, in: IEEE CVPR Workshops 2019, 2019, pp. 1–11.
[3] J. Wang, Z. Liu, Y. Wu, J. Yuan, Learning actionlet ensemble for 3D human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 36 (5) (2014) 914–927.
[4] S.X. Wang, Current trends in computer science and mechanical automation vol. 1: Selected papers from CSMA2016, Walter de Gruyter GmbH & Co KG, 2018.
[5] J. Cai, X. Tang, RGB video based tennis action recognition using a deep weighted long short-term memory, arXiv preprint arXiv:1808.00845 (2018).
[6] J. Zang, L. Wang, Z. Liu, Q. Zhang, G. Hua, N. Zheng, Attention-based temporal weighted convolutional neural network for action recognition, in: IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, 2018, pp. 97–108.
[7] S. Wang, G. Zhou, A review on radio based activity recognition, Digital Communications and Networks 1 (1) (2015) 20–29.
[8] L. Atallah, B. Lo, R. Ali, R. King, G.-Z. Yang, Real-time activity classification using ambient and wearable sensors, IEEE Trans. Inf. Technol. Biomed. 13 (6) (2009) 1031–1039.
[9] T. Hossain, M.S. Islam, M.A.R. Ahad, S. Inoue, Human activity recognition using earable device, in: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, 2019, pp. 81–84.
[10] S. Gaglio, G.L. Re, M. Morana, Human activity recognition process using 3-D posture data, IEEE Trans. Hum. Mach. Syst. 45 (5) (2015) 586–597.
[11] J. Shan, S. Akella, 3D human action segmentation and recognition using pose kinetic energy, in: 2014 IEEE international workshop on advanced robotics and its social impacts, IEEE, 2014, pp. 69–75.
[12] E. Cippitelli, S. Gasparrini, E. Gambi, S. Spinsante, A human activity recognition system using skeleton data from RGBD sensors, Comput. Intell. Neurosci. 2016 (2016) 21.
[13] L. Xia, C.-C. Chen, J.K. Aggarwal, View invariant human action recognition using histograms of 3D joints, in: IEEE CVPR Workshops, IEEE, 2012, pp. 20–27.
[14] X. Yang, Y. Tian, Effective 3D action recognition using eigenjoints, J. Vis. Commun. Image Represent. 25 (1) (2014) 2–11.
[15] L. Piyathilaka, S. Kodagoda, Gaussian mixture based hmm for human daily activity recognition using 3D skeleton features, in: 2013 IEEE 8th conference on industrial electronics and applications (ICIEA), IEEE, 2013, pp. 567–572.
[16] D.R. Faria, C. Premebida, U. Nunes, A probabilistic approach for human everyday activities recognition using body motion from RGB-D images, in: The 23rd IEEE international symposium on robot and human interactive communication, IEEE, 2014, pp. 732–737.
[17] Y. Zhu, W. Chen, G. Guo, Evaluating spatiotemporal interest point features for depth-based action recognition, Image Vis. Comput. 32 (8) (2014) 453–464.
[18] J. Sung, C. Ponce, B. Selman, A. Saxena, Unstructured human activity detection from RGBD images, in: 2012 IEEE international conference on robotics and automation, IEEE, 2012, pp. 842–849.
[19] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: 2012 IEEE CVPR, IEEE, 2012, pp. 1290–1297.
[20] H.S. Koppula, R. Gupta, A. Saxena, Learning human activities and object affordances from RGB-D videos, Int. J. Rob. Res. 32 (8) (2013) 951–970.
[21] A. Franco, A. Magnani, D. Maio, A multimodal approach for human activity recognition based on skeleton and RGB data, Pattern Recognit. Lett. (2020).
[22] F. Battistone, A. Petrosino, TGLSTM: a time based graph deep learning approach to gait recognition, Pattern Recognit. Lett. 126 (2019) 132–138.
[23] L. Bao, S.S. Intille, Activity recognition from user-annotated acceleration data, in: International conference on pervasive computing, Springer, 2004, pp. 1–17.
[24] O. Oreifej, Z. Liu, HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences, in: IEEE CVPR, IEEE, 2013, pp. 716–723.
[25] L. Seidenari, V. Varano, S. Berretti, A. Bimbo, P. Pala, Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses, in: IEEE CVPR Workshops, 2013, pp. 479–485.
[26] A. Franco, A. Magnani, D. Maio, Joint orientations from skeleton data for human activity recognition, in: International Conference on Image Analysis and Processing, Springer, 2017, pp. 152–162.
[27] Y. Zhu, W. Chen, G. Guo, Fusing spatiotemporal features and joints for 3D action recognition, in: IEEE CVPR Workshops, IEEE, 2013, pp. 486–491.
[28] L. Gan, F. Chen, Human action recognition using APJ3D and random forests, Journal of Software 8 (9) (2013) 2238–2246.
[29] I. Theodorakopoulos, D. Kastaniotis, G. Economou, S. Fotopoulos, Pose-based human action recognition via sparse representation in dissimilarity space, J. Vis. Commun. Image Represent. 25 (1) (2014) 12–23.
[30] W. Ding, K. Liu, F. Cheng, J. Zhang, Stfc: spatio-temporal feature chain for skeleton-based human action recognition, J. Vis. Commun. Image Represent. 26 (2015) 329–337.
[31] M. Jiang, J. Kong, G. Bebis, H. Huo, Informative joints based human action recognition using skeleton contexts, Signal Process. Image Commun. 33 (2015) 29–40.
[32] A.-A. Liu, W.-Z. Nie, Y.-T. Su, L. Ma, T. Hao, Z.-X. Yang, Coupled hidden conditional random fields for RGB-D human action recognition, Signal Processing 112 (2015) 74–82.
[33] R. Anirudh, P. Turaga, J. Su, A. Srivastava, Elastic functional coding of human actions: From vector-fields to latent variables, in: IEEE CVPR, 2015, pp. 3147–3155.
[34] T. Huynh-The, B.-V. Le, S. Lee, Describing body-pose feature-poselet-activity relationship using pachinko allocation model, in: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2016, pp. 000040–000045.
[35] Z. Liu, C. Zhang, Y. Tian, 3D-based deep convolutional neural network for action recognition with depth sequences, Image Vis. Comput. 55 (2016) 93–100.
[36] X. Li, Y. Zhang, D. Liao, Mining key skeleton poses with latent svm for action recognition, Applied Computational Intelligence and Soft Computing 2017 (2017).
[37] H.A. El-Ghaish, A.A. Shoukry, M.E. Hussein, CovP3DJ: Skeleton-parts-based-covariance descriptor for human action recognition., in: VISIGRAPP (5: VISAPP), SciTePress, 2018, pp. 343–350.
[38] D. Avola, M. Bernardi, G.L. Foresti, Fusing depth and colour information for human action recognition, Multimed. Tools Appl. 78 (5) (2019) 5919–5939.
[39] G. Huang, Q. Yan, Optimizing features quality: a normalized covariance fusion framework for skeleton action recognition, IEEE Access 8 (2020) 211869–211881.
[40] M. Liu, Q. He, H. Liu, Fusing shape and motion matrices for view invariant action recognition using 3D skeletons, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 3670–3674.

[41] C. Youssef, et al., Spatiotemporal representation of 3D skeleton joints-based action recognition using modified spherical harmonics, Pattern Recognit. Lett. 83 (2016) 32–41.
[42] Y.-F. Song, Z. Zhang, L. Wang, Richly activated graph convolutional network for action recognition with incomplete skeletons, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 1–5.
[43] Y. Ding, Y. Zhu, Y. Wu, F. Jun, Z. Cheng, Spatio-temporal attention LSTM model for flood forecasting, in: 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), IEEE, 2019, pp. 458–465.
[44] C. Plizzari, M. Cannici, M. Matteucci, Spatial temporal transformer network for skeleton-based action recognition, arXiv preprint arXiv:2008.07404 (2020).
[45] C. Li, Q. Zhong, D. Xie, S. Pu, Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation, arXiv preprint arXiv:1804.06055 (2018).
[46] R. Zhao, W. Xu, H. Su, Q. Ji, Bayesian hierarchical dynamic model for human action recognition, IEEE CVPR, IEEE, 2019, pp. 7733–7742.
[47] S. Li, W. Li, C. Cook, C. Zhu, Y. Gao, Independently recurrent neural network (IndRNN): Building a longer and deeper RNN, IEEE CVPR, IEEE, 2018, pp. 5457–5466.
[48] J. Ling, L. Tian, C. Li, 3D human activity recognition using skeletal data from RGBD sensors, in: International Symposium on Visual Computing, Springer, 2016, pp. 133–142.
[49] C. Dhiman, D.K. Vishwakarma, A robust framework for abnormal human action recognition using transform and Zernike moments in depth videos, IEEE Sens. J. 19 (13) (2019) 5195–5203.
[50] N.E. El Madany, Y. He, L. Guan, Integrating entropy skeleton motion maps and convolutional neural networks for human action recognition, in: 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2018, pp. 1–6.
[51] X. Yang, C. Zhang, Y. Tian, Recognizing actions using depth motion maps-based histograms of oriented gradients, in: Proceedings of the 20th ACM international conference on Multimedia, ACM, 2012, pp. 1057–1060.
[52] O. Oreifej, Z. Liu, HON4d: Histogram of oriented 4D normals for activity recognition from depth sequences, in: IEEE CVPR, IEEE, 2013, pp. 716–723.
[53] R. Vemulapalli, F. Arrate, R. Chellappa, R3DG Features: relative 3D geometry-based skeletal representations for human action recognition, Comput. Vision Image Understanding 152 (2016) 155–166.
[54] L. Cai, C. Liu, R. Yuan, H. Ding, Human action recognition using lie group features and convolutional neural networks, Nonlinear Dyn. (2020) 1–11.
[55] D.-T. Pham, T.-N. Nguyen, T.-L. Le, H. Vu, Spatio-temporal representation for skeleton-based human action recognition, in: 2020 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), IEEE, 2020, pp. 1–6.
[56] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, A. Del Bimbo, 3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold, IEEE Trans. Cybern. 45 (7) (2014) 1340–1352.
[57] R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3D skeletons as points in a lie group, in: IEEE CVPR, IEEE, 2014, pp. 588–595.
[58] Y. Yang, C. Deng, D. Tao, S. Zhang, W. Liu, X. Gao, Latent max-margin multitask learning with skelets for 3-D action recognition, IEEE Trans. Cybern. 47 (2) (2016) 439–448.