Multimed Tools Appl (2019) 78:6329–6353
https://doi.org/10.1007/s11042-018-6370-1

Effective human action recognition using global and local offsets of skeleton joints

Bin Sun[1] · Dehui Kong[1] · Shaofan Wang[1] · Lichun Wang[1] · Yuping Wang[1] · Baocai Yin[2]

Received: 1 December 2017 / Revised: 14 May 2018 / Accepted: 3 July 2018 /
Published online: 20 July 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract Human action recognition based on 3D skeleton joints is an important yet challenging task. While many research work are devoted to 3D action recognition, they mainly suffer from two problems: complex model representation and low implementation efficiency. To tackle these problems, we propose an effective and efficient framework for 3D action recognition using a global-and-local histogram representation model. Our method consists of a global-and-local featuring phase, a saturation based histogram representation phase, and a classification phase. The global-and-local featuring phase captures the global feature and local feature of each action sequence using the joint displacement between the current frame and the first frame, and the joint displacement between pairwise fixed-skip frames, respectively. The saturation based histogram representation phase captures the histogram representation of each joint considering the motion independence of joints and saturation of each histogram bin. The classification phase measures the distance of each joint histogram-to-class. Besides, we produce a novel action dataset called BJUT Kinect dataset, which consists of multi-period motion clips and intra-class variations. We compare our method with many state-of-the-art methods on BJUT Kinect dataset, UCF Kinect dataset, Florence 3D action dataset, MSR-Action3D dataset, and NTU RGB+D Dataset. The results show that our method achieves both higher accuracy and efficiency for 3D action recognition.

✉ Dehui Kong
kdh@bjut.edu.cn

Bin Sun
sunbin1357@emails.bjut.edu.cn

[1] Beijing Advanced Innovation Center for Future Internet Technology, Beijing Key Laboratory of Multimedia and Intelligent Software Technology, BJUT Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China

[2] College of Computer Science and Technology, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, 116024, China

Keywords Action recognition · Skeleton joints · Offsets · Histogram representation · Naive-Bayes-Nearest-Neighbor

1 Introduction

Human action recognition has been a significant issue for the past three decades due to its practical applications in many critical fields, e.g., human-computer interaction, video surveillance, motion retrieval, and health-care [6, 9, 21, 27, 34, 36]. Much progress related to the research on intensity image based action recognition has been made [11, 15, 18, 30, 39]. However, intensity image based action recognition methods vastly suffer from lots of difficult situation, such as the illumination and viewpoint variation, the cluttered background, the camera movement, and the partial occlusion. Moreover, these methods also encounter difficulties to extract discriminative features because of the intra-class variability and inter-class similarity of the action sequences, which are critical for achieving a high recognition accuracy.

Range information, provided by the Microsoft Kinect liked somatosensory devices, has been proved to be useful for solving these problems, which can improve the recognition accuracy of the actions that are hard to be recognized by intensity images due to their similarity in 2D projections space. Significant improvements [7, 20, 32, 41] have shown promising applications of depth maps in the field of human action recognition. Furthermore, the sequences of 3D skeleton joints of an action video clip also can be obtained in real-time using Microsoft Kinect SDK toolkit [57]. Since human skeleton can be viewed as an articulated system connected by hinged joints, the human actions are essentially embodied in skeletal motions in the 3D space. As a result, many skeleton based methods are springing up [2, 5, 10, 26, 49, 54].

Despite fruitful research work on 3D skeleton based action recognition methods, existing methods still endure some drawbacks, especially when representing the structure of actions. Some methods [1, 12, 17, 51] treat the feature of each joint by stacking them together, which produce a great amount of computing cost and high dimension feature, others [8, 14, 19, 31, 48] are devoted to subtle feature pruning and accompanied by a complicated classification models, which are usually time consuming and supervised. These two kinds of methods improve recognition accuracy while reducing computational efficiency, which is more important in practical usage. Considering the presented problems, in this paper, we propose a novel joint offsets based histogram representation model for each joint, which is simple to implement, sufficiently efficient in recognition tasks, and unsupervised at training process. At the same time, the characteristics of the displacements in the local movement and the global movement are comprehensively considered, thus improving the recognition accuracy.

The flowchart of the proposed framework is illustrated in Fig. 1, which includes the training phase and the testing phase. The main idea of this paper comes from the observation that the offset determined by the displacement of each skeleton joint corresponding to two different frames reflects the movement characteristic of the joint during the time interval of these two frames. Moreover, the joint offset from the first frame to the current frame, named after global offset feature, manifest the global movement of the joint, and the joint offset from a pair of frames with a fixed interval, named after local offset feature, manifest the local movement of the joint. The strategy to hybrid local feature and global feature together can promote the action recognition accuracy without increasing the complication of the model. Then the joint histogram representation is generated by clustering and coding the global

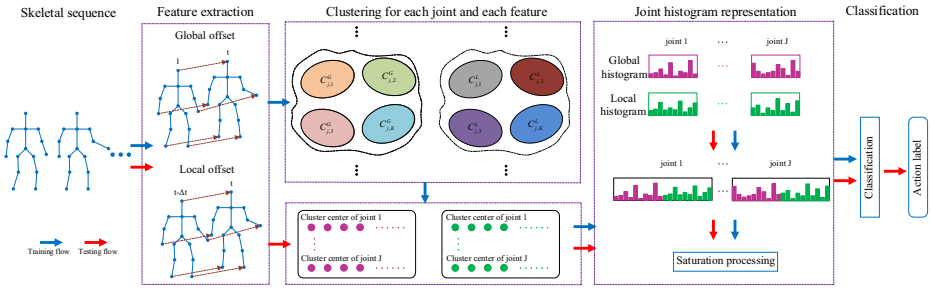


Fig. 1 The general framework of the proposed method, where purple and green solid points are cluster centers of global and local offsets, respectively. For clarity, we only illustrate global offsets and local offsets of five joints

and local offsets of each joint, respectively. Furthermore, because the effect of each bin of histogram cannot be neglected, saturation based histogram representation is proposed. This strategy based on histogram representation model is motivated by [26]. While Lu et al. [26] proposed a histogram model by clustering local offset vectors of all joints together, our model differs from their work in three aspects: (1) We propose the global offset feature that captures the global characteristic of an action, and integrate it with the local offset feature together to construct the motion representation model; (2) We improve the method performance by clustering the offset vectors for each joint independently instead of clustering for all joints together; (3) We use saturation for histogram representation model to enhance the discrimination ability of the features. Afterwards, we employ two different classifiers to classify the different actions, respectively, i.e., Naive-Bayes-Nearest-Neighbor(NBNN) and Sparse Representation-based Classifier(SRC). The experiments are run over five datasets with different characteristic, including BJUT dataset captured by ourselves with Kinect, UCF Kinect dataset, Florence 3D action dataset, MSR-Action3D dataset, and NTU RGB+D Dataset. The results show that our method achieves both higher accuracy and efficiency.

In summary, the main contributions of our work include four aspects as follows:

(1) A novel action feature that consists of global and local position offsets of joints is used, which synergistically reflects the spatial and temporal properties of the action video.
(2) The action representation model is generated using a set of joint histograms. The motion independence of skeleton joints is embedded with the representation by applying K-means clustering to all offset vectors of each joint separately. Moreover, saturation of each histogram bin is considered to enhance the discrimination ability.
(3) The effect of two different classifiers on our proposed method, i.e., NBNN and SRC, are verified separately, which maintain the effectiveness of the spatial independence of joints by a distance measurement principle of joint histogram-to-class. The former is a non-parametric classifier, which does not require learning process, and is easy to implement for practical usage. The latter is a parametric classifier, which requires learning process.
(4) A novel action dataset is provided which consists of ten classes, in which each video sequence poses an action multi-period.

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 elaborates on the global-and-local featuring phase and the saturation based

histogram representation phase. Section 4 describes two classifiers, including Naive-Bayes-Nearest-Neighbor(NBNN) and Sparse representation-based classifier(SRC). Experimental results are presented in Section 5. The conclusion is given in Section 6.

2 Related work

Various range information based human action recognition approaches have been proposed in the past decades. According to the types of the raw data the approaches rely on, these methods fall into three classes, which are depth map based, skeleton joints based, and multiple data modalities based methods.

The first class of methods only used depth maps or 3D point clouds converted from depth maps for action recognition. Li et al. [20] constructed an action graph to represent the actions, and used a bag of 3D points to characterize the postures, which is obtained by projection based sampling method. Yang et al. [53] applied HOG to depth motion maps which were generated by accumulating motion energy of depth maps projected onto three orthogonal Cartesian planes. Wang et al. [47] proposed random occupancy pattern features for action recognition, and weighted random sampling to explore an extremely large dense sampling space. Similarly, Vieira et al. [45] proposed a space-time occupancy patterns, and divided space and time axes into multiple segments to define a 4D grid. Oreifej et al. [33] encoded the distribution of surface normal orientation in the 4D space. In the above methods, they model the action by the discrete points in the depth map and fail to consider every map as an overall set with inherent structure of human body, which limits the recognition accuracy.

As for skeleton joints based methods, the features were extracted to capture the essential structure of actions. Ellis et al. [12] presented a logistic regression learning framework that automatically determined distinctive pose representation of each action. Xia et al. [50] used histograms of 3D skeleton joint locations (HOJ3D) as a compact representation of postures. Posture vocabularies were built by clustering HOJ3D vectors calculated from a large collection of postures, and discrete hidden Markov model was used for action classification. Zhou et al. [56] presented a skeleton induced discriminative approximate rigid part model for human action recognition, which not only captured the human geometrical structure, but also took rich human body surface cues into consideration. Yang et al. [51] proposed another action feature descriptor called eigenjoints by calculating the differences of joints, which includes relative position information, consecutive information, and offset information. Similarly, Jiang et al. [17] presented a method with consecutive information and relative position information, and employed weighted graphs to organize these information. Li et al. [19] also used a graph-based model to characterize the actions, but only used relative position feature. The proposed top-K relative variance of joint relative distances determined which joint pairs should be selected in the resulting graph. In addition, the temporal pyramid covariance descriptors were adopted to represent joint locations. Qiao et al. [35] proposed trajectorylet, which captured static and dynamic information in a short interval of joints, and generated a representation for actions by learning a set of distinctive trajectorylet detectors. In order to reduce the feature space, Luvizon et al. [29] proposed extracting sets of spatial and temporal local features from subgroups of joints, used the Vector of Locally Aggregated Descriptors (VLAD) represent an action, and then proposed a metric learning method which can efficiently combine the feature vectors. Lu et al. [26] extracted feature by computing local position offsets of joints. However, this method didn't fully exploit temporal relationship of action sequences, which fails to capture the continuous

information of each action; moreover, it didn't consider the motion independence of each joint in the codebook formation phase which mattered in action recognition. Alternatively, we construct a global offset feature as well as K-means clustering of offsets of each joint for compensating the deficiency of the method of Lu et al. [26]. Besides, some works have presented satisfactory results using skeletal features in RNN [42] and Long Short-Term Memory (LSTM) networks [55]. Due to the relatively small number of training samples, neural networks methods usually leads to strong overfitting.

As for multiple data modalities based methods, more than one type of data source was used for action recognition. Ohn-bar et al. [32] proposed two descriptors including joint angle similarity and modified HOG algorithm. Similarly, Zhu et al. [58] fused the spatio-temporal interest points extracted from depth sequence and skeleton joint feature with random forests. Luo et al. [28] proposed a framework fusing pairwise relative position feature extracted from skeleton joints and center-symmetric motion local ternary pattern feature extracted from RGB sequences. Besides RGB and skeleton joints, Sung et al. [41] added depth maps for action recognition. A two-layer maximum entropy Markov model was presented for classification. Wang et al. [46] combined the pairwise relative position feature and local occupancy pattern, and employed Fourier temporal pyramid to represent the actions. In the above methods, they have complex models and require long computing time.

As the extension of human action, human activity can be considered as the composition of some actions. There has been relatively little work on bridging the gap between actions and activities, Liu et al. [22, 23] provided temporal pattern mining, which encoded temporal relatedness among actions, and captured the intrinsic properties of activities. Furthermore, Liu et al. [24] presented a probabilistic interval-based model where the Chinese restaurant process model is incorporated to capture the inherent structural varieties of complex activities. Due to the difficult collection of annotated or labelled training data for sensor-based supervised human activity recognition, Lu et al. [27] proposed an unsupervised method for recognizing physical activities using smartphone sensors. Since action recognition is the basis of activity recognition, in this paper, we focus on discussing action recognition.

From this related work, we can conclude three important facts. First, most methods concentrating on high efficiency are skeleton-based method. Second, both spatial and temporal information are important for action recognition. It is not trivial to fuse global and local temporal information. Third, the trajectory feature of each joint is independent and important for action recognition, and the importance of each joint feature is not equal. In our work, we only use skeletal joints as input data, and our method characterizes both the global and local movements. The combination of the joints can improve the recognition accuracy, but this requires class label information. However, the training process in this paper is unsupervised, so the labels-related combination of joints cannot be realized. Therefore, we propose the following compromise: We separately represent the trajectory of each joint using saturation based histogram representation, allowing further classification by measuring the distance of joint feature to class.

3 Feature extraction and action representation

In this section, the proposed representation model based on global and local offsets of skeleton joints is described. The main idea is first to extract the low-level feature of an action by computing the position offset of corresponding joint in two assigned frames, and then to construct the histogram representation of the action by clustering and coding the global and local offsets of each joint, respectively.

3.1 Joint-based spatial-temporal feature extraction

Let Ψ denote a set including N video sequences:

$$\Psi \equiv \{F_r | F_r = [f_r(1), f_r(2), \ldots, f_r(n_r)], r = 1, 2, \ldots, N\}, \tag{1}$$

where F_r represents the rth video with n_r frames. Suppose that J joints are acquired in each frame, the tth frame $f_r(t)$ can be denoted by the 3D coordinates of joints as follows:

$$f_r(t) = \{\theta_{1,r}(t), \theta_{2,r}(t), \ldots, \theta_{J,r}(t)\}, t = 1, 2, \ldots, n_r, \tag{2}$$

where $\theta_{j,r}(t) = (x_{j,r}(t), y_{j,r}(t), z_{j,r}(t))$ denotes the 3D position of the jth joint in $f_r(t)$, $j = 1, 2, \ldots, J$.

Obviously, the joint coordinates reveal the spatial feature of the action, while the joint displacements characterize the temporal feature of the action. Therefore, we calculate the corresponding joint offset between the tth frame and the $(t - \Delta t)$th frame to present the spatial-temporal feature of the action.

$$\phi^L_{j,r}(t) = \theta_{j,r}(t) - \theta_{j,r}(t - \Delta t), \tag{3}$$

where Δt is the time difference which can balance the precision of the offset and the ability of robustness to noise. If Δt becomes greater, noise fluctuations are more robust, but computation precision becomes lower, and vice versa. Upper label L is used to indicate that the feature describes the local movement of the joint during the time interval$[(t - \Delta t), t]$. However, (3) only characterizes the local spatial-temporal property, and fails to express the global movement of the joint related to the original pose in the first frame. Therefore, in order to enhance the spatial-temporal property, we introduce the global offset, which is computed as the displacement from the joint position in the first frame to the position of the corresponding joint in the tth frame.

$$\phi^G_{j,r}(t) = \theta_{j,r}(t) - \theta_{j,r}(1). \tag{4}$$

Thus, $f_r(t)$ can be represented as follows:

$$\begin{aligned} \Phi^G_r(t) &= [\phi^G_{1,r}(t), \phi^G_{2,r}(t), \ldots, \phi^G_{J,r}(t)], \\ \Phi^L_r(t) &= [\phi^L_{1,r}(t), \phi^L_{2,r}(t), \ldots, \phi^L_{J,r}(t)]. \end{aligned} \tag{5}$$

In other words, the combination of two features forms the preliminary feature representation of each frame as follows:

$$\Phi_r(t) = [\Phi^G_r(t), \Phi^L_r(t)], t = \Delta t + 1, \Delta t + 2, \ldots, n_r. \tag{6}$$

Therefore, an action can be described as follow:

$$F'_r = \begin{bmatrix} \phi^G_{1,r}(\Delta t + 1) & \phi^G_{2,r}(\Delta t + 1) & \cdots & \phi^G_{J,r}(\Delta t + 1) \\ \phi^G_{1,r}(\Delta t + 2) & \phi^G_{2,r}(\Delta t + 2) & \cdots & \phi^G_{J,r}(\Delta t + 2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi^G_{1,r}(n_r) & \phi^G_{2,r}(n_r) & \cdots & \phi^G_{J,r}(n_r) \\ \phi^L_{1,r}(\Delta t + 1) & \phi^L_{2,r}(\Delta t + 1) & \cdots & \phi^L_{J,r}(\Delta t + 1) \\ \phi^L_{1,r}(\Delta t + 2) & \phi^L_{2,r}(\Delta t + 2) & \cdots & \phi^L_{J,r}(\Delta t + 2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi^L_{1,r}(n_r) & \phi^L_{2,r}(n_r) & \cdots & \phi^L_{J,r}(n_r) \end{bmatrix}. \tag{7}$$

For the method of Lu et al. [26], only the local offset is extracted to represent an action, while our proposed method improves its representation method by introducing the global

offset as above. Figure 2 shows that the difference between the global offset of the tth frame and the local offset of the tth frame equals the global offset of the $(t - \Delta t)$th frame, i.e.,

$$\phi^G_{j,r}(t - \Delta t) = \phi^G_{j,r}(t) - \phi^L_{j,r}(t), \tag{8}$$

which means the partial feature of the current frame is embraced by the feature of the subsequent frame.

3.2 Joint-based histogram representation model

After each training action has been represented by a set of low-level features from all body joints according to (7), histogram of occupation frequency(HOF) representation method based on offset vectors clustering is used to generate the action representation model. Inspired by Luo et al. [28] which indicated that each joint plays a different role for different actions, we maintain the motion independence of joints using K-means clustering for the offset vectors of each joint respectively as illustrated in Fig. 3 rather than for all joints together.

Firstly, we group together the global offset vectors of each joint of training action sequences, and denote it by $\Omega^G_j = \{\phi^G_{j,r}(t)\}_{j=1,2,\ldots,J,r=1,2,\ldots,N,t=1,2,\ldots,n_r}$, where N is the number of video sequences, and n_r is the frame number of the rth sequence. Here Ω^G_j corresponds to the global feature set of the jth body joint of all frames in all training action sequences. In the same way, $\Omega^L_j = \{\phi^L_{j,r}(t)\}_{j=1,2,\ldots,J,r=1,2,\ldots,N,t=1,2,\ldots,n_r}$ denotes the local feature set of each body joint. Then we use K-means clustering algorithm for Ω^G_j and Ω^L_j to form cluster centers $C^G_{j,k}$ and $C^L_{j,k}, k = 1, 2, \ldots, K$, where K is the number of clusters, which is very important to balance the discrimination and robustness of representation model. The Euclidean distance is used as the clustering metric.

Then, a given video sequence F'_r with n_r frames described as (7) can be further expressed by a set of histograms, which represent occupation frequencies of assigned clusters.

$$\begin{aligned}
\alpha^G_{j,r}(k') &= \frac{\#\{\phi^G_{j,r}(t) | k' = \arg\min_k \|\phi^G_{j,r}(t) - C^G_{j,k}\|\}}{n_r - \Delta t}, k' = 1, 2, \ldots, K, \\
\alpha^L_{j,r}(k') &= \frac{\#\{\phi^L_{j,r}(t) | k' = \arg\min_k \|\phi^L_{j,r}(t) - C^L_{j,k}\|\}}{n_r - \Delta t}, k' = 1, 2, \ldots, K,
\end{aligned} \tag{9}$$

where $\alpha^G_{j,r}(k')$ and $\alpha^L_{j,r}(k')$ represent the histograms of global and local offsets of the jth joint, respectively, and $\#\{\}$ denotes the cardinality of a set, $k = 1, 2, \ldots, K$, $t = \Delta t + 1, \Delta t + 2, \ldots, n_r$. Thus, the movement of the jth joint can be represented by a histogram, i.e., $\alpha_{j,r} = [\alpha^G_{j,r}, \alpha^L_{j,r}] \in \mathbb{R}^{2K}$.

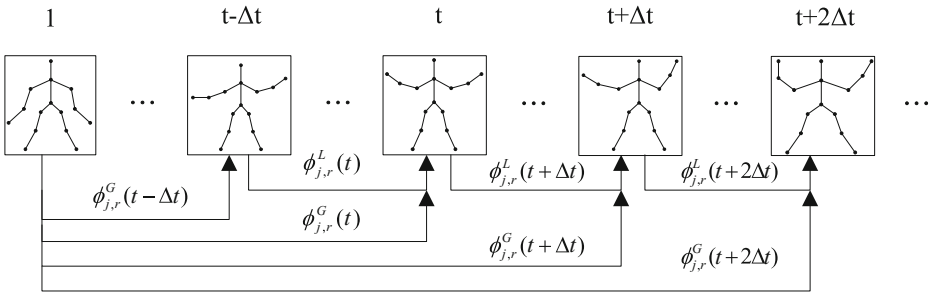


Fig. 2 Illustration of temporal sequence property

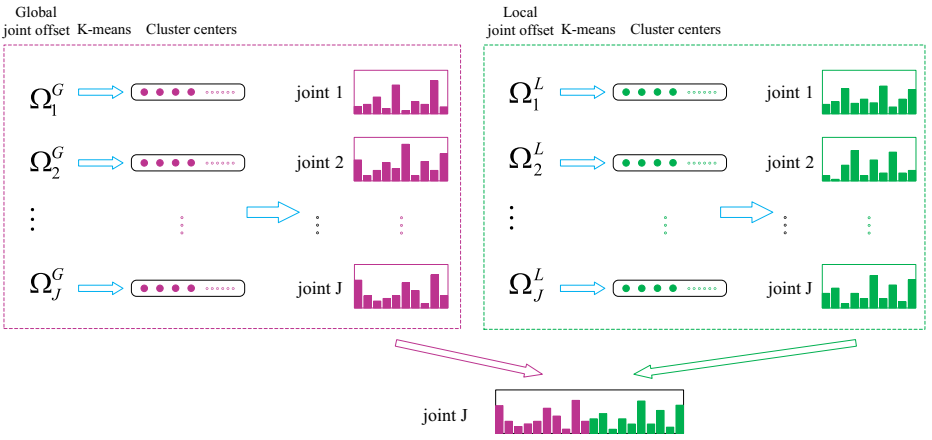


Fig. 3 Illustration of histogram representation using K-means clustering for Ωj^G and Ωj^L

Furthermore, we notice the existence of nonuniform distribution of occupation frequencies among the histograms. For instance, the majority of clusters in a histogram have a few occupation frequencies. In the situation, if we don't consider saturation, and directly use the histogram feature, the effect of clusters that have low frequencies would be diminished in the classification. Their frequencies are very low relative to the other clusters with high frequencies, nevertheless these clusters with low frequencies are usually very relevant for action recognition. Therefore, saturation based histogram of occupation frequency(SHOF) is proposed. we set a parameter ε to truncate the high occupation frequencies, then the histogram can be represented as follows:

$$\begin{aligned}
\alpha'^{G}_{j,r}(k) &= \frac{\min\{\alpha^{G}_{j,r}(k), \varepsilon\}}{\sum_{k'=1}^{K} \min\{\alpha^{G}_{j,r}(k'), \varepsilon\}}, k = 1, 2, \ldots, K, \\
\alpha'^{L}_{j,r}(k) &= \frac{\min\{\alpha^{L}_{j,r}(k), \varepsilon\}}{\sum_{k'=1}^{K} \min\{\alpha^{L}_{j,r}(k'), \varepsilon\}}, k = 1, 2, \ldots, K,
\end{aligned} \tag{10}$$

where ε is empirically selected to maximize recognition accuracy. Thus, the movement of the jth joint can be represented by a histogram, i.e., $\alpha'_{j,r} = [\alpha'^{G}_{j,r}, \alpha'^{L}_{j,r}] \in \mathbb{R}^{2K}$. Finally, an action sequence can be represented by SHOF based on the joint movement, i.e., $F''_r = [\alpha'_{1,r}, \alpha'_{2,r}, \ldots, \alpha'_{J,r}] \in \mathbb{R}^{2K \times J}$. Our proposed framework is presented in Algorithm 1.

Algorithm 1 Process of proposed histogram representation method

Require: Training set $\{F_r\}_{r=1}^{N}$; parameters Δt, K, ε

1: **while** $i \leq joints\ number$ **do**
2: Extract global and local feature via (3 – 4);
3: Cluster for global and local feature sets, respectively;
4: Extract HOF via (9);
5: Extract SHOF via (10);
6: **end while**
7: **return** SHOF

4 Classification

Suppose an action sequence is represented by a set of histograms of all joints, i.e., $V = [h_1, h_2, \ldots, h_J]$. To remain the movement independence of each joint, which often provides additional clues for action discrimination, we classify an action video by measuring the distance of joint histogram-to-class rather than the distance of video-to-video or video-to-class. The action recognition based on the distance of histogram-to-class is performed according to the following equation:

$$c^* = \arg\min_c \sum_{j=1}^{J} \|h_j - U_j^c(h_j)\|^2, \tag{11}$$

where c^* reflects the class that the testing video sequence V belongs to, h_j is the histogram of the jth joint of V, $U_j^c(h_j)$ is the nearest histogram with h_j in the class c. We apply two different classifiers, i.e., NBNN and SRC, to classify the actions based on the above distance measurement principle, respectively. The difference is that the former is employed to classify proposed histogram representation, and the latter is employed to classify the sparse representation transformed from histogram representation.

4.1 Naive-bayes-nearest-neighbor classifier

Naive-Bayes-Nearest-Neighbor (NBNN) [4] is employed by measuring the distance of histogram-to-class defined as above. NBNN is a non-parametric classifier and equips with the following four advantages compared with other learning-based classifiers. (1) NBNN doesn't require learning process; (2) NBNN can avoid the over-fitting problem; (3) NBNN can deal with a large number of classes; (4) NBNN is easy to implement for practical usage.

The action video is classified according to (11) with $U_j^c(h_j)$ represented as follows:

$$U_j^c(h_j) = NN_j^c(h_j) = \arg\min_{h'} |h_j - h'_j(c)|, \tag{12}$$

where $h'_j(c)$ denotes the histogram of jth joint of an action in the class c.

4.2 Sparse representation-based classifier

Assume that there are C classes in the training set. For the jth joint in the cth class, gathering all histograms of sample videos, we can learn a dictionary to represent the histogram feature of the jth joint. In this way, we learn $C \times J$ dictionaries. The jth joint histograms in the training videos of the cth class are arranged as columns of matrix $A_j^c = \{h_j^p\}_{p=1,2,\ldots,P}$, where P is the number of training videos in the cth class. We wish to learn a dictionary $D_j^c \in \mathbb{R}^{2K \times P}$ over which A_j^c has a sparse representation $X_j^c = \{x_1, x_2, \ldots, x_P\}$. It is modeled as the following optimization problem:

$$\min_{D,X}\{\|A_j^c - D_j^c X_j^c\|_F^2\} \quad s.t. \|x\|_0 \le q_1. \tag{13}$$

For a testing video sequence, $V = [h_1, h_2, \ldots, h_J]$. One way to classify V is to find approximations of $\{h_j\}_{j=1,2,\ldots,J}$, given by each of the learned dictionaries and their corresponding reconstruction errors. The following (14) defines the item of $U_j^c(h_j)$.

$$U_j^c(h_j) = D_j^c \hat{x}_j^c = \min_{D_j^c \tilde{x}_j^c} \|h_j - D_j^c \tilde{x}_j^c\|_2^2 \quad s.t. \|\tilde{x}_j^c\|_0 \le q_2, \tag{14}$$

where $\hat{x}_j^c$ is the sparse representation of V over D_j^c, $j = 1, 2, \ldots, J$. The $U_j^c(h_j)$ is then substituted into (11) to execute the classification.

5 Experimental results

In this paper, we evaluate our method on five datasets, including a new dataset captured by ourselves called BJUT Kinect dataset and four publicly available datasets: UCF Kinect dataset [12], Florence 3D action dataset [37], MSR-Action3D dataset [20], and NTU RGB+D Dataset [38] . The experiments are run on a Core (TM) i7-4790 3.6GHz machine with 8GB RAM using Matlab R2016a.

5.1 Databases

5.1.1 BJUT Kinect dataset

We introduce a new action dataset by Kinect sensor called BJUT Kinect dataset, which we collected in order to emphasize two points: First, each video sequence of the dataset is multi-period, and each actor performed a requested action different times in each sequence. This dataset is useful to evaluate how well the feature descriptors for multi-period actions. Second, each individual performed actions freely without standard action demo so that this dataset has a certain diversity, which brings difficulty for recognition. This dataset has 159 video sequences in total and 10 classes listed in Table 1. In each frame, the 3D coordinates of 25 joints are available. The dataset is captured from 12 individuals including 9 males and 3 females whose ages range from 24 to 35. The actions of this dataset are illustrated in Fig. 4.

5.1.2 UCF Kinect dataset

UCF Kinect dataset [12] is a publicly available dataset including 16 classes: `balance`, `climb ladder`, `climb up`, `duck`, `hop`, `kick`, `leap`, `punch`, `run`, `step back`, `step forward`, `step left`, `step right`, `twist left`, `twist right`, `vault`. This dataset is captured by Kinect sensor and the OpenNI platform, gathered from 16 individuals (13 males and 3 females whose ages range from 25 to 35), and has 1280 video sequences in total. In each frame, the 3D coordinates of 15 joints are available. Every one performed 16 actions five times. The dataset is used to measure the latency possible, and how quickly a method can overcome the ambiguity in initial poses when performing an action.

5.1.3 Florence 3D action dataset

Florence 3D action dataset [37] is captured by Kinect camera including 215 action sequences. It includes 9 action classes: `wave`, `drink from a bottle`, `answer phone`, `clap`, `tight lace`, `sit down`, `stand up`, `read watch`, `bow`. 10 subjects

Table 1 The list of actions on the BJUT Kinect dataset

1. Open arms	2. Shooting
3. Goalkeeping	4. Kick
5. Wave	6. Twist
7. Step forward-back	8. Warm-up
9. Head snaking	10. Step left-right

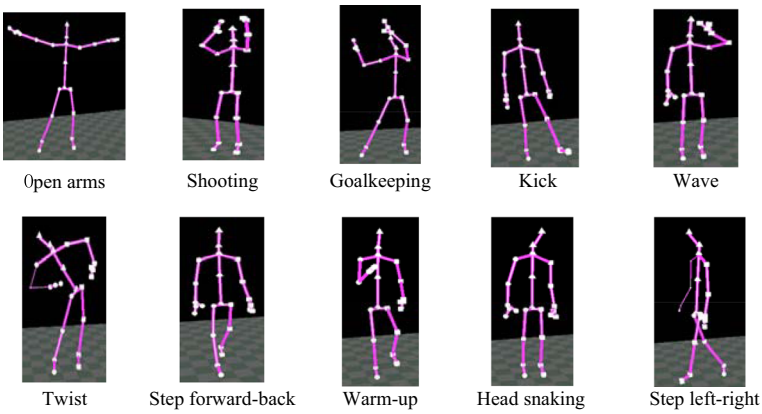


Fig. 4 Several poses associated with different actions on the BJUT Kinect dataset

were asked to perform the above actions twice or three times. The 3D positions of 15 joints are provided in each frame. Moreover, the dataset has high intra-class variations, and most activities involve human-object interactions, which is challenging for recognition only by 3D joints.

5.1.4 MSR-Action3D dataset

MSR-Action3D dataset [20] is a publicly available dataset including 567 action sequences, and is performed by 10 individuals. It includes 20 action classes: `high wave`, `horizontal arm wave`, `hammer`, `hand catch`, `forward punch`, `high throw`, `draw X`, `draw tick`, `draw circle`, `hand clap`, `hand wave`, `side boxing`, `bend`, `forward kick`, `side kick`, `jogging`, `tennis swing`, `tennis serve`, `golf swing`, `pickup throw`. The data was recorded with a depth sensor similar to Kinect. Each individual performed each action 2 or 3 times. The dataset provides 3D skeleton joints and depth maps. We use 3D skeleton joints only, and the 3D coordinates of 20 joints are available in each frame.

5.1.5 NTU RGB+D dataset

NTU RGB+D Dataset [38] is a large human action dataset, which provides more than 56000 sequences and 4 million frames. There are 60 action classes performed by 40 distinct subjects, including 40 daily actions(e.g., `drinking`, `reading`, `writing`), 9 health-related actions (e.g., `sneezing`, `staggering`, `falling down`) and 11 mutual actions (e.g., `handshaking`, `hugging`, `punching`). The dataset used three cameras to capture these actions, which were placed at different locations and viewpoints. In each frame, the 3D coordinates of 25 joints are available. Due to the large number of viewpoints, intra-class and sequence length variations, the dataset is very challenging.

5.2 Parameters evaluation

For action representation, we need tune three parameters. Time difference Δt balances the precision and robustness to noise, the number of clusters K balances the discrimination of

Table 2 Parameters setting of each dataset, where Δt is time difference, K is clusters number, and ε is saturation parameter

Dataset	Δt	K	ε
BJUT Kinect dataset	2:2:10	10:10:40	0.1:0.1:1
UCF Kinect dataset	2:2:10	10:10:40	0.1:0.1:1
Florence 3D action dataset	2:2:8	10:10:40	0.1:0.1:1
MSR-Action3D dataset	2:2:10	10:10:40	0.1:0.1:1
NTU RGB+D dataset	2:2:10	80:10:150	0.1:0.1:1

our method, and saturation parameter ε balances the effect of each bin of the histogram. We empirically determine the intervals and the step size of each parameter. Table 2 shows the parameters setting of each dataset. Due to different protocols of each dataset, we describe the tuning process for three of the evaluated datasets. We tune Δt and K jointly, and evaluate all combinations of the two parameter values to find the optimal values according to the accuracy. Figure 5 shows the performance with different values of Δt and K for three dataset, where BJUT1 represents original BJUT dataset, and BJUT2 represents BJUT dataset with the video segmented to contain only one motion in a clip. There is an optimum spot in each dataset for Δt and K, which gives the best performance. We tune ε on two datasets as shown in Fig. 6. It is clear that, when ε is 0.5 and 0.6, we achieve the best performance on UCF Kinect dataset and Florence 3D action dataset, respectively. For sparse representation-based classifier (SRC), we need tune q_2 to investigate the impact of the sparsity of joint histogram feature. The experiments are performed on the UCF Kinect dataset

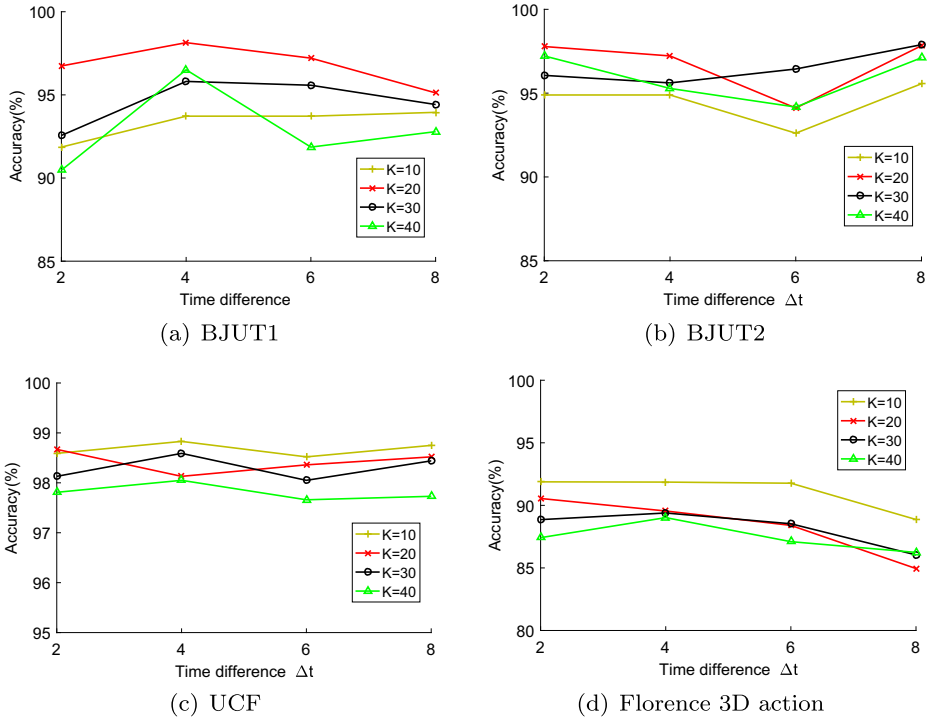


Fig. 5 The recognition accuracies with different combinations of parameter values for each dataset

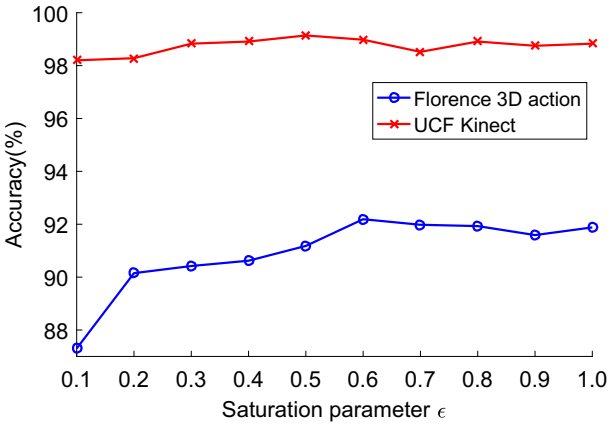


Fig. 6 The recognition accuracies with different values of ε

with different values of q_2 as shown in Fig. 7. From the figure, we can see that $q_2 = 4$ achieves the best performance.

5.3 Experimental results on BJUT Kinect dataset

For this dataset, an action is performed more times in a video sequence. Therefore, the evaluation is performed with two protocols. We first test on the dataset without video segmentation(BJUT1), and then test on the dataset with the video segmented to contain only one motion in a clip(BJUT2). The ratio of training video and testing video is both 3:1, and the repeat count is 10 and 20, respectively. Table 3 shows the comparable results. Here we use "LF", "GF", "AT", and "ER" to denote "local feature", "global feature", "clustering offsets of all joints together", and "clustering offsets of each joint respectively", for convenience. For SRC-based method, $q_1 = 2$, $q_2 = 4$. The results verify that the modeling strategy of clustering offset vectors of each joint respectively is effective and outperforms the strategy of clustering offset vectors of all joints together. We also can reach the conclusion that our model improves recognition accuracy without excessive increase

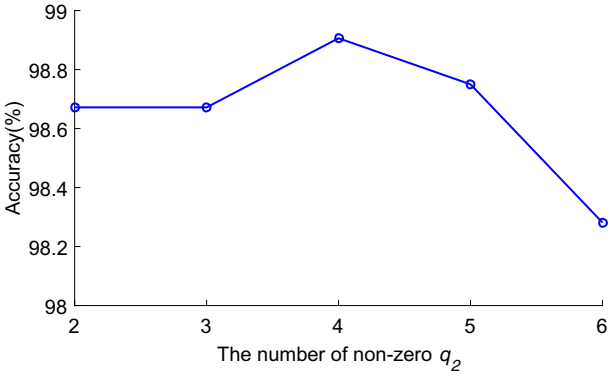


Fig. 7 The recognition accuracies with different values of q_2

Table 3 Performance comparison on the BJUT Kinect dataset. we use "LF", "GF", "AT", and "ER" to denote "local feature", "global feature", "clustering offsets of all joints together", and "clustering offsets of each joint respectively", respectively. (8,40) represents $\Delta t = 8$ and $K = 40$

Method		BJUT1	BJUT2
LF+AT+NBNN [26]	Parameter	(8,40)	(8,30)
	Accuracy	92.09%	93.78%
	Dimensionality	1000	750
LF+GF+AT+NBNN	Parameter	(8,30)	(6,10)
	Accuracy	96.05%	95.44%
	Dimensionality	1500	500
LF+GF+ER+NBNN	Parameter	(4,20)	(8,30)
	Accuracy	98.14%	97.89%
	Dimensionality	1000	1500
LF+GF+ER+SRC	Parameter	(4,20)	(8,30)
	Accuracy	99.07%	96.94%
	Dimensionality	1000	1500

dimensionality. The effectiveness comes from employing the global offset feature to intensify the temporal property of the model. Furthermore, the accuracy of BJUT1 is better than BJUT2 employing the global feature, which further illustrate that global feature intensifies the temporal property. The recognition accuracy of the SRC-based method is slightly better than NBNN-based method. Our method based on SRC outperforms 0.93% and 0.08% than our method based on NBNN classifier on the BJUT1 and the UCF, respectively. However, our method based on NBNN outperforms 0.95% than SRC on the BJUT2.

5.4 Experimental results on UCF Kinect dataset

For this dataset, We use 5-fold cross validation estimation method to evaluate our method. We obtain the best performance when Δt is 4, K is 10, and ε is 0.5. Table 4 shows the accuracy comparison with the state-of-the-art methods. In the case of UCF Kinect dataset, the average accuracy of the proposed method is 99.14%. Our method is comparable to the method of Jiang et al. [17], and better than other methods. Figure 8 shows the confusion

Table 4 Performance comparison on the UCF kinect dataset

Method	Accuracy(%)
LTSS [25]	91.70
BoW [12]	94.06
CRF [12]	94.29
LAL [12]	95.94
EigenJoint [51]	97.10
JAS [32]	97.37
Local offset [26]	97.58
Grassmann Manifold [40]	97.91
MvMF+HMM [3]	98.90
Ours(SHOF+SRC)	98.91
Ours(SHOF+NBNN)	99.14
Weighted graphs [17]	99.30

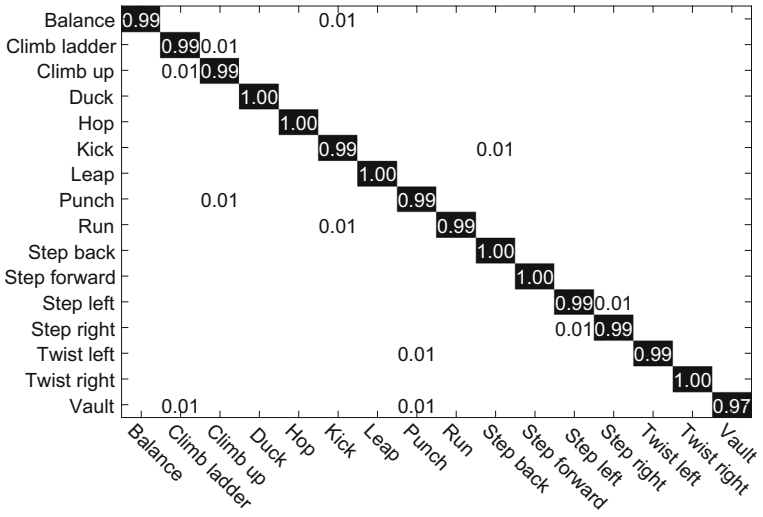


Fig. 8 Confusion matrix of proposed method(SHOF+NBNN) for the UCF kinect dataset

matrix of action recognition corresponding to the best accuracy of our method. The result shows that the recognition accuracies of most actions are 100%, such as `climb ladder`, `duck`, `hop`, `step back`, `step forward` and `twist right`. As for `balance`, `kick`, `leap`, `step left`, `step right` and `twist left`, their recognition accuracies are no less than 99%, even the poor recognition accuracies for `climb up`, `punch` and `vault` are as high as 95%. Table 5 shows the average testing runtime of each phase of our method. The average frame number of each video is 66 for the UCF. Our method based on NBNN only costs approximate 0.018s for one video sequence, and our method based on SRC costs approximate 0.088s for one video sequence. Both two methods are highly efficient, and NBNN is faster than SRC.

5.5 Experimental results on Florence 3D action dataset

For this dataset, we follow the standard leave-one-out-cross-validation protocol as described in [37] to evaluate our method. We obtain the best performance when Δt is 2, K is 10, and ε is 0.6. Table 6 shows the accuracy comparison with the state-of-the-art methods. We can see that our proposed method, comparable to the method of Yang et al. [54] which is supervised-based method, achieves better performance than other methods. We can obtain a recognition accuracy of 92.19%, which is very good performance. Figure 9 illustrates the confusion matrix on the Florence 3D action dataset. We can see that the proposed method performs very well on most of the actions, except some actions, such as `drink from a bottle` and `answer phone`, which are often misclassified each other. The reason is that for these human-object interactions, object information is not available from the skeleton joints data making these interactions look almost the same.

Table 5 The average testing runtime for each phase of our method per action sequence on the UCF dataset

Phase	Video representation	NBNN	SRC
Time(ms)	8.147	9.781	79.36

Table 6 Performance comparison on the florence 3D action dataset

Method	Accuracy(%)
Multi-part bag-of-poses [37]	82.15
Riemannian manifold [10]	87.04
RF-PCA [2]	89.67
Lie Group [43]	90.88
Wang et al. [49]	91.63
Ours(SHOF+NBNN)	92.19
Skelets [54]	93.42

5.6 Experimental results on MSR-Action3D dataset

For this whole dataset, we follow the cross-subject evaluation as described in [33], where the samples of half of the subjects are utilized for training, and the others are employed as testing data. We obtain the best performance when Δt is 6, K is 30, and ε is 0.3. Table 7 shows the comparison results with the state-of-the-art unsupervised methods. Our proposed method achieves acceptable performance. For this dataset, the accuracy of supervised methods [28, 48] can reach 93.8%, which are outperform ours, but this result should be viewed in the context of the accuracy/latency trade-off. These methods require that the entire action be viewed before recognition can occur. Insight into the performance of our method can be gained by examining the accuracies for specific action classes. Figure 10 shows the comparisons of recognition accuracies of our method and the method of Luvizon et al. [29], which is a state-of-the-art unsupervised method. From the figure, we can see that the two unsupervised methods can all exactly distinguish the actions with different body poses, but get into trouble when distinguishing the actions with similar poses, such as `draw x` and `draw tick`. The immediate reason is that our representation model is based on the primary skeleton joints related with torso and limbs like "big" part, and ignore the detailed joints related with fingers like "little" part. As a result, the actions distinguished by subtle detail are difficult to recognize. Therefore, using more detailed joints to represent action is our future research plan.

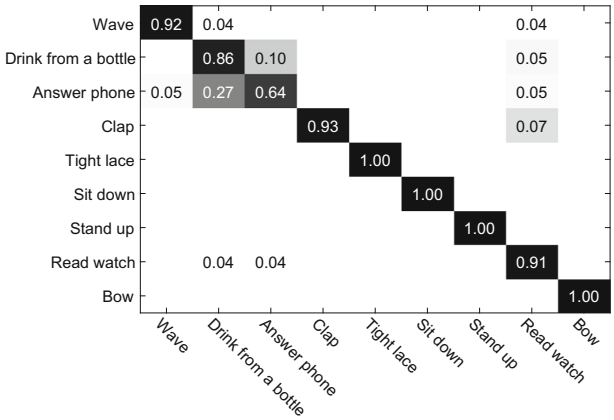


Fig. 9 Confusion matrix of our method on the Florence 3D action dataset

Table 7 Performance comparison on the MSR-Action3D dataset

Method	Accuracy(%)
LAL [12]	65.7
Bag of 3D points [20]	74.7
HOJ3D [50]	79
STOP [44]	81.43
EigenJoint [51]	82.3
Local feature+VLAD [29]	83.2
Ours(SHOF+NBNN)	83.6
HON4D [33]	85.8

5.7 Experimental results on NTU RGB+D dataset

For this dataset, the evaluation is performed with two standard protocols as described in [38], i.e., cross-subject evaluation and cross-view evaluation. For cross-subject evaluation, the samples of 20 subjects are used for training and the samples from 20 other subjects are used for testing. For cross-view evaluation, the samples captured by two cameras are used for training and the others are used for testing. We obtain the best performance when Δt is 2, K is 100, and ε is 1. The comparison results with the state-of-the-art handcrafted methods on this dataset are reported in Table 8. We can find out that our proposed method achieves acceptable performance when features are calculated only using skeleton joints without using multi-modal fusion such as [16]. The method [16] also employed supervised learning for their features, whose performance improvement coincided with an increase in computational cost particularly in the training phase.

5.8 Efficiency analysis

Because fair execution under the same condition are almost impossible, we cannot compare the actual computation time of other methods. Therefore, the efficiency analysis is discussed from two aspects, i.e., the computational complexity analysis and the latency analysis.

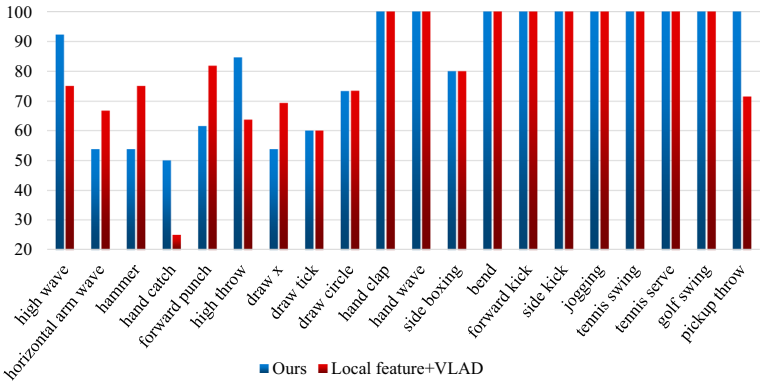


Fig. 10 Recognition accuracy (per action) for the MSR-Action3D dataset obtained by local feature+VLAD and our method

Table 8 Performance comparison on the NTU RGB+D dataset

Method	Accuracy(%)	
	Cross subject	Cross view
HON4D [33]	30.56	7.26
Super normal vector [52]	31.82	13.61
JAS [32]	32.24	22.27
Skeletal Quads [13]	38.62	41.36
Lie Group [43]	50.08	52.76
Ours(SHOF+NBNN)	51.43	56.16
FTP Dynamic skeletons [16]	60.23	65.22

For the computational complexity analysis, we compare our method with the methods of [12] and [26], which are also concentrate on computational efficiency while keeping a satisfactory recognition accuracy. Table 9 shows the comparison of computational complexity. From the table, we can see that our method has a comparable computational complexity with the compared methods. Our feature has lower dimension and our method has higher recognition accuracy. Therefore, our method is more effective with high computational efficiency while keeping a better recognition accuracy than the compared methods. In a word, our proposed method has low computational complexity, which can be implemented in real-time.

For the latency analysis, the goal here is to investigate how many frames are sufficient to enable accurate action recognition. We evaluate our method on sequences of varying frame lengths. From the original dataset, new datasets are created by varying a parameter termed maxFrames. The sub-sequences are extracted from the first maxFrames frames of the video. If the video is shorter than maxFrames, the entire video is used. The comparison of recognition accuracies using different number of first maxFrames frames are illustrated in Fig. 11, where LAL, CRF and BOW are the methods of Ellis et al. [12], and Local offset is the method of Lu et al. [26]. From Fig. 11 we can see that our method clearly outperforms other methods. All of the methods perform poorly given a small number of frames, and well given a large number of frames. However, in the middle range, i.e., from 20 frames to 40 frames, our approach achieves a much higher accuracy than all other methods. In a word, our method can recognize actions at the desired accuracy with a lower latency.

5.9 Discussion

Unlike many methods [14, 29] using supervised methods, our method is an unsupervised method. When using unsupervised techniques to extract features, there is no need to rely

Table 9 Computational complexity of different methods, where feature dimension and accuracy are on the UCF Kinect dataset, and n is frame number

Method	Computational complexity	Feature dimension	Accuracy(%)
LAL [12]	O(n)	2776	95.94
Local offset [26]	O(n)	300	97.58
Ours	O(n)	300	99.14

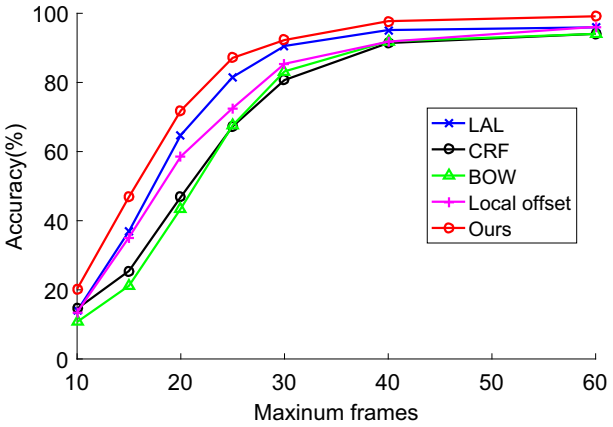


Fig. 11 Accuracy vs. state-of-the-art methods over videos truncated at varying maximum frames

on prior knowledge, and no data inadaptability problem, since the features are learned from the data. Based on our experiments, time difference Δt has the effect on the precision and robustness to noise, the number of clusters K has the effect on the discrimination, and saturation parameter ε has the effect on balancing the effect of each bin of the histogram. We find that the three parameters of action representation have different values for different datasets. Tuning the three parameters is an important task, which has significant effect on the recognition accuracy of our method(see Figs. 5, 6 and 7). We also come to a conclusion that each component of our proposed method improves the recognition accuracy. Compared with Lu et al. [26], which is very close to ours, our method not only characterize both the global and local movements of the joints in an action sequence, but also improve temporal sequence property (see Table 3). Furthermore, our method, comparable to some methods which employ supervised techniques and complex learning model, achieves better performance than many other methods on five different types of datasets. For BJUT Kinect dataset, each video sequence is multi-period. For UCF Kinect dataset, it is a relatively large and clean dataset, and general measure the efficiency of methods. For Florence 3D action dataset, it has many human-object interactions and high intra-class variations. For MSR-Action3D dataset, it has a great amount of noise and high intra-class variations. For NTU RGB+D Dataset, it is perhaps the largest human action dataset, and has a large number of viewpoints and intra-class variations. The evaluations in terms of efficiency have clearly revealed our method can recognize actions in real-time. It is possible to recognize actions up to 92% using only 30 frames which is a good performance comparing to state-of-the-art methods (see Fig. 11). Thus, our approach can be used for interactive systems.

However, our method has some limitations. Our proposed method is a 3D joint-based framework for human action recognition from skeletal joints sequences. Therefore, for actions involving human-object interactions, our method does not provide any relevant information about objects and thus, action with different objects are confused. In future, this limitation can be improved by leveraging complementary information, which can be extracted from depth or color images associated with 3D joint locations. Besides, if a dataset has both a great amount of noise and high intra-class variations, such as MSR-Action3D dataset, our method cannot accurate recognition. Further study is needed to determine

precisely how important low latency is in these types of abstract actions. More detailed joint information is also an extension of future research.

6 Conclusion

This paper presents a novel framework for action recognition focusing on the computational efficiency. In the framework, an action feature is designed based on offsets of skeleton joints including global offset feature and local offset feature that can intensify the temporal sequence property. A novel histogram representation model based on global and local offsets of joints is introduced to represent actions considering the spatial independence of joints. K-means clustering algorithm is used for the global or local offset vectors of each joint, respectively. This method can get higher accuracy than the method of K-means based on offset vectors of all joints together. Histogram of occupation frequency based high-level representation model is constructed to represent a video sequence. The saturation scheme is presented to modify the model, in case that the majority of the clusters with low occupation frequencies would be overlooked. Two classifiers based on measuring the histogram-to-class distance are designed, including NBNN and SRC. Two classifiers achieve approximate recognition accuracies, and NBNN is much faster than SRC. A novel dataset for the purpose of our experiments called BJUT dataset by Kinect and four publicly available datasets including UCF Kinect dataset, Florence 3D action dataset, MSR-Action3D dataset, and NTU RGB+D Dataset are introduced to testify our framework. The experiments on five datasets show that our method is effective, and achieves a comparable or a better performance compared with the state-of-the-art methods.

In conclusion, the motion feature proposed in this paper is concise and intuitive, and the action representation model is facile and discriminative. However, the actions with similar body poses or human-object interactions cannot be recognized precisely using our method. To improve our framework, exploring more discriminative features with low dimensionality is the coming work.

Acknowledgments This work was supported by National Natural Science Foundation of China (No. 61772048, 61632006), Beijing Natural Science Foundation (No. 4162009), Beijing Municipal Science and Technology Project(No. Z161100001116072, Z171100004417023).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

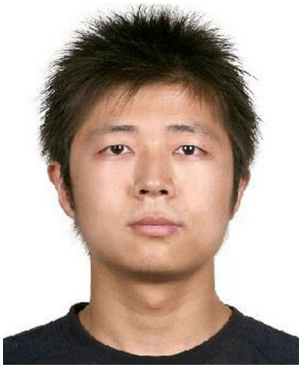
1. Agahian S, Negin F, Köse C (2018) Improving bag-of-poses with semi-temporal pose descriptors for skeleton-based action recognition. Vis Comput. https://doi.org/10.1007/s00371-018-1489-7
2. Anirudh R, Turaga P, Su J, Srivastava A (2017) Elastic functional coding of Riemannian trajectories. IEEE Trans Pattern Anal Mach Intell 39(5):922–936
3. Beh J, Han D, Durasiwami R, Ko H (2014) Hidden markov model on a unit hypersphere space for gesture trajectory recognition. Pattern Recogn Lett 36:144–153
4. Boiman O, Shechtman E, Irani M (2008) In defense of nearest-neighbor based image classification. In: IEEE conference on computer vision and pattern recognition, pp 1–8
5. Chaaraoui A, Padilla-Lopez J, Climent-Perez P, Florez-Revuelta F (2014) Evolutionary joint selection to improve human action recognition with RGB-d devices. Expert Syst Appl 41:786–794

6. Chen H, Hwang J (2011) Integrated video object tracking with applications in trajectory-based event detection. J Vis Commun Image Represent 22:673–685
7. Chen W, Guo G (2015) Triviews: a general framework to use 3D depth data effectively for action recognition. J Vis Commun Image Represent 26:182–191
8. Chen H, Wang G, Xue J-H, He L (2016) A novel hierarchical framework for human action recognition. Pattern Recogn 55:148–159
9. Cui J, Liu Y, Xu Y, Zhao H, Zha H (2013) Tracking generic human motion via fusion of low-and high-dimensional approaches. IEEE Trans Syst Man Cybern Syst 43:996–1002
10. Devanne M, Wannous H, Berretti S, Pala P, Daoudi M, Del Bimbo A (2015) 3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold. IEEE Trans Cybern 45(7):1340–1352
11. Dong J, Sun C, Yang W (2015) A supervised dictionary learning and discriminative weighting model for action recognition. Neurocomputing 158:246–256
12. Ellis C, Masood S, Tappen M, Laviola J, Sukthankar R (2013) Exploring the trade-off between accuracy and observational latency in action recognition. Int J Comput Vis 101:420–436
13. Evangelidis G, Singh G, Horaud R (2014) Skeletal quads: human action recognition using joint quadruples. In: International conference on pattern recognition, pp 4513–4518
14. Eweiwi A, Cheema MS, Bauckhage C, Gall J (2014) Efficient pose-based action recognition. In: Asian conference on computer vision, pp 428–443
15. Fathi A, Mori G (2008) Action recognition by learning mid-level motion features. In: IEEE conference on computer vision and pattern recognition
16. Hu J-F, Zheng W-S, Lai J, Zhang J (2015) Jointly learning heterogeneous features for RGB-D activity recognition. In: IEEE conference on computer vision and pattern recognition, pp 5344–5352
17. Jiang X, Zhong F, Peng Q, Qin X (2016) Action recognition based on global optimal similarity measuring. Multimedia Tools and Applications 75:11019–11036
18. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: IEEE conference on computer vision and pattern recognition, pp 1–8
19. Li M, Leung H (2017) Graph-based approach for 3D human skeletal action recognition. Pattern Recogn Lett 87:195–202
20. Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3D points. In: IEEE computer society conference on computer vision and pattern recognition workshops, pp 9–14
21. Liu Y, Cui J, Zhao H, Zha H (2012) Fusion of low-and high-dimensional approaches by trackers sampling for generic human motion tracking. In: International conference on pattern recognition, pp 898–901
22. Liu Y, Nie L, Han L, Zhang L, Rosenblum DS (2015) Action2activity: recognizing complex activities from sensor data. In: International joint conferences on artificial intelligence, pp 1617–1623
23. Liu Y, Nie L, Liu L, Rosenblum DS (2016) From action to activity: sensor-based activity recognition. Neurocomputing 181:108–115
24. Liu L, Cheng L, Liu Y, Jia Y, Rosenblum DS (2016) Recognizing complex activities by a probabilistic interval-based model. In: AAAI conference on artificial intelligence, pp 1266–1272
25. Lu G, Zhou Y (2013) Extraction of action patterns using local temporal self-similarities of skeletal body-joints. In: International congress on image and signal processing, pp 96–100
26. Lu G, Zhou Y, Li X, Kudo M (2016) Efficient action recognition via local position offset of 3D skeletal body joints. Multimed Tools Appl 75:3479–3494
27. Lu Y, Wei Y, Liu L, Zhong J, Sun L, Liu Y (2017) Towards unsupervised physical activity recognition using smartphone accelerometers. Multimed Tools Appl 76(8):10701–10719
28. Luo J, Wang W, Qi H (2014) Spatio-temporal feature extraction and representation for RGB-d human action recognition. Pattern Recogn Lett 50:139–148
29. Luvizon DC, Tabia H, Picard D (2017) Learning features combination for human action recognition from skeleton sequences. Pattern Recogn Lett 99:13–20
30. Matikainen P, Hebert M, Sukthankar R (2009) Trajectons: Action recognition through the motion analysis of tracked features. In: IEEE 12th international conference on computer vision workshops, pp 514–521
31. Negin F, Özdemir F, Akgül CB, Yüksel KA, Erçil A (2013) A decision forest based feature selection framework for action recognition from rgb-depth cameras. In: International conference image analysis and recognition, pp 648–657
32. Ohn-bar E, Trivedi M (2013) Joint angles similiarities and HOG2 for action recognition. In: IEEE international conference of computer vision and pattern recognition workshops, pp 465–470
33. Oreifej O, Liu Z (2013) Hon4d: histogram of oriented 4D normals for activity recognition from depth sequences. In: IEEE conference on computer vision and pattern recognition, pp 716–723

34. Pirsiavash H, Ramanan D (2012) Detecting activities of daily living in first-person camera views. In: IEEE conference on computer vision and pattern recognition, pp 2847–2854
35. Qiao R, Liu L, Shen C, van den Hengel A (2017) Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition. Pattern Recogn 66:202–212
36. Roccetti M, Marfia G, Semeraro A (2012) Playing into the wild: a gesture-based interface for gaming in public spaces. J Vis Commun Image Represent 23:426–440
37. Seidenari L, Varano V, Berretti S, Del Bimbo A, Pala P (2013) Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In: IEEE conference on computer vision and pattern recognition workshops, pp 479–485
38. Shahroudy A, Liu J, Ng T-T, Wang G (2016) NTU RGB+ D: a large scale dataset for 3D human activity analysis. In: IEEE conference on computer vision and pattern recognition, pp 1010–1019
39. Sheng B, Yang W, Sun C (2015) Action recognition using direction-dependent feature pairs and non-negative low rank sparse model. Neurocomputing 158:73–80
40. Slama R, Wannous H, Daoudi M, Srivastava A (2015) Accurate 3D action recognition using learning on the grassmann manifold. Pattern Recogn 48:556–567
41. Sung J, Ponce C, Selman B, Saxena A (2012) Unstructured human activity detection from RGBD images. In: IEEE international conference on robotics and automation, pp 842–849
42. Veeriah V, Zhuang N, Qi G-J (2015) Differential recurrent neural networks for action recognition. In: IEEE international conference on computer vision, IEEE, pp 4041–4049
43. Vemulapalli R, Arrate F, Chellappa R (2014) Human action recognition by representing 3D skeletons as points in a lie group. In: IEEE conference on computer vision and pattern recognition, pp 588–595
44. Vieira AW, Nascimento ER, Oliveira GL, Liu Z, Campos MF (2012) STOP: space-time occupancy patterns for 3D action recognition from depth map sequences. In: Iberoamerican congress on pattern recognition, pp 252–259
45. Vieira A, Nascimento E, Oliveira G, Liu Z, Campos M (2014) On the improvement of human action recognition from depth map sequences using space-time occupancy patterns. Pattern Recogn Lett 36:221–227
46. Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. In: IEEE conference on computer vision and pattern recognition, pp 1290–1297
47. Wang J, Liu Z, Chorowski J, Chen Z, Wu Y (2012) Robust 3D action recognition with random occupancy patterns. In: European conference on computer vision, pp 872–885
48. Wang J, Liu Z, Wu Y (2014) Learning actionlet ensemble for 3D human action recognition. IEEE Trans Pattern Anal Mach Intell 36(5):914–927
49. Wang P, Yuan C, Hu W, Li B, Zhang Y (2016) Graph based skeleton motion representation and similarity measurement for action recognition. In: European conference on computer vision, pp 370–385
50. Xia L, Chen C, Aggarwal J (2012) View invariant human action recognition using histograms of 3D joints. In: IEEE computer society conference on computer vision and pattern recognition workshops, pp 20–27
51. Yang X, Tian Y (2014) Effective 3D action recognition using eigenjoints. J Vis Commun Image Represent 25:2–11
52. Yang X, Tian Y (2014) Super normal vector for activity recognition using depth sequences. In: IEEE conference on computer vision and pattern recognition, pp 804–811
53. Yang X, Zhang C, Tian Y (2012) Recognizing actions using depth motion maps-based histograms of oriented gradients. In: ACM international conference on multimedia, pp 1057–1060
54. Yang Y, Deng C, Tao D, Zhang S, Liu W, Gao X (2017) Latent max-margin multitask learning with skelets for 3-D action recognition. IEEE Trans Cybern 47(2):439–448
55. Zhang S, Liu X, Xiao J (2017) On geometric features for skeleton-based action recognition using multilayer LSTM networks. In: IEEE winter conference on applications of computer vision, pp 148–157
56. Zhou Y, Ming A (2016) Human action recognition with skeleton induced discriminative approximate rigid part model. Pattern Recogn Lett 83:261–267
57. Zhu Y, Dariush B, Fujimura K (2010) Kinematic self retargeting: a framework for human pose estimation. Comput Vis Image Underst 114:1362–1375
58. Zhu Y, Chen W, Guo G (2013) Fusing spatiotemporal features and joints for 3D action recognition. In: IEEE conference on computer vision and pattern recognition workshops, pp 486–491

Bin Sun is currently a Ph.D student in Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing University of Technology. His research interests is pattern recognition.

Dehui Kong received the M.S. degree and Ph.D. degree both from Beihang University in 1992 and 1996, respectively. She is a professor of Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing University of Technology. Her research interests include virtual reality, computer graphics and pattern recognition.



Shaofan Wang received the B.S. degree and Ph.D. degree in computational mathematics both from Dalian University of Technology in 2003 and 2010, respectively. He is an associate professor of Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing University of Technology. His research interests include computer graphics and pattern recognition.

Lichun Wang received the M.S degree from Harbin Institute of Technology in 1998 and Ph.D degree from Nanjing University in 2001. Currently she is a professor of Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing University of Technology. Her research interest is human computer interaction.

Yuping Wang received the B.S degree from Northeast Normal University in 2008, M.S and Ph.D degrees from Beijing University of Technology in 2011 and 2016. Currently she is a post-doctor of Beijing University of Technology. Her research area is computational photography and image processing.

Baocai Yin received the B.S. degree, M.S. degree and Ph.D. degree from Dalian University of Technology in 1985, 1988 and 1993, respectively. He is a professor of the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology. He is currently the editorial member for Journal of Information and Computational Science. His research interests include digital multimedia, multi-functional perception, virtual reality and computer graphics.