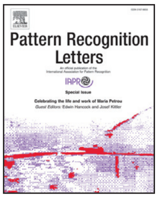
Pattern Recognition Letters 87 (2017) 195–202

Contents lists available at ScienceDirect

Pattern Recognition Letters

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



Graph-based approach for 3D human skeletal action recognition[☆]

Meng Li[*], Howard Leung

Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Available online 3 August 2016

Keywords:
Action recognition
Graph-based method
Kernel matching

ABSTRACT

Human action recognition is a challenging task due to the articulated and complex nature of actions. Recently developed commodity depth sensors coupled with the skeleton estimation algorithm have generated a renewed interest in human skeletal action recognition. In this paper, we characterize the human actions with a novel graph-based model which preserves complex spatial structure among skeletal joints according to their activity levels as well as the spatio-temporal joint features. In particular, the proposed top-K Relative Variance of Joint Relative Distance (RVJRD)s determine which joint pairs should be selected in the resulting graph according to normalized activity levels. In addition, the temporal pyramid covariance descriptors are adopted to represent joint locations. The graph kernel is used for measuring the similarity between two graphs by matching the walks from each of the two graphs to be matched. We evaluate the proposed approach on three challenging action recognition datasets captured by depth sensors. The experimental results show our proposed approach outperforms several state-of-the-art human skeletal action recognition approaches.

1. Introduction

Human action recognition has many applications in surveillance, video games, sports video analysis, robotics, etc. Despite significant research efforts over the past few decades, action recognition still remains a challenging problem due to the complex articulated essence of human movements.

It is generally agreed that knowing the 3D joint positions is helpful for action recognition. Recently introduced cost-effective depth sensors coupled with the real-time skeleton estimation algorithm of [24] largely ease the task of action recognition. With the 3D locations of skeletal joints in the scene, action recognition can be performed by classifying their variations over time. These recent advances have resulted in a renewed interest in 3D human skeletal action recognition.

Most of the existing skeletal action recognition approaches consider the human skeleton as either a set of points or a connected set of rigid segments. These methods do not scale very well for skeletal action recognition as they do not take into account the complex structure among different body parts which can reflect the spatial variations of human actions.

Graphs are extremely effective in modeling the complex structured data as reported in Borgwardt et al. [6] and Gauzere et al. [14]. However, not many researches construct the graphs to model the complex spatial structures of joints for skeletal action representation. There are two main reasons: (1) It is a challenging task to characterize the large spatial variations of joints in actions. (2) It is a challenging task to compute the similarities of the structure, the vertex attributes and edge attributes between two constructed graphs. In this paper, we focus on these two tasks, and propose a novel graph-based approach for human skeletal action recognition.

We first construct our proposed graph model based on spatial variations of human action, where the vertices of each graph represent the skeletal joints in a human body, the edges represent the connections of joint pairs. The weights of edges reflect the importance of the corresponding joint pairs for certain action. In our consideration, inspired by the observation of human skeletal actions, the relative spatial variations between various skeletal joints (though not directly connected by a bone) provides a more meaningful description than their absolute movements (clapping is more intuitively described using the relative spatial variations between the two hand joints). In view of this, we construct a graph-based model to represent the relative spatial structure among various joints in our action recognition approach. There are a large number of skeletal joint pairs in a human body, but a given action is usually only associated with and characterized by a subset of those joints. If we want to determine which skeletal joint pairs are more important according to their activity levels

[☆] "This paper has been recommended for acceptance by Cheng-Lin Liu".

[*] Corresponding author.
E-mail address: mli269-c@my.cityu.edu.hk (M. Li).

http://dx.doi.org/10.1016/j.patrec.2016.07.021
0167-8655/© 2016 Elsevier B.V. All rights reserved.

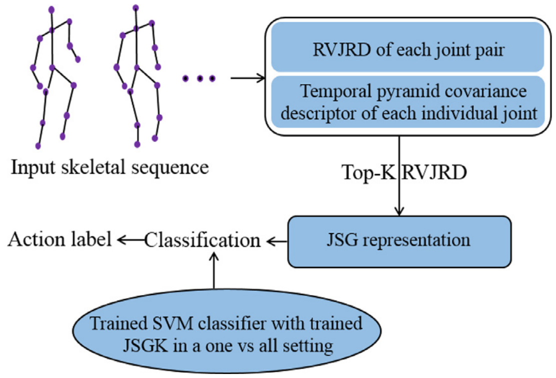

RVJRD of each joint pair
Temporal pyramid covariance descriptor of each individual joint
Input skeletal sequence
Top-K RVJRD
JSG representation
Action label
Classification
Trained SVM classifier with trained JSGK in a one vs all setting

Fig. 1. Flow chart of the proposed approach.

over the duration of the human movement, the Variance of Joint Relative Distance (VJRD) proposed by Tang and Leung [27] may not give a fair ranking because it is more likely for too farther away skeletal joints to give larger VJRD values. As a result, we use the Relative Variance of Joint Relative Distance (RVJRD) which indicates the normalized activity levels among the joint pairs. We rank the joint pairs according to this measure to obtain the top-K RVJRDs which indicates the joint pairs with the largest normalized spatial variations. Based on the top-K RVJRDs, we propose the Joint Spatial Graph (JSG) to model 3D human skeletal actions.

Then, we use the proposed spatio-temporal joint relative location features and the values of RVJRDs to attribute the proposed JSG. The vertex attributes in JSG are the the temporal pyramid covariance descriptors on 3D joint relative locations. The edge attributes in JSG are the edge weights, i.e., the values of RVJRDs. Hence, the proposed JSG is made adaptively to capture the spatio-temporal pattern of a given action. Compared with other existing skeletal action recognition methods, our approach can preserve the complex spatial structure among joints as well as the spatio-temporal features of individual joints.

In order to compute the similarity of the complex structures, the vertex attributes and the edge attributes between two JSGs, we propose the JSG Kernel (JSGK) which is a bridge between the structured action representation and the similarity measurement. The proposed JSGK is actually a series of decomposition kernels. That is, the JSG is first decomposed into a number of the walk groups with different lengths, and the JSGK is then obtained by matching the walk groups of the two graphs based on the walk graph kernels. A multiple kernel learning method is applied for obtaining the optimal JSGK. Final multi-class classification task is performed using usual SVM classifier with JSGK in a one-versus-all setting as adopted in Shawe-Taylor and Cristianini [23]. Fig. 1 presents an overview of the proposed approach.

The proposed approach is evaluated on three benchmark datasets: MSR-Action3D dataset created by Li et al. [19], UTKinect-Action dataset created by Xia et al. [32] and Florence3D-Action dataset created by Seidenari et al. [22]. We also show that the proposed approach outperforms several state-of-the-art human skeletal action recognition methods.

Our main contributions include the following three aspects. First, we propose the JSG model as a new way of characterizing 3D human skeletal actions. Second, we propose the JSGK to measure the similarity of two JSGs. The JSGK performs this measurement in three aspects, including the structures, the vertex attributes and edge attributes of the two JSGs. Third, the proposed temporal pyramid covariance descriptor is a new representation of spatio-temporal pattern to preserve the individual joint features.

After a brief review of the related work in Section 2, the proposed JSG is described in Section 3. Section 4 presents the JSGK. The action recognition method using JSGK is given in Section 5. Section 6 applies the multiple kernel learning for obtaining the optimal JSGK. In Section 7, we analyse the runtime complexity of our approach. In Section 8, we conduct the experimental analysis and evaluation of our approach. In Section 9, we give a conclusion and future work.

2. Related work

In this section, we briefly review various human action recognition approaches. Our review mainly focus on the skeletal action recognition. The readers are referred to [1] and [35] for a systematic review of approaches based on video and depth map, respectively.

In the existing approaches for human skeletal action recognition, one way to represent the human skeleton in actions is to consider human skeleton as a set of points. Wang et al. [31] represented a human skeleton using pairwise relative positions of the joints, and the temporal evolutions of this representation were modeled using a hierarchy of Fourier coefficients. Furthermore, an actionlet-based approach was used to select discriminative joint combinations by using a multiple kernel learning approach. Yang and Tian [33] represented a human skeleton using relative joint positions, temporal displacement of joints and offset of the joints with respect to the initial frame. They performed action classification based on the Naive-Bayes nearest neighbor rule in a lower dimensional space constructed using Principal Component Analysis (PCA). In Xia et al. [32], human skeleton was represented by histograms of 3D joint locations in a frame, the temporal evolutions of this view-invariant representation were modeled using HMMs. In Gowayyed et al. [16], histograms of Oriented Displacements (HOD) were used to represent the 3D trajectories of skeletal joints for human action recognition. Du et al. [11] proposed an end-to-end hierarchical recurrent neural network (RNN) for the skeletal representation construction, where the raw locations of skeletal joints are directly used as the input to the RNN. Zhu et al. [36] used raw 3D skeletal joint locations as the input to a RNN with Long Short-Term Memory (LSTM) to learn human actions.

Another way for representing the human skeleton is to consider human skeleton as a connected set of rigid segments. For example, Chaudhry et al. [8] divided a human skeleton into smaller parts and represented each part using certain bio-inspired shape features. They modeled the temporal evolutions of these bio-inspired features using linear dynamical systems. Since joint angles measure the geometry between directly connected pairs of body parts, they can also be considered as a set of skeleton segments. Ofli et al. [20] selected the informative skeletal joints at each time instance based on highly interpretable measures such as mean or variance of the joint angles. They represented human actions as sequences of these informative joints, which were compared using the Levenshtein distance. Ohn-Bar and Trivedi [21] represented skeletal sequences using pairwise affinities between joint angle trajectories, and then performed classification task by linear SVM. In Vemulapalli et al. [29], human actions were modeled as curves in a Lie group, then a linear SVM is used for recognition task. Amor et al. [2] modeled human actions using shape evolutions of skeletons on Kendalls shape manifold, and then use an SVM classifier to obtain action recognition results.

Different from these approaches, we propose Joint Spatial Graph (JSG) to model the relative spatial structure among various joints with temporal pyramid covariance descriptors in our skeletal representation. The proposed JSG is a descriptive model. It can model the complex articulated structure of a human action without complicated learning as in the generative models like HMM.

Graphs are an extremely effective tool modeling complex structured objects. Recently, many graph-based methods are applied for

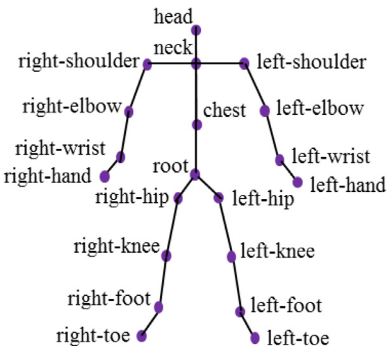


Fig. 2. Skeletal joints captured by the depth sensor.

human action recognition on video data. In Borzeshi et al.[7], each frame is represented as a graph with vertices corresponding to the spatial local features extracted from this frame. In Gaur et al. [13], a string of feature graphs are constructed for the spatio-temporal layout of local features. Each graph in the string models the spatial configuration of local features in a small temporal segments. In Ta et al. [26], a hypergraph is constructed to model the extracted spatiotemporal local features and a hypergraph matching algorithm is used for activity recognition. In Aoun et al. [3], a graph-based video representation is used for modeling the spatio-temporal relations among the commonly used Space-Time Interest Points (STIP) local features. These approaches construct graphs to model local features in video data. However, when we want to use data captured by depth sensors, since there is no texture in the depth data, these local features are not suitable. Our work is the first attempt to use graph to model the spatial structures among joints for skeletal action recognition on depth data, and achieves better performance than several state-of-the-art human skeletal action recognition approaches on multiple benchmark datasets.

3. Skeletal action representation

3.1. Relative variance of joint relative distance

Suppose that the human body is represented by N joints, the illustration of an example skeleton with 20 joints are shown in Fig. 2. The skeletal action $A = \{A_1, A_2, \ldots, A_T\}$ is performed over T frames. Let $J_i(t) = (x_i(t), y_i(t), z_i(t))$ $(1 \le i \le N)$ be the ith joint at the tth $(1 \le t \le T)$ frame, and each frame contain P joint pairs. The N joints at tth frame are depicted as $\{J_1(t), J_2(t), \ldots, J_N(t)\}$. The Joint Relative Distance (JRD) [27] of the pth $(1 \le p \le P)$ joint pair at the tth frame is calculated by pairwise Euclidean distances (i.e., L_2-norm) between two joints J_i and J_j at frame t:

$$\mathrm{JRD}(t,p) = \mathrm{d}_{L_2}(J_i(t), J_j(t)). \tag{1}$$

The relative variance of each JRD over the duration of the movement is considered to characterize the action, and we name it as RVJRD. The value of RVJRD measures the importance of variations of each JRD for description of actions. Specifically, the K pairs of joints in the top-K RVJRDs according to descending order and the structure of these joints can reflect the main characteristic of the actions. We denote the mean of $\mathrm{JRD}_A(t, p)$ over T frames be $\overline{\mathrm{JRD}_A(p)}$, the RVJRD of the pth joint pair is formulated as:

$$\mathrm{RVJRD}_A(p) = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\mathrm{JRD}_A(t,p) - \overline{\mathrm{JRD}_A(p)}}{\overline{\mathrm{JRD}_A(p)}}\right)^2. \tag{2}$$

3.2. Joint spatial graph

Given a skeletal action A, a Joint Spatial Graph (JSG) based on the top-K RVJRDs is constructed to model the spatial layout relationships of the adaptively selected set of skeletal joints. This JSG is denoted as $G = (V, E, H)$, where V is the vertex set which represents the adaptively selected set according to the top-K RVJRDs of action A, E is the edge set which describes the joints' spatial layout relationships based on the top-K RVJRDs in action A, and K is the number of edges. $H \in R^{n\times n}$ is the affinity matrix of this graph, where n is the number of the vertices of the graph. We apply the ε-Graph model to construct G:

$$H(i,j) = \begin{cases} 1 & \text{if } \mathrm{RVJRD}_A(p) > \varepsilon, \\ 0 & \text{otherwise}. \end{cases} \tag{3}$$

The parameter $\varepsilon > 0$ is the threshold of the RVJRDs of joint pairs in the action A, and its value is set according to the top-K RVJRDs of the action A in descending order. When $H(i,j) = 1$, it means $(v_i, v_j) \in E$. Hence, it is clear that this graph is constructed entirely depending on the spatial variations of the action A. Moreover, we propose the temporal pyramid covariance joint descriptors to attribute the vertices (see Section 5), and use the values of the RVJRDs to attribute edges.

In the end, each action is adaptively represented by the special attributed graph based on the top-K RVJRDs. This representation has two properties. First, the graph is adaptively constructed for each action to reflect its special spatial structure of joints. Second, the graph preserves the spatio-temporal features of the joints. The structures of JSGs for several skeletal actions are shown in Fig. 3.

Fig. 3 shows several spatial structures of JSGs according to skeletal actions in three action datasets. In each subfigure, the purple points connected with others by blue lines represent the vertices of graph. The blue lines between the purple points represent the edges of the graph. It is clear that for the same type of actions (e.g., two forward punch actions), their JSGs have similar structures, and for the different types of actions, their JSGs have different structures. Hence, our proposed JSG representations of actions are very discriminative.

4. Joint spatial graph kernel

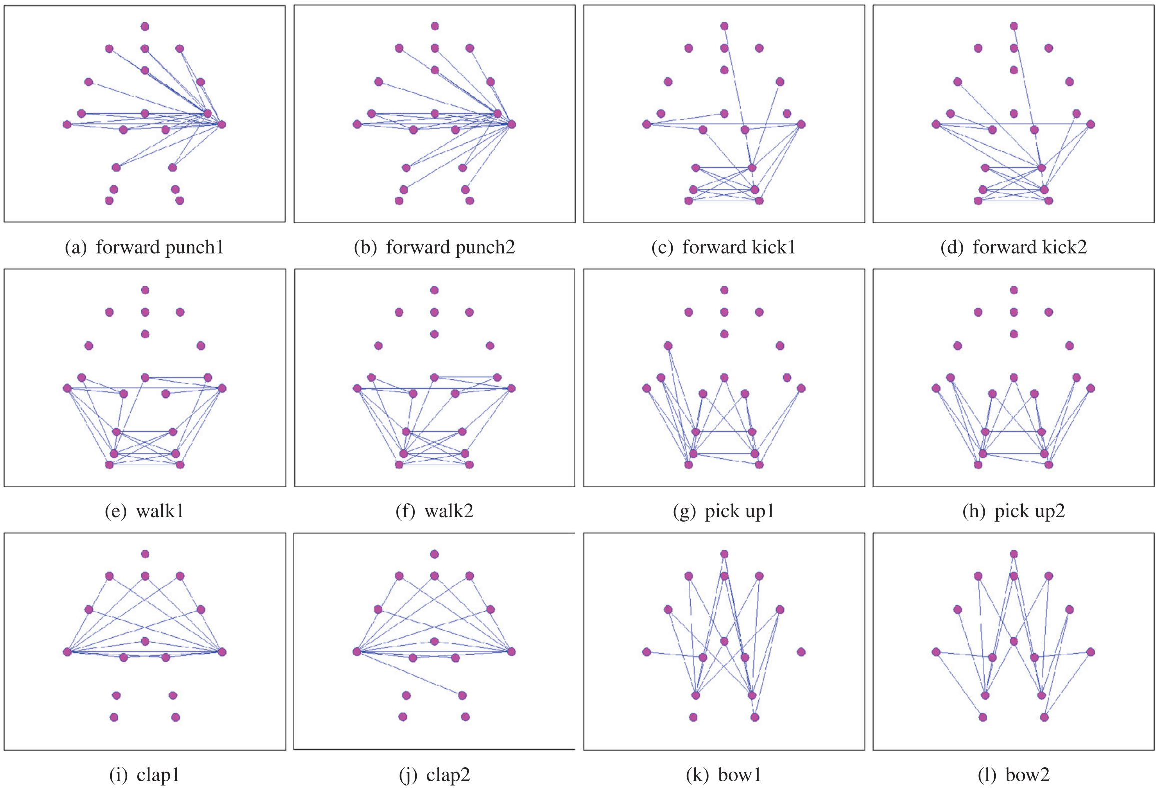
Given a skeletal action, let $G = (V, E)$ be its JSG with n vertices, where $V = \{v_i\}_{i=1}^{n}$ is the vertex set and E is the edge set. A vertex represents a joint selected by top-K RVJRDs. Inspired by Hussein et al. [18], we use the sample covariance matrix (see Section 5) instead. An edge represents the connection of a pair of joints selected by the top-K RVJRDs. The attribute of this edge is the variations of JRD between the pair of joints in the action which is represented by the value of RVJRD.

In order to measure the similarity between two joint spatial graphs, we should evaluate the similarity of their structures, their vertex attributes and edge attributes. Hence, we propose an Joint Spatial Graph Kernel (JSGK) to perform this evaluation.

We first decompose the two JSGs into a number of the walk groups, and then JSGK is obtained by the walk group matching.

The decomposition of our JSG is based on walks. A walk with length q in graph G is defined as a sequence of vertices connected by edges, $w = (v_{w_0}, e_{w_1}, \ldots, e_{w_q}, v_{w_q})$, where $e_{w_i} = (v_{w_{i-1}}, v_{w_i}) \in E$, $1 \le i \le q$. Considering the problem of the tottering and halting, we just focus on the walks with different vertices in the JSG. Hence, these walks are actually paths. Let φ_G^q be the set of total walks with length q in the JSG, and $\varphi_G^q(i,j) \subset \varphi_G^q$ be a walk group, which is a subset of φ_G^q which contains walks starting at vertex v_i and ending at v_j. It means that for a walk $w \in \varphi_G^q(i,j)$, we have $v_{w_0} = v_i$ and $v_{w_q} = v_j$. Hence, we can see that a walk w with length $q = 0$ is a vertex. Hence, we have $\varphi_G^0 = V$, and $\varphi_G^0(i,i) = v_i$. The walk group can be regarded as subgraph of the JSG.

Given two actions, let $G = (V, E)$ and $G' = (V', E')$ be their JSGs, where $k_v(v, v')$ and $k_e(e, e')$ are defined as the vertex kernel and


(a) forward punch1 (b) forward punch2 (c) forward kick1 (d) forward kick2

(e) walk1 (f) walk2 (g) pick up1 (h) pick up2

(i) clap1 (j) clap2 (k) bow1 (l) bow2

Fig. 3. Spatial structures of several JSGs according to skeletal actions in three datasets. The first row corresponds to actions represented by JSGs based on top-20 RVJRD in MSR-Action3D dataset. The second row corresponds to actions represented by JSGs based on top-20 RVJRD in UTKinect-Action dataset. The third row corresponds to actions represented by JSGs based on top-15 RVJRD in Florence3D-Action dataset.

edge kernel on the two JSGs, respectively, where the vertices v and v' in $k_v(v, v')$ represent the joints of human body. $k_v(v, v')$ and $k_e(e, e')$ are designed for measuring the similarity of the vertex attributes and edge attributes. The specific formulations of the $k_v(v, v')$ and $k_e(e, e')$ are designed in Section 5. Let $k_w(w, w')$ be defined as the walk kernel on two walks with the same length. For the case $q = 0$, we have $k_w\big(w, w'\big) = k_v\big(v_w, v'_{w'}\big)$. For the case $q \geq 1$, we have

$$k_w\big(w, w'\big) = \prod_{i=0}^{q} k_v(v_{w_i}, v'_{w'_i}) \prod_{j=1}^{q} k_e(e_{w_j}, e'_{w'_j}). \tag{4}$$

In the two JSGs G and G', the kernels on two walk groups with length q are defined as a summation of walk kernels on all matched walk pairs from the two walk groups with length q, respectively:

$$k_g^q(\varphi_G^q(i, j), \varphi_{G'}^q(r, s)) = \sum_{\substack{w \in \varphi_G^q(i, j) \\ w' \in \varphi_{G'}^q(r, s)}} k_w\big(w, w'\big). \tag{5}$$

The proposed JSGK on two JSGs are computed as a summation of similarities between all pairs of groups in two JSGs. For the two JSGs G and G', we denote the JSGK with walk length q, as the qth order JSGK which is defined as

$$k_G^q\big(G, G'\big) = \frac{1}{M_{GG'}^q} \sum_{\substack{\varphi_G^q(i, j) \subset \varphi_G^q \\ \varphi_{G'}^q(r, s) \subset \varphi_{G'}^q}} k_g^q(\varphi_G^q(i, j), \varphi_{G'}^q(r, s)), \tag{6}$$

where $M_{GG'}^q$ is the number of matched walk pairs in walk groups with length q in G and G'. With a series of weight coefficients (λ_0, λ_1, λ_2, ...) to emphasize the importance of each order $k_G^q\big(G, G'\big)$, the final JSGK is calculated as a weighted summation of different qth order JSGKs:

$$k_G\big(G, G'\big) = \sum_{q=0} \lambda_q k_G^q(G, G'). \tag{7}$$

From Eq. (7), we can see that if all the $k_v(v, v')$, $k_e(e, e')$ and λ_q equal to one, JSGK will be simplified to count the number of matched walks which is a graph kernel proposed by Gärtner et al.[12]. In this case, the vertex attributes and edge attributes in the two JSGs are exactly the same, and JSGK actually just measures the similarity of the structures for the two JSGs.

5. Action recognition based on JSGK

We further apply our JSGK for 3D human skeletal action recognition. Hence, we should define the vertex kernel and edge kernel.

As mentioned above, the vertex kernel is designed for measuring the similarity of the vertex attributes. The vertex attribute is actually the probability distribution of joint location. Instead of using absolute coordinates, we use Laplacian coordinates as adopted in Sorkine et al. [25] to represent the joint locations for encoding spatial relationship among the joints in each frame of action.

Laplacian coordinates is known to describe the local details using the adjacent vertex data and is a popular tool in the field of 3D shape modeling. Using the Delaunay tetrahedralization as adopted in Ho et al. [17] to produce a volumetric mesh that connects the joints in close proximity by edges, the Laplacian coordinates of joint J_i is given by

$$L(J_i) = J_i - \sum_{J_j \in S_i} w_j^i J_j, \tag{8}$$

where S_i is the neighborhood of J_i in the Delaunay mesh. The weight w_j^i of each neighborhood edge is set to be the reciprocal of

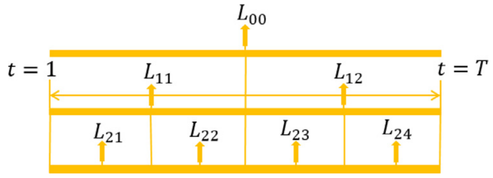


Fig. 4. Temporal construction of the covariance descriptor. L_{li} is the ith covariance matrix in the lth layer temporal pyramid. A covariance matrix at the lth layer covers $\frac{T}{2^l}$ frames of the sequence, where T is the length of the entire human action.

the Euclidean distance between the pair of joints, hence the close proximity joint pairs become more influential while the impact by further apart pairs is suppressed.

Typically, the probability distribution of joint location is not known, inspired by Hussein et al. [18], we give its sample covariance matrix to represent the vertex attribute. Suppose that an action is represented by the JSG with N vertices which denote N joints of the human body, and the action is performed over T frames. Then the sample covariance of $L(J_i)$ $(1 \leq i \leq N)$ is denoted as

$$C(L(J_i)) = \frac{1}{T-1}\sum_{t=1}^{T}\left(L(J_i) - \overline{L(J_i)}\right)\left(L(J_i) - \overline{L(J_i)}\right)', \tag{9}$$

where $\overline{L(J_i)}$ is the sample mean of $L(J_i)$, and the $'$ is the transpose operator. The sample covariance matrix $C(J_i)$ is a symmetric 3×3 matrix. For this descriptor, we only use the upper triangle. Hence, we can see that the upper triangle of the covariance matrix contains $3(3+1)/2 = 6$ elements, which is the length of the vector of the descriptor. This 3D joint descriptor captures the dependence of locations of this joint during the performance of the action, however, it does not capture its temporal properties. Hence, we present the temporal pyramid covariance on 3D joint location as the vertex attribute partly inspired by Fourier Temporal Pyramid proposed by Wang et al. [31].

In the temporal pyramid, the top layer of the 3D joint descriptor is calculated over the entire 3D human action. The lower layers are computed over smaller time windows of the entire action. In Fig. 4, we just show the 3-layer temporal pyramid. In this temporal pyramid, each covariance matrix is identified by two indices: the first is the temporal pyramid layer index, and the second is the index within the layer. The top layer matrix covers the entire action and is denoted by L_{00}. The covariance matrix at layer l is calculated over $T/2^l$ frames of the action. For the descriptor configurations in Fig. 4, the length of joint covariance descriptor with 3-layer temporal pyramid is $7 \times 6 = 42$.

Using this temporal joint descriptor as the vertex attribute, the vertex kernel is defined as

$$k_v\left(v, v'\right) = \exp\left(-\frac{\left\|d_1 - {d_1}'\right\|_2^2}{2{\sigma_1}^2}\right), \tag{10}$$

where d_1 and ${d_1}'$ are the corresponding 3D joint descriptors for vertices v and v' which represent the same joint of human body. $\sigma_1 > 0$ is a scale parameter for the Gaussian function. At the same time, we also define the edge kernel as

$$k_e\left(e, e'\right) = \exp\left(-\frac{\left\|d_2 - {d_2}'\right\|_2^2}{2{\sigma_2}^2}\right), \tag{11}$$

where d_2 and ${d_2}'$ are the values of the RVJRDs which represent the attributes of the edge e and e'. $\sigma_2 > 0$ is a scale parameter for Gaussian function.

Let A and A' be two human action sequences, these actions are represented by their JSGs G and G', respectively. m is set for the maximal order of JSGK. Hence, according to Eq. (7), the final JSGK matching on the two action sequences is denoted as

$$k\left(A, A'\right) = \sum_{q=0}^{m} \lambda_q k_G^q(G, G'), \tag{12}$$

where $\Omega = [\lambda_0, \lambda_1, \ldots .\lambda_m]$, $\Omega > 0$, is the weight coefficient vector.

6. Multiple kernel learning

It is believed that combining multiple kernels in a linear way seems more reasonable than nonlinear combination in fusing information provided by complex Gaussian kernels as pointed out in Varma and Babu [28]. As different order JSGKs may contain redundant information for action recognition, we apply the multiple kernel learning formulation proposed by Bach et al.[4] to learn the sparse linear combination of the different order JSGKs. Specifically, if we have M kernels k_1, ..., k_M, then the multiple kernel learning method learns the optimal linear combination $\sum_{j=1}^{M} \lambda_j K_j$ and the optimal SVM classifier for action recognition simultaneously, where K_j is the kernel matrix according to k_j. The publicly available code of [5] is used in this paper.

7. Runtime complexity analysis

Given two JSGs, G and G' with n_G and $n_{G'}$ vertices, using the direct product graph as adopted in Golub and Van Loan [15] for graph kernel $k_G(G, G')$, the time complexity of JSGK calculation is $O(|V|^3)$, where $|V|$ is the number of corresponding vertices in two JSGs. Hence, $|V| \leq \min\{n_G, n_{G'}\}$.

8. Experiments

8.1. Experimental setting

We evaluate the discrimination power of our graph-based descriptor for 3D human action recognition. We perform this evaluation on three publicly available datasets, MSR-Action3D dataset created by Li et al. [19], UTKinect-Action dataset created by Xia et al. [32] and Florence3D-Action dataset created by Seidenari et al. [22]. JSGK is used to measure the similarity between the JSGs of actions and an SVM classifier is used for action classification. For the multi-class classification task, the usual SVM classifier with JSGK was used in a one-versus-all setting as adopted in Shawe-Taylor and Cristianini [23].

MSR-Action3D dataset is an action dataset of depth sequences captured by a depth camera. The 3D locations of 20 joints are provided with the dataset. This dataset contains twenty types of actions: horizontal arm wave, hammer, hand catch, high arm wave, two hand wave, forward punch, high throw, draw X, draw circle, draw tick, hand clap, side-boxing, bend, side kick, forward kick, jogging, tennis swing, tennis serve, pick up throw, golf swing. Every action was performed by 10 subjects 3 times each. UTKinect-Action dataset consists of depth sequences captured using a single stationary Kinect. The 3D locations of 20 joints are provided with the dataset. This dataset contains action types: walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands, and clap hands. There are 10 subjects, each subject performs each action twice. Florence3D-Action dataset has been captured using a Kinect camera. The 3D locations of 15 joints are provided with the dataset. It includes 9 activities: wave, drink from a bottle, answer phone, clap, tight lace, sit down, stand up, read watch, and bow. During acquisition, 10 subjects were asked to perform the above actions for 2 or 3 times.

For MSR-Action3D dataset, we evaluated the proposed approach based on the top-42 RVJRDs with 6-layer temporal pyramid. For

Table 1
The accuracy results of the two approaches.

MSR-Action3D dataset	
JSG (top-K VJRDs)+JSGK	55.4%
JSG (top-K RVJRDs)+JSGK	92.2%
UTKinect-Action dataset	
JSG (top-K VJRDs)+JSGK	63.6%
JSG (top-K RVJRDs)+JSGK	98.3%
Florence3D-Action dataset	
JSG (top-K VJRDs)+JSGK	60.8%
JSG (top-K RVJRDs)+JSGK	93.2%

Table 2
Comparison with the state-of-the-art results.

MSR-Action3D dataset	
Sequence of most information joints [20]	47.1%
Histogram of 3D joints [32]	78.9%
Eigenjoints [33]	82.3%
Actionlets [31]	88.2%
Lie group [29]	89.5%
Pose-based [30]	90.2%
Skeletal shape trajectories [2]	90%
Histogram of oriented displacements [16]	91.3%
Proposed approach	92.2%
UTKinect-Action dataset	
Random forests [37]	87.9%
Histogram of 3D joints [32]	90.9%
HSOM [10]	94.5%
Lie group [29]	97.1%
Proposed approach	98.3%
Florence3D-Action dataset	
Multi-part bag-of-poses [22]	82.0%
Motion Trajectories [9]	87.0%
Lie group [29]	90.9%
Proposed approach	93.2%

Table 3
The runtime results of our approach on the three datasets.

Dataset	Average testing runtime (ms)
MSR-Action3D dataset	337.16
UTKinect-Action dataset	318.57
Florence3D-Action dataset	206.92

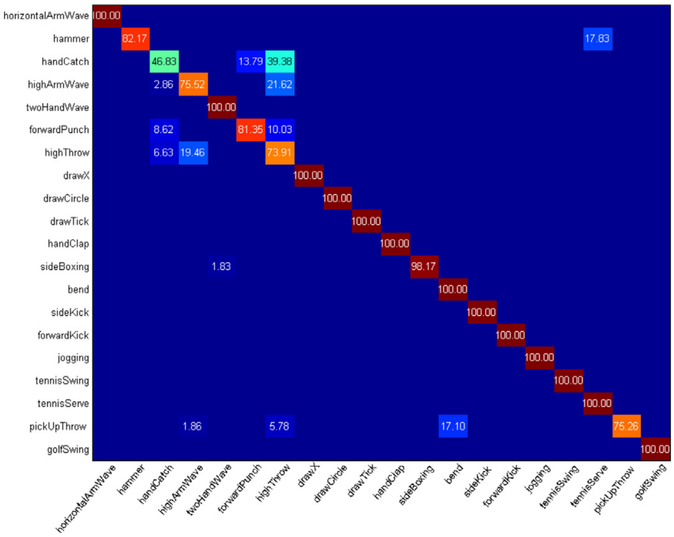


Fig. 5. Confusion matrix for MSR-Action3D dataset.

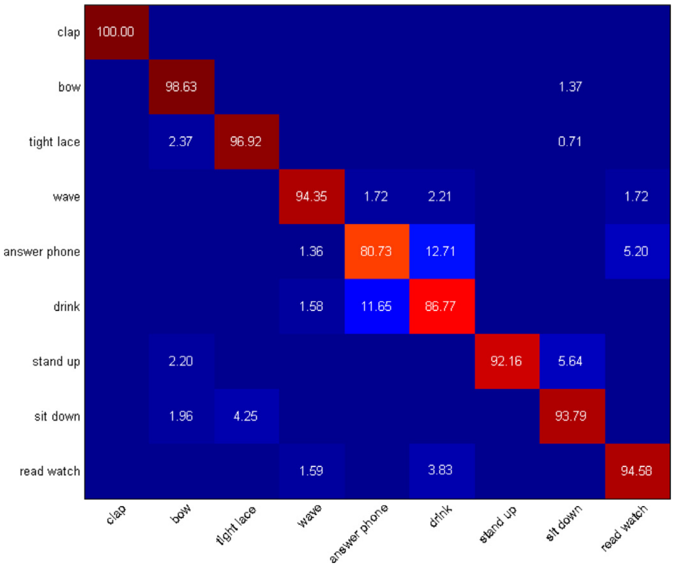


Fig. 6. Confusion matrix for Florence3D-Action dataset.

UTKinect-Action dataset, we evaluated the proposed approach based on the top-43 RVJRDs with 6-layer temporal pyramid. For Florence3D-Action dataset, we evaluated the proposed approach based on the top-26 RVJRDs with 5-layer temporal pyramid. We followed the cross-subject test setting of [19] for MSR-Action3D dataset, and [37] for UTKinect-Action and Florence3D-Action datasets, respectively, in which half of the subjects were used for training and the other half were used for testing.

8.2. Results

First, in order to illustrate the improvement of our proposed RVJRD, two approaches are designed in order to evaluate and contrast with the performance of our algorithm in recognizing human actions. In the first approach, we use JSG based on the top-K VJRDs to model each human action, and apply the JSGK to evaluate the similarity between two JSGs. In the second approach, we use our proposed approach for action recognition.

The above two approaches are referred to as “JSG (top-K VJRDs)+JSGK”, “JSG (top-K RVJRDs)+JSGK”. The parameters K and numbers of temporal pyramid layers in the two approaches are for the best accuracy performance on the three datasets.

Table 1 shows that our proposed approach “JSG (top-K RVJRDs)+JSGK” achieves the much better accuracies on the two action datasets. The performance of our proposed approach illustrates that the proposed RVJRD works much better than the VJRD in action recognition, since RVJRD can reflect the activity levels of the joint pairs for each type of actions better than VJRD.

Then, we compare the proposed approach with several state-of-the-art 3D human skeletal action recognition methods. The comparison of the recognition accuracy is shown in Table 2. We can see that the proposed approach achieves the best results on all datasets. Some recent approaches like [21] and [37] have reported recognition around 94.5% for MSR-Action3D dataset by combining skeletal features with additional depth-based features. Since the focus of this paper is not on combining multiple features, we only use skeletal results reported in Ohn-Bar and Trivedi [21] and Zhu et al.[37] for comparison.

Table 3 shows the average testing runtime of our approach for each input action on the three datasets, where the testing running time includes all steps in the testing mode. All the experiments were conducted on a desktop computer with Intel Core 2 Duo 3.17 GHz processor. The efficiency in runtime is attributed to the limited number of the vertices in the proposed JSG.

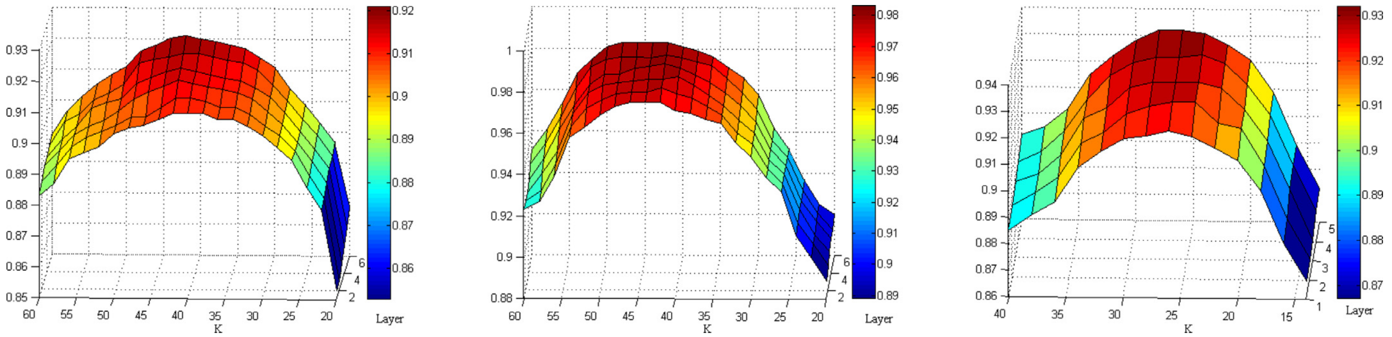


Fig. 7. Recognition accuracy performance of our approach with respective to the parameters K and layer on three datasets: Left-MSR-Action3D dataset; Center-UTKinect-Action dataset; Right-Florence3D-Action dataset.

Figs. 5 and 6 show the confusion matrices for MSR-Action3D dataset and Florence3D-Action dataset, respectively. We can see that our approach works very well. The classification confusions occur when the two actions are highly similar to each other like “hand catch” and “high throw” in the case of MSR-Action3D dataset, “drink” and “answer phone” in the case of Florence3D-Action dataset. We skip the UTKinect-Action dataset as the corresponding action recognition accuracy is very high.

Fig. 7 shows the performance of our proposed approach under different values of K and numbers of layers on the three datasets. It is clear that the recognition accuracies remain high over a large range of the value of K and the number of layer, which demonstrates that our recognition approach is robust to these parameters. However, when the value of K is too small, the JSGs do not have enough edges to reflect the spatial structures of joints in actions. When the value of K is too large, some edges with low values of RVJRDs may be selected in JSGs. Thus, the JSGs may not be discriminative enough to describe the actions. Therefore, we should choose proper value of K for our approach. For the same value of K, our approach can achieve higher accuracy with lager number of layer, since lager number of layer means that our proposed joint covariance descriptor can capture more temporal information of locations of joints in actions. However, the maximum number of temporal pyramid layers in our experiment is limited according to the minimum length of actions. In general, we can deduce that adding more layers can improve the recognition accuracy.

9. Conclusion and future work

In this paper, we have proposed a novel graph-based approach for 3D human skeletal action recognition. First, we have constructed a novel graph model based on the top-K RVJRDs which were proposed to model spatial structures among joints in different skeletal actions. Second, we have adopted the temporal pyramid of covariance descriptors to preserve certain layers of spatio-temporal joint features. Third, we have computed the graph kernel by matching the walks from each of the two graphs to be matched. Our experimental results have shown that our graph-based approach has superior performance in comparison with several state-of-the-art methods on three different action datasets.

Future work includes the application of our graph-based approach on 2D video data from an arbitrary viewpoint, preprocessed to provide partial 2D joint positions. In order to apply our approach to action recognition in 2D videos, we can use the similar idea in Yao and Fei-Fei [34] to estimate the 2D key points of human skeleton from the image.

Acknowledgment

The work described in this paper was fully supported by a grant from City University of Hong Kong (Project No. 7004548).

References

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

[2] B.B. Amor, J. Su, A. Srivastava, Action recognition using rate-invariant analysis of skeletal shape trajectories, IEEE Trans. Pattern Anal. Mach. Intell. 38 (1) (2016) 1–13.

[3] N.B. Aoun, M. Mejdoub, C.B. Amar, Graph-based approach for human action recognition using spatio-temporal features, J. Vis. Commun. Image Represent. 25 (2) (2014) 329–338.

[4] F.R. Bach, G.R. Lanckriet, M.I. Jordan, Multiple kernel learning, conic duality, and the SMO algorithm, in: Proceedings of the Twenty-First International Conference on Machine Learning, ACM, 2004, p. 6.

[5] F.R. Bach, R. Thibaux, M.I. Jordan, Computing regularization paths for learning multiple kernels, Adv. Neural Inf. Process. Syst. 17 (2005) 73–80.

[6] K.M. Borgwardt, C.S. Ong, S. Schönauer, S. Vishwanathan, A.J. Smola, H.-P. Kriegel, Protein function prediction via graph kernels, Bioinformatics 21 (Suppl 1) (2005) i47–i56.

[7] E.Z. Borzeshi, M. Piccardi, R.Y.D. Xu, A discriminative prototype selection approach for graph embedding in human action recognition, in: Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, 2011, pp. 1295–1301.

[8] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, R. Vidal, Bio-inspired dynamic 3d discriminative skeletal features for human action recognition, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2013, pp. 471–478.

[9] 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) (2015) 1340–1352.

[10] W. Ding, K. Liu, F. Cheng, J. Zhang, Learning hierarchical spatio-temporal pattern for human activity prediction, J. Vis. Commun. Image Represent. 35 (2016) 103–111.

[11] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1110–1118.

[12] T. Gärtner, P. Flach, S. Wrobel, On graph kernels: hardness results and efficient alternatives, in: Learning Theory and Kernel Machines, Springer, 2003, pp. 129–143.

[13] U. Gaur, Y. Zhu, B. Song, A. Roy-Chowdhury, A string of feature graphs model for recognition of complex activities in natural videos, in: Proceedings of 2011 IEEE International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 2595–2602.

[14] B. Gauzere, L. Brun, D. Villemin, M. Brun, Graph kernels based on relevant patterns and cycle information for chemoinformatics, in: Proceedings of 21st International Conference on Pattern Recognition (ICPR), IEEE, 2012, pp. 1775–1778.

[15] G.H. Golub, C.F. Van Loan, Matrix Computations, vol. 3, JHU Press, 2012.

[16] M.A. Gowayyed, M. Torki, M.E. Hussein, M. El-Saban, Histogram of oriented displacements (hod): Describing trajectories of human joints for action recognition., in: IJCAI, 2013.

[17] E.S. Ho, T. Komura, C.-L. Tai, Spatial relationship preserving character motion adaptation, ACM Trans. Graph. vol. 29 (2010) 33.

[18] M.E. Hussein, M. Torki, M.A. Gowayyed, M. El-Saban, Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations, in: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, AAAI Press, 2013, pp. 2466–2472.

[19] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: Proceedings of the 2010 IEEE Computer Society Conference onComputer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2010, pp. 9–14.

[20] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Sequence of the most informative joints (SMIJ): a new representation for human skeletal action recognition, J. Vis. Commun. Image Represent. 25 (1) (2014) 24–38.

[21] E. Ohn-Bar, M.M. Trivedi, Joint angles similarities and hog2 for action recognition, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2013, pp. 465–470.

[22] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, P. Pala, Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2013, pp. 479–485.

[23] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.

[24] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, Comput. Vis. Pattern Recognit. 10 (June 2011) 5–32.

[25] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, H.-P. Seidel, Laplacian surface editing, in: Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing, ACM, 2004, pp. 175–184.

[26] A.-P. Ta, C. Wolf, G. Lavoue, A. Baskurt, Recognizing and localizing individual activities through graph matching, in: Proceedings of the 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2010, pp. 196–203.

[27] J.K. Tang, H. Leung, Retrieval of logically relevant 3d human motions by adaptive feature selection with graded relevance feedback, Pattern Recognit. Lett. 33 (4) (2012) 420–430.

[28] M. Varma, B.R. Babu, More generality in efficient multiple kernel learning, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 1065–1072.

[29] R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3d skeletons as points in a lie group, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 588–595.

[30] C. Wang, Y. Wang, A. Yuille, An approach to pose-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 915–922.

[31] J. Wang, Z. Liu, Y. Wu, J. Yuan, Learning actionlet ensemble for 3d human action recognition, IEEE Trans. on Pattern Anal. Mach.Intell., 36 (5) (2014) 914–927.

[32] L. Xia, C.-C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2012, pp. 20–27.

[33] X. Yang, Y. Tian, Eigenjoints-based action recognition using naive-bayes-nearest-neighbor, in: Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2012, pp. 14–19.

[34] B. Yao, L. Fei-Fei, Action recognition with exemplar based 2.5 d graph matching, in: Proceedings of the 12th European Conference on Computer Vision–ECCV 2012, Springer, 2012, pp. 173–186.

[35] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, J. Gall, A survey on human motion analysis from depth data, in: Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications, Springer, 2013, pp. 149–187.

[36] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks, in: Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2015.

[37] Y. Zhu, W. Chen, G. Guo, Fusing spatiotemporal features and joints for 3d action recognition, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2013, pp. 486–491.