This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2803838, IEEE Transactions on Big Data

Human Activity Recognition with Posture Tendency Descriptors on Action Snippets

Yaqiang Yao, Yan Liu, Zhenyu Liu, Huanhuan Chen, *Senior Member, IEEE*

Abstract— Human activity recognition is a challenging problem in computer vision due to large resemblance across classes and variance within an individual class. A routine way to recognize human activity from 3D skeleton sequences can be divided into two tasks, discriminative features representation and temporal dynamics modeling. During the past few years, temporal pyramid is widely used for capturing temporal dynamics after extracting discriminative features from frames. However, this uninformative dividing method could destroy the geometric structure of meaningful action snippets within skeleton sequence. To resolve this problem efficiently, we propose a novel and intuitive method in this paper. First, based on a more realistic assumption that adjacent postures in action sequences are more similar and activity can be depicted with several action snippets, a dividing algorithm is designed to encode the temporal information. Second, an interpretable and discriminative descriptor named posture tendency descriptor (PTD) is constructed to represent one action snippet. Finally, multiple PTDs along the entire skeleton sequence are concatenated in a hierarchical and temporal order forming the representation of a human activity. Experimental results on three benchmark datasets demonstrate that the proposed approach with an off-the-shelf classification algorithm achieves highly competitive performance in comparison with the state-of-the-art approaches.

Index Terms—Activity Recognition, Manifold Dividing, Major Posture Feature, Main Dynamic Tendency

1 Introduction

Human activity recognition has been one of the most popular research in computer vision field [1]–[3] for its wide applications, including video surveillance, human-computer entertainment, health care and social assistance, and so on. The main objective of human action recognition is to enable machines to analyze and recognize human activities in videos automatically, which is associated with two issues, representing the spatio-temporal features and modeling dynamical patterns. Despite significant efforts have been made in the past decade, recognizing human action accurately is still a challenging problem.

Traditional studies for human action recognition focus on RGB videos [1]. However, this RGB video-based methods are sensitive to the inherent attributes of video frames, such as background clutter, partial occlusion, changes in scale, viewpoint, lighting, and appearance. Recently, the emergence of depth camera such as Microsoft Kinect and its corresponding real-time skeleton extraction method [4] promote the study of human action recognition from both aspects of depth maps-based methods [5]–[7] and skeleton-based methods [8]–[10]. Compared with 2D frame images, depth images reflect pure geometric or shape clues and are insensitive to lighting conditions, which are more robust in practical application. Different from approaches based on features extraction from RGB or depth images, skeleton-based methods require 3D coordinates of human body joints (human posture) and then model posture sequences to recognize human action. Yao et al. [11] validated that 3D skeleton information outperforms other low-level appearance features for human activity recognition.

In the skeleton-based representation, a human body is viewed as an articulated system of rigid segments, which are connected by several joints, and a human action is considered as a continuous evolution of the spatial configuration of these segments (i.e. postures) [12]. In this way, skeleton-based human activity recognition can be treated as a problem of structured time series analysis [13]. According to the approach of modeling temporal dynamics, existing skeleton-based action recognition methods can be divided into three main categories [10]: approaches based on latent variable models, approaches based on recurrent neural networks, and approaches based on temporal pyramid.

Latent variable model-based methods extract local features from each frame first, and then try to capture the dynamical patterns with sequential models such as hidden Markov model (HMM) [14], linear dynamic system (LDS) [15], and conditional restricted Boltzmann machine (CRBM) [16]. Recurrent neural network-based methods combine convolutional neural network (CNN) with recurrent neural network (RNN) to classify human action directly. CNN is employed to capture the multiscale features of individual or partial combination of body joints, while RNN is trained to learn the contextual information of action sequences [17]. Although these two kinds of methods can obtain excellent performance, they need a large amount of data samples and time epochs to estimate the large number of parameters accurately.

Different from the above approaches, temporal pyramid-based methods capture the temporal dependency of joint locations by dividing the action sequence into several sub-sequences, and every sub-sequence is represented by cor-

• *Y. Yao, Y. Liu, and H. Chen are with School of Computer Science and Technology, University of Science and Technology of China (USTC), Hefei, Anhui 230027, China. E-mail: {yaoyaq,ready}@mail.ustc.edu.cn, hchen@ustc.edu.cn. Z. Liu is with Department of Science and Technology Teaching, China University of Political Science and Law (CUPL), Beijing 100088, China. E-mail: lzhy@cupl.edu.cn. Corresponding Author: Huanhuan Chen.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2803838, IEEE Transactions on Big Data

responding discriminative features [18]–[20]. For example, Wang et al. [19] proposed to describe the 3D appearance with local occupancy pattern (LOP) features, which can capture the interaction between a human body and associative objects, and then Fourier temporal pyramid (FTP) is utilized to represent the temporal structure. Recently, a Covariance descriptor on 3D Joint locations with temporal hierarchy construction (Cov3DJ) was proposed in [20]. The authors first represent a sequence with the covariance matrix of skeleton joints over time, and then model temporal dynamics with multiple covariance matrices of sub-sequences in a hierarchical way.

Despite its simplicity, Cov3DJ outperformed complicated methods such as random occupancy patterns [21] and actionlets ensemble [19] on several datasets. On the other hand, Elgammal and Lee [22] proposed learning the view-based representation of human activity using manifold embedding, and indicated that temporal relation could not preserve the geometric structure of the manifold. Therefore, there are still some problems for Cov3DJ method. First and foremost, dividing a sequence into several sub-sequences with equal length would destroy the geometric structure of action snippets within an activity. Second, Cov3DJ ignores the mean vector of skeleton sequence, which seems useless at first sight, but each mean vector of sub-sequences can be viewed as a major posture of one action. Finally, the covariance matrix used in Cov3DJ contains some redundant information, and using vectorized covariance matrix as features might impact the discriminative ability of descriptor.

To resolve the above problems, we propose a novel and intuitive approach to recognize human activity in this paper. First, based on a realistic assumption that adjacent postures in action sequence is more similar and activity can be depicted with several action snippets, we design an intuitive dividing algorithm combining the advantages of temporal relation and manifold learning to divide an action sequence into two compact sub-sequences (action snippets), where the compactness is measured by linear perturbation of postures in a sub-sequence [23]. Second, a major posture feature (MTF), which is computed by the mean of skeleton sequence, is appended to the descriptor to enhance the discriminative capacity. Finally, instead of vectorizing the whole covariance descriptor, a main dynamical tendency feature (MPF) that combines the top-T eigenvectors of the obtained covariance matrix is constructed to improve the robustness of the descriptor. The concatenation of MTF and MPF consists of the proposed posture tendency descriptor (PTD), which is interpretable and discriminative for human activity recognition. The dividing algorithm leads to a hierarchical temporal description of an initial human action sequence. We deploy multiple PTDs over an entire action sequence and its sub-sequences, and the final representation of a human activity is the concatenation of PTDs in hierarchical and temporal order. With the help of an off-the-shelf classification method, we conduct experiments on three publicly available benchmarks, including the KARD dataset [24], the UTKinect dataset [25], and the Florence3D dataset [26]. The results demonstrate our proposed approach outperforms the state-of-the-art approaches. In summary, the major contributions of this paper are:

1) We propose an intuitive dividing algorithm to divide the action sequence into several meaningful action snippets, which can preserve the geometric structure within the activity.
2) An interpretable and discriminative posture tendency descriptor consisted of major posture feature and main dynamical tendency feature is designed to represent the action snippet.
3) Experimental results demonstrate that the proposed approach with an off-the-shelf classification algorithm achieves competitive performance on several benchmark datasets.

The remainder of this paper is organized as follows. We review the skeleton-based human activity recognition methods and manifold learning in Section 2. Section 3 gives an introduction to some preliminaries, the problem formulation of skeleton-based human activity recognition and manifold assumption on human action sequence. The proposed approach is presented in Section 4, including posture tendency descriptor construction and hierarchical temporal dividing. Section 5 discusses the experimental studies, and finally, the paper is concluded in Section 6.

2 Related Work

The proposed approach is based on the representation and measurement of 3D skeleton sequence with manifold dividing, therefore, we review some related works of human activity recognition based on skeleton sequence, and the applications of manifold learning on activity recognition in this section.

2.1 Activity Recognition with Skeleton Sequences

With the emergence of depth cameras such as Microsoft Kinect and corresponding real-time skeleton extraction methods [4], acceptable accuracies of 3D joint locations have led to a boom in research into human activity recognition. Johansson et al. demonstrated that human actions can be identified by the joint position of 3D skeleton individually [27]. Skeleton-based action sequence can be treated as the evolution of human skeleton. In terms of the methods of modeling temporal dynamics, existing skeleton-based approaches can be divided into three main classes [10], approaches based on latent variable models, approaches based on neural networks, and approaches based on hierarchical temporal pyramid.

Latent variable model-based approaches extract discriminative features from individual frame first, and then try to capture the dynamical patterns with transitions between latent variables such as HMM [13], [14], [28], LDS [15], and CRBM [16]. For example, by replacing Gaussian mixture models with deep neural networks in the emission part of HMM, Wu and Shao [28] extracted high level features of 3D joint positions in a generative way to learn the emission probabilities, which is demonstrated to be beneficial to action sequences inference. In consideration of the complex non-linearity contained in the components of motion sequences, Taylor et al. [16] proposed a model with distributed hidden state. In particular, to incorporate the temporal information in motion sequences, the authors

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2803838, IEEE Transactions on Big Data

extend the RBM to a conditional RBM by adding two kinds of directed connections, connections from past visible units to current visible units and connections from past units to current hidden units.

Recurrent neural networks-based approaches model the temporal relationship with directed cyclical connections and classify human action sequences with RNN directly [10], [29]–[31]. Du et al. [10] proposed a hierarchical bidirectional recurrent neural network (BRNN) to perform action recognition. The entire skeleton is divided into five parts first, and then the representations obtained from previous layers are merged as the inputs of next layers. In order to capture the salient dynamics, Veeriah et al. [29] proposed a differential RNN to learn the salient spatio-temporal information contained in actions. However, the importance of joints varies over different frames, and the conventional LSTM does not consider this attention mechanism. Some work brings attention mechanism into LSTM [32], [33]. On the other hand, other work employs CNN to capture the multiscale features of individual or partial combination of body joints, and models contextual information of action sequences with RNN [17], [34].

The above approaches based on latent variable models and neural networks are widely used for human activity recognition in compute vision community, however, there are some problem with these two approaches. In the latent variable model-based approaches, the model parameters of HMM and LDS are often learned with EM algorithm, which is sensitive to initialization and easy to get stuck in local optimal solutions. As for neural networks-based approaches, the model training process need a large amount of data samples and is very time and computation resource consuming.

Temporal pyramid-based approaches model the temporal dependency of joint locations by dividing the action sequence into several sub-sequences, and then represent the sequence with the concatenation of discriminative features extracted from these sub-sequences [19], [20], [35]. Wang et al. [19] proposed describing the 3D appearance with LOP feature, which can capture the interaction between human body and associative objects, and then FTP was utilized to represent the temporal structure. A dictionary learning method combined with temporal pyramid matching was proposed in [35] to learn the set of representative features and capture the temporal information respectively. Moreover, an actionlet ensemble model that combined features of a subset of joints is proposed to enhance the discriminative ability. Hussein et al. [20] first represented a sequence with the covariance matrix of skeleton joints over time, and then modeled temporal dynamics with multiple covariance matrices of sub-sequences in a hierarchical way.

The posture tendency descriptor in our approach is motivated by Hussein et al. [20]. However, the posture tendency descriptor not only enhance the discriminative capacity of the descriptor by taking into consideration of the mean of skeleton sequence, but also improve the robustness of the descriptor by replacing the vectorized covariance matrix with the top-T eigenvectors of the covariance matrix.

On the other hand, there is a trend of representing skeleton-based action sequences as curves in Lie group in recent years [9], [36], [37]. Vemulapalli et al. [9] modeled the relative 3D geometric relationships of skeleton joints with rotations and translations in a matrix Lie group, and then combine dynamic time warping and Fourier temporal pyramid to capture the temporal dynamics. Considering that human activities represented with high-dimensional trajectories in Lie group are in a non-Euclidean space, the authors combined logarithm map with rolling map to unwrap these curves onto the Lie algebra to perform classification in [36].

2.2 Manifold Learning and Action Recognition

In a nutshell, the manifold can be seen as a superposition of many local patches, which are constructed by a series of low-dimensional planes embedding into a high-dimensional space. The main purpose of manifold learning is extracting a set of low-dimensional bases that reflect the inherent dimensions of the manifold to describe high-dimensional data. Two classic manifold learning algorithms named Isometric Feature Mapping (Isomap) and Local Linear Embedding (LLE) were proposed by Tenenbaum et al. [23] and Roweis et al. [38], respectively.

In Isomap [23], the geodesic distance between pairwise points are computed by Dijkstra's algorithm based on k-nearest neighbors graph, and Multi Dimensional Scale analysis (MDS) are employed to process distance data for extracting low-dimensional bases. Isomap maintains the global structure of the manifold and requires less parameters, but the algorithm is sensitive to noise and has a high time complexity. Different from Isomap, to characterize the local geometric properties of a manifold, LLE [38] used the linear representation of the local point under its least squares and the final form combined all of the local geometric properties (linear representation coefficients). Moreover, LLE is additive and has low time complexity, which is also invariant to translation and rotation.

Manifold based representation and related algorithms have also made great progress in the field of human activity recognition [22], [39]–[42]. Instead of modeling temporal dynamics in sequence explicitly, manifold based learning methods embed the sequence into a low-dimensional space by preserving the local geometric properties. Before the emergence of cost-effective depth cameras, the human silhouettes are commonly used for activity recognition [22], [39]. For example, Elgammal et al. [22] proposed learning view-based representation of activity manifolds and infer 3D body poses directly from human silhouettes. The activity manifolds embedding was implemented with LLE. Skeleton 3D joint positions provide more intrinsic and robust motion structure for activity recognition [40], [42]. In particular, taking into consideration of temporal dimension, Gong et al. [40] proposed a model named Spatio-Temporal Manifold (STM) to analyze non-linear multivariate time series with latent spatial structure, which is utilized to recognize actions in the joint-trajectories space.

Based on the intuitive assumption that adjacent postures in action sequence are more similar and activity can be depicted with several action snippets, the dividing algorithm in our approach combines the advantages of temporal relation and manifold learning to divide an action sequence into two compact action snippets. This dividing algorithm can preserve the geometric structure of action snippets within an activity.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2803838, IEEE Transactions on Big Data

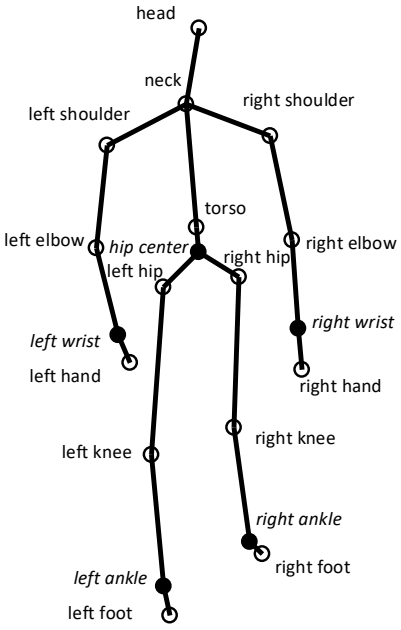

Fig. 1. Human skeleton labeled with 20 joints (5 optional joints are indicated with "•" and others are denoted with "◦").

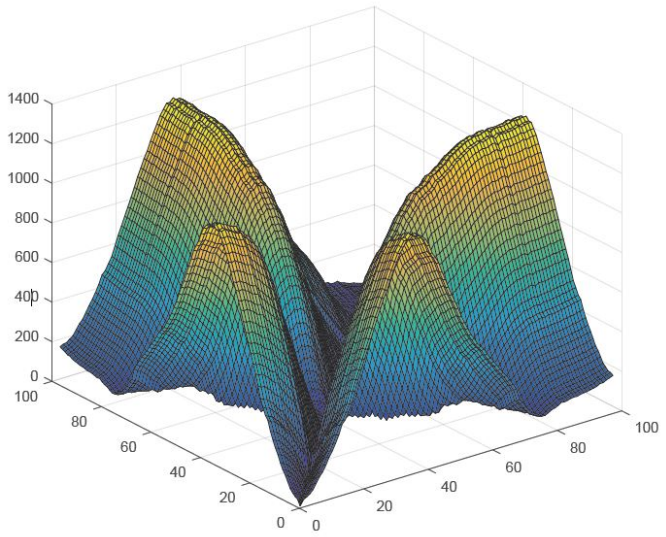

Fig. 2. An illustration of manifold assumption on human action sequence that a human activity denoted by a series of ordered points in the high dimensional space extends to a manifold.

3 PRELIMINARIES

This section presents preliminaries for the proposed approaches, the 3D coordinates of human body joints and manifold assumption on human action sequence. The joints coordinates describe a human body posture, which characterizes the transformation invariance of human activity. The posture sequence of a human activity is regarded as a manifold which captures the intrinsic structure of motion. These preparations are the basis of our proposed approach for human activity recognition.

3.1 Human Activity Representation

The human body is an articulated system of rigid segments connected by joints, and human activity is considered as a continuous evolution of the spatial configuration of these segments (i.e. body postures) [12]. In this paper, we build a representation of human activity based on a sequence of 3D joint locations. With the emergence of depth camera Microsoft Kinect and corresponding real-time skeleton extraction method [4], 15 (or 20) acceptable accurate skeletal 3D joint locations are provided: head, neck, torso, L/R shoulders, L/R elbows, L/R hands, L/R hips, L/R knees, L/R feet (hip center, L/R wrists and L/R ankles), as illustrated in Fig. 1.

For a human activity video, there are J joint locations and each joint j has 3 coordinates $\hat{\boldsymbol{p}}_f^j = [x_f^j, y_f^j, z_f^j]^T$ at the f-th frame. For each joint j, we extract the relative position to joint *hip center* by taking the difference between the coordinates of joint j to hip center joint c

$$\boldsymbol{p}_f^j = \hat{\boldsymbol{p}}_f^j - \hat{\boldsymbol{p}}_f^c \quad (j = 1, ..., J).$$

In the case of $J = 20$, $\hat{\boldsymbol{p}}_f^c$ is obtained from the tracked coordinates of *hip center* directly, while for $J = 15$, we set $\hat{\boldsymbol{p}}_f^c$ as the average coordinates of *left hip* and *right hip*. The 3D joint positions for frame f is defined as

$$\boldsymbol{p}_f = [\boldsymbol{p}_f^1; \boldsymbol{p}_f^2; \cdots; \boldsymbol{p}_f^J],$$

where the derived $\boldsymbol{p}_f \in \mathbb{R}^{3J}$ describes human action state (posture) at frame f. Yao et al. [11] validated that the pose-based features outperform other low-level appearance features. In addition, pose-based features is a natural and intrinsic representation for human activity as it conforms to the study of how human understand actions [43]. Finally, a human activity is represented by a sequence of postures $\boldsymbol{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_F\}$, where each posture $\boldsymbol{p}_f$ is a high dimensional vector consisting of a set of body joint locations, and F is the number of postures.

3.2 Manifold Assumption on Action Sequence

For a human action sequence during a certain period of time, the actual number of human joints participated in and the direction of movement in the three-dimensional space is limited. In other words, the human activity can be described by part of the dimensionality of data, which forms a low-dimensional plane and has the Euclidean space property. On the other hand, human joints and their movement direction associated with an action sequence vary smoothly over time. Therefore, an action sequence can be decomposed into several sub-sequences, which are defined as action snippets in this paper. In summary, a human activity denoted by a series of ordered points in the high dimensional space extends to a manifold, which consists of several local planes (action snippets).

Each action snippet C is a local linear patch and embodies Euclidean property. We resort to local linearity to extract representative action snippets from a human action posture sequence. Specifically, the local linearity is determined by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2803838, IEEE Transactions on Big Data

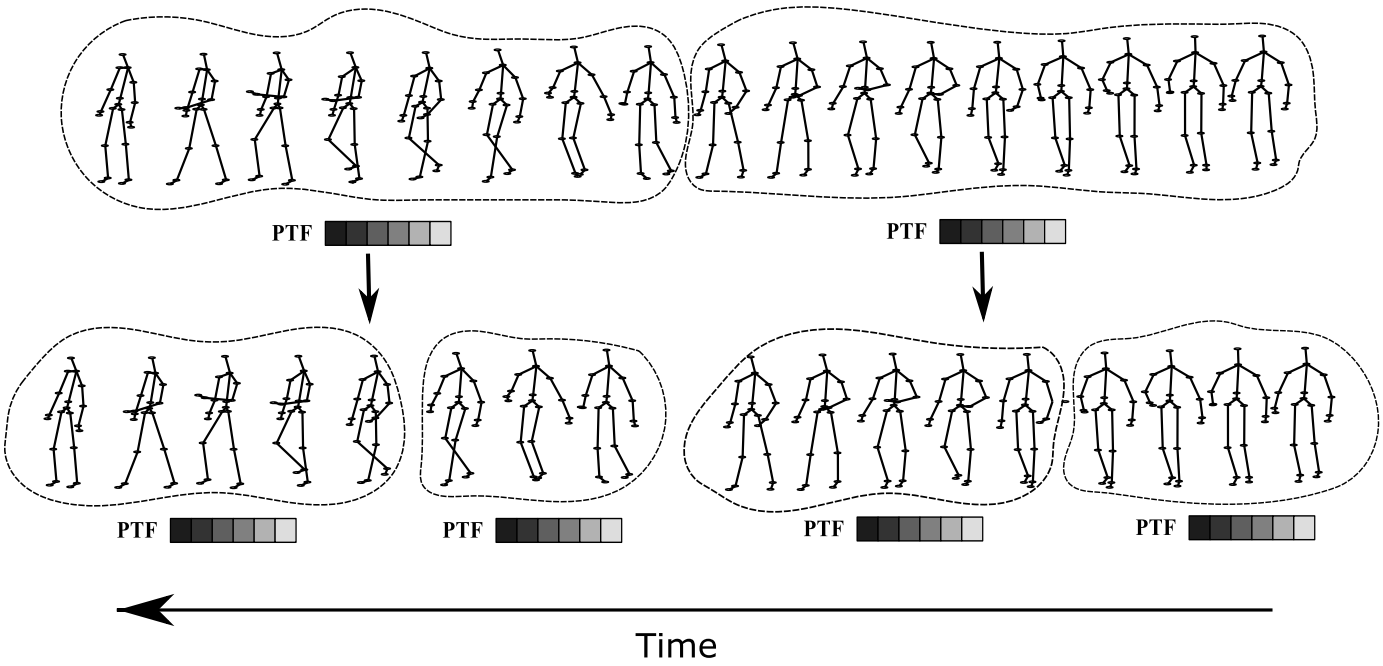

Fig. 3. An illustration of posture tendency descriptor (PTF) on action snippets. The proposed dividing algorithm leads to a meaningful segmentation on sequence of action $walk$ of the second subject from UTKinect Dataset. The action snippets in dotted local patches indicate the separation in the next level. The length of the posture sequence is 54, and one third of those postures is uniformly sampled (i.e. the $\{1, 4, \cdots, 52\}$-th posture) to make the illustration clear and not overlapping. In addition, the 52-th posture is ignored due to its distortion in 2D plane. The result shows that the proposed algorithm can preserve the completeness of action snippets. In particular, the posture sequence is first divided into two action snippets, each of which corresponds to a continuous movement of left and right foot, and then each action snippet is divided into two action snippets that correspond to the movement of left foot and right foot, respectively. The traditional dividing algorithm would segment the posture sequence into two sub-sequence of the same length and destroy the geometric structure of action snippets. Multiple PTDs are deployed over the entire action sequence and its sub-sequences to construct the final representation.

the ratio of the geodesic distance to the Euclidean distance of all pairwise posture points in C [44],

$$\beta = \frac{1}{|\mathrm{C}|^2} \sum_{i \in \mathrm{C}} \sum_{j \in \mathrm{C}} r(\boldsymbol{p}_i, \boldsymbol{p}_j), \tag{1}$$

where $|\mathrm{C}|$ denotes the number of postures in an action snippet C, and $r(\boldsymbol{p}_i, \boldsymbol{p}_j)$ is the ratio between posture points $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$,

$$r(\boldsymbol{p}_i, \boldsymbol{p}_j) = \frac{d_G(\boldsymbol{p}_i, \boldsymbol{p}_j)}{d_E(\boldsymbol{p}_i, \boldsymbol{p}_j)}. \tag{2}$$

Euclidean distance $d_E(\boldsymbol{p}_i, \boldsymbol{p}_j)$ is computed by the L_2 norm and the geodesic distance $d_G(\boldsymbol{p}_i, \boldsymbol{p}_j)$ is computed by Dijkstra's algorithm based on k nearest neighbors graph, which is defined as:

Definition 1. *k **nearest neighbors** (k**-NN**) **graph**: A connected graph where each node links to its nearest k nodes based on Euclidean distance.*

k-NN graph is able to construct the feature relations among disordered images and is widely used in Face Recognition with Image Sets (FRIS) problem. However, the construction of k-NN graph neglects the order of image sequence, it can not be used for computing geodesic distance of postures in sequence since the temporal relationship of human postures is significant for activity recognition.

As we have mentioned, a human activity is represented by a sequence of postures $\boldsymbol{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_F\}$, and continuous postures tend to have smaller Euclidean distance. As illustrated in Fig 2, the x, y axes in the horizontal plane denote frame number, and the vertical height is the distance between frames. The concave diagonal validates the hypothesis that adjacent frames have more similar feature. Therefore, for the purpose of holding the temporal relationship of human postures, we present the following definition of k sequential neighbors graph to replace k-NN graph:

Definition 2. *k **sequential neighbors** (k**-SN**) **graph**: A connected graph where each node links to its previous and next k nodes based on action postures order.*

All geodesic distances of pairwise human postures in this paper are computed by Dijkstra's algorithm based on the proposed k-SN graph.

4 The Proposed Approach

In this approach, the initial 3D skeleton sequence of a human activity is represented in a hierarchical temporal fashion, as illustrated in Fig. 3. For a specific human action snippet, we compute the mean and the covariance matrix for the sub-sequence, where the mean represents the major posture feature (MPF) and the main dynamical tendency feature (MTF) is characterized by top-T eigenvectors of the obtained covariance matrix. Our proposed posture tendency descriptor (PTD) is the concatenation of MPF and MTF, which is interpretable and discriminative for human activity recognition.

To encode the temporal information, we design a dividing algorithm to divide an action sequence into two compact

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2803838, IEEE Transactions on Big Data

sub-sequences (action snippets), where the compactness is measured by the linear perturbation of postures in a sub-sequence [23]. The dividing algorithm leads to a hierarchical temporal description of an initial human action sequence. We deploy multiple PTDs over an entire action sequence and its sub-sequences, and the final representation of a human action is the concatenation of PTDs in hierarchical and temporal order.

4.1 Posture Tendency Descriptor

For a human activity formulated as a sequence of postures $\boldsymbol{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_F\}$, the MPF $\bar{\boldsymbol{p}}$ is defined as the average of all posture vectors,

$$\bar{\boldsymbol{p}} = \frac{1}{F}\sum_{f=1}^{F} \boldsymbol{p}_f, \tag{3}$$

where the derived $\bar{\boldsymbol{p}} \in \mathbb{R}^N$, $N = 3 \times J$. MPF describes the most representative posture of corresponding posture sequence.

The sample covariance matrix is given by the following equation,

$$\boldsymbol{C}_p = \frac{1}{F-1}\sum_{f=1}^{F} (\boldsymbol{p}_f - \bar{\boldsymbol{p}})(\boldsymbol{p}_f - \bar{\boldsymbol{p}})^T. \tag{4}$$

The obtained $\boldsymbol{C}_p \in \mathbb{R}^{N\times N}$ is a symmetric matrix. Since every symmetric matrix can be orthogonally diagonalizable, there exists an orthogonal matrix $\boldsymbol{Q}$ such that,

$$\boldsymbol{Q}^{-1}\boldsymbol{C}_p\boldsymbol{Q} = \boldsymbol{\Lambda}, \tag{5}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with real eigenvalues of $\boldsymbol{C}_p$ on the diagonal elements, and $\boldsymbol{Q}$ is the eigenvectors corresponding to the eigenvalues in $\boldsymbol{\Lambda}$, the MTF $\boldsymbol{T}$ is defined as the top-T eigenvectors,

$$\boldsymbol{T} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_T], \tag{6}$$

where $\boldsymbol{q}_t \in \mathbb{R}^N$ is the eigenvector corresponding to the t-th largest eigenvalue.

During a certain period of time, the actual number of human joints participated in a specific human action snippet is limited, as well as the direction of movement in the three-dimensional (x, y, and z) space. For example, the participated human body joints of action "draw circle" are left hand/wrist/elbow (or right hand/wrist/elbow), and the main motion direction is in the x and y directions (where z-axis is in the direction perpendicular to the camera's imaging plane), so the action can be characterized by the changes in directions of x and y axes of the associated human body joints. In this respect, the proposed definition of MTF is capable of describing the movement trend of human action snippet.

MPF indicates the major posture and MTF denotes the main dynamical tendency of an action snippet. The proposed PTD is the combination of MPF and MTF,

$$\boldsymbol{d}_{PT} = [\bar{\boldsymbol{p}}; \boldsymbol{q}_1; \boldsymbol{q}_2; \cdots; \boldsymbol{q}_T], \tag{7}$$

Thus the dimensionality of our proposed PTD is $N + T \times N$. A larger T means more information of the original sample covariance matrix but with worse robustness. Our preliminary experiments indicate that $T = 2$ is a good trade-off between accuracy and robustness, and in this case, the dimensionality of PTD under the skeleton recorded with 20 joints is $3 \times 20 + 2 \times (3 \times 20) = 180$.

Compared with the Cov3DJ descriptor presented in [20], the proposed PTD is more discriminative with the supplement of MPF. On the other hand, the low-dimensional MTF consisted of principal eigenvectors is more compact and robust. Therefore, PTD is an interpretable descriptor that takes both MPF and MTF of an action sequence into consideration.

4.2 Hierarchical Temporal Dividing Algorithm

The proposed PTD captures both MPF and MTF information for a human action sequence. However, it ignores the order of action snippets in a human action. For example, "$stand\ up$" and "$sit\ down$" are two actions in reverse order of motions, but their PTDs are identical even though they are totally different actions. To encode the temporal information, an intuitive and practical dividing algorithm is proposed to divide a sequence into two compact sub-sequences (action snippets). The compactness of snippet is measured by linear perturbation of postures, which is naturally reflected by the deviation between Euclidean distances and geodesic distances.

A human activity represented by a postures sequence $\boldsymbol{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_F\}$ is divided into two action snippets corresponding to two sub-sequences $\boldsymbol{P}_L$ and $\boldsymbol{P}_R$ using the proposed dividing algorithm, as described in Algorithm 1. We utilize two indexes l and r to indicate two positions on initial sequence, thus the current lengths of left sub-sequence $\boldsymbol{P}_L$ and right sub-sequence $\boldsymbol{P}_R$ are $M_L = l-1$ and $M_R = F - r$, respectively. At the start of dividing, both the nonlinear scores and the length of sub-sequences are equal, and we concatenate $\boldsymbol{p}_l$ to the left sub-sequence $\boldsymbol{P}_L$ in this case. If the nonlinear score $\beta_L < \beta_R$, we concatenate $\boldsymbol{p}_l$ to the left sub-sequence $\boldsymbol{P}_L$ and update the related parameters. Otherwise, if the nonlinear score $\beta_L > \beta_R$, we concatenate $\boldsymbol{p}_r$ to the right sub-sequence $\boldsymbol{P}_R$. In the case of $\beta_L = \beta_R$, we concatenate posture to the shorter sub-sequence in order to balance their length. Specifically, we concatenate $\boldsymbol{p}_l$ to the left sub-sequence $\boldsymbol{P}_L$ if $M_L < M_R$, and concatenate $\boldsymbol{p}_r$ to the right sub-sequence $\boldsymbol{P}_R$ if $M_L > M_R$. The Euclidean distance matrix $\boldsymbol{D}_{EL}$, the geodesic distance matrix $\boldsymbol{D}_{GL}$, and the distance ratio matrix $\boldsymbol{R}_L$ for the left sub-sequence are all $M_L \times M_L$ matrices. After concatenating $\boldsymbol{p}_l$ to $\boldsymbol{P}_L$, the expansion of $\boldsymbol{R}_L$ can be illustrated as follows,

$$\begin{pmatrix} r_{11} & \cdots & r_{1,l-1} \\ \vdots & \ddots & \vdots \\ r_{l-1,1} & \cdots & r_{l-1,l-1} \end{pmatrix} \Longrightarrow \begin{pmatrix} r_{11} & \cdots & r_{1,l-1} & r_{1l} \\ \vdots & \ddots & \vdots & \vdots \\ r_{l-1,1} & \cdots & r_{l-1,l-1} & r_{l-1,l} \\ r_{l,1} & \cdots & r_{l,l-1} & r_{l,l} \end{pmatrix}$$

where $r_{ij} = r(\boldsymbol{p}_i, \boldsymbol{p}_j)$. The matrices for the right sub-sequence are expanded in the same way.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2803838, IEEE Transactions on Big Data

Algorithm 1 Hierarchical Temporal Dividing Algorithm.

Input: A human activity represented by postures sequence $\boldsymbol{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_F\}$; the sequential neighbors k.
Output: Two sub-sequences $\boldsymbol{P}_L = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_d\}$ and $\boldsymbol{P}_R = \{\boldsymbol{p}_{d+1}, \boldsymbol{p}_{d+2}, \cdots, \boldsymbol{p}_F\}$ represent the left and the right part of the initial action sequence, respectively.
Initialization:
$\boldsymbol{P}_L \leftarrow \boldsymbol{p}_1$, $\boldsymbol{P}_R \leftarrow \boldsymbol{p}_F$;
Compute Euclidean distance matrix $\boldsymbol{D}_{EL}$ and $\boldsymbol{D}_{ER}$;
Compute geodesic distance matrix $\boldsymbol{D}_{GL}$ and $\boldsymbol{D}_{GR}$;
Compute distance ratio matrix $\boldsymbol{R}_L$ and $\boldsymbol{R}_R$;
$l \leftarrow 2$, $r \leftarrow F-1$, $\beta_L \leftarrow 1$, $\beta_R \leftarrow 1$;
while $l < r$ **do**
  if $\beta_L < \beta_R$ **then**
    Update $\boldsymbol{P}_L \leftarrow [\boldsymbol{P}_L; \boldsymbol{p}_l]$;
    Expand $\boldsymbol{D}_{EL}$, $\boldsymbol{D}_{GL}$, $\boldsymbol{R}_L$ to include $\boldsymbol{p}_l$;
    Compute β_L with Eq. (1);
    $l \leftarrow l+1$;
  else if $\beta_L > \beta_R$ **then**
    Update $\boldsymbol{P}_R \leftarrow [\boldsymbol{P}_R; \boldsymbol{p}_r]$;
    Expand $\boldsymbol{D}_{ER}$, $\boldsymbol{D}_{GR}$, $\boldsymbol{R}_R$ to include $\boldsymbol{p}_r$;
    Compute β_R with Eq. (1);
    $r \leftarrow r-1$;
  else
    if $M_L \leq M_R$ **then**
      Execute line 6—9;
    else if $M_L > M_R$ **then**
      Execute line 11—14;
    end if
  end if
end while
return $\boldsymbol{P}_L$, $\boldsymbol{P}_R$;

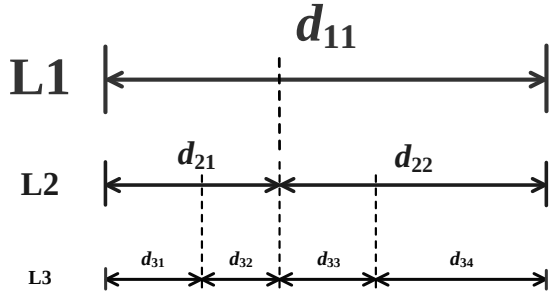

Fig. 4. Hierarchical temporal description of human action sequence. $\boldsymbol{d}_{li}$ is the i^{th} PTD in the l^{th} level over corresponding sub-sequence.

The expansion step makes the algorithm efficient since we just compute the relative distances between the current posture $\boldsymbol{p}_l$ (or $\boldsymbol{p}_r$) and each posture in $\boldsymbol{P}_L$ (or $\boldsymbol{P}_R$). Besides, the defined k sequential neighbors graph leads to a faster computation of $\boldsymbol{D}_G$ as it is not necessary to use Dijkstra's algorithm, so the computation of β_L and β_R can be efficiently carried out. Moreover, the balance step is applied to ensure a more balanced dividing, that is, if $\beta_L = \beta_R$, we tend to increase the shorter sub-sequence and update relevant information. Compared with the common average dividing method [20], [45], the proposed dividing algorithm takes full account of the internal geometric structure of human activity and enhances the interpretability, which leads to better experimental performance.

The dividing algorithm obtains a hierarchical temporal description of an initial human action sequence as shown in Fig. 4, where three levels are annotated. The first level PTD $\boldsymbol{d}_{11}$ is computed over the entire action sequence, and the second level PTDs $\boldsymbol{d}_{21}$ and $\boldsymbol{d}_{22}$ are computed over two divided sub-sequences, and so on. We deploy multiple PTDs over an entire action sequence and its sub-sequences, the final representation of a human action is the concatenation of PTDs in hierarchical and temporal order $[\boldsymbol{d}_{11}; \boldsymbol{d}_{21}; \boldsymbol{d}_{22}; \boldsymbol{d}_{31}; \cdots]$. It can be conjectured that adding more levels would enhance the discriminative ability of hierarchical temporal PTDs to represent a human activity, where both coarse (level one) and detailed (level two, three, and et al.) descriptions are taken into consideration. However, more levels increase the dimensionality of obtained features, and may impact classification performance due to over-fitting. Our preliminary experiments indicate that the level $L = 3$ is a good compromise between accuracy and generalization. In this case, a skeleton recorded with 20 joints and a MTF with $T = 2$ obtain the final action representation feature, whose dimensionality $(1+2+4) \times 180 = 1260$ is much lower than Cov3DJ.

Note that the sequence can be divided into more than two subsequences by computing the entire distance ratio matrix, in which case, a threshold parameter $\sigma > 1$ needs to be set in advance. Starting from the first posture, the dividing algorithm concatenates the current posture $\boldsymbol{p}_t$ to the current subsequence $\boldsymbol{P}_c$ (initialized as empty set) and calculates the nonlinear score β of the subsequence $[\boldsymbol{P}_c; \boldsymbol{p}_t]$. If nonlinear score $\beta_c \leq \sigma$, concatenate the current posture to the current subsequence $\boldsymbol{P}_c = [\boldsymbol{P}_c; \boldsymbol{p}_t]$. Otherwise, the current subsequence is separated from the entire sequence to form an action snippet, and the dividing algorithm restarts from the current posture. This process continues until the entire sequence is divided into action snippets. Compared with the proposed dividing algorithm, the number of action snippets obtained with this dividing algorithm is sensitive to the threshold parameter σ, whose value varies to different action sequence and is difficult to set appropriately.

4.3 Complexity Analysis

The computational complexity of the proposed approach is composed of the following three parts.

1) The complexity of computing sample covariance matrix $\boldsymbol{C}_p$ of a sequence of postures of length F is $\mathcal{O}(FN^2)$, where N is number of coordinates of all joints.
2) After obtaining the samples covariance matrix, the complexity of eigendecomposition of $\boldsymbol{C}_p$ is $\mathcal{O}(N^3)$.
3) Since the complexity of computing geodesic distance matrix of $F \times F$ is $\mathcal{O}(F^2)$, the computational complexity of dividing algorithm is $\mathcal{O}\left(\sum_{n=1}^{F} n^2\right) = \mathcal{O}\left(F^3\right)$.

The number of coordinates of all joints $N = 3 \times 20 = 60$, which is smaller than the length of a sequence of postures

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2803838, IEEE Transactions on Big Data

TABLE 1
Parameters search scopes for all methods.

Methods	Parameters	Parameter Scope
Gaglio et al. [24]	k, N	$k \in \{15, 16, \cdots, 51\}$, $N \in \{3, 4, \cdots, 17\}$
Cippitelli et al. [46]	k	$k \in \{3, 5, 10, 15, \cdots, 35\}$
Xia et al. [25]	k, N	$k = 125$, $N = 6$
Hussein et al. [20]	L	$L = 3$

k is the number of clusters.
N is the number of the HMM states.
L is the number of levels of descriptor.

TABLE 2
Activity sets grouping different and similar activities from the KARD dataset. Actions are in bold.

Activity Set 1	Activity Set 2	Activity Set 3
Horizontal arm wave	High arm wave	Draw tick
Two-hand wave	Side kick	**Drink**
Bend	**Catch cap**	**Sit down**
Phone call	Draw tick	**Phone call**
Stand up	Hand clap	**Take umbrella**
Forward kick	Forward kick	**Toss paper**
Draw X	Bend	High throw
Walk	**Sit down**	Horizontal arm wave

F. Therefore, the total complexity of the proposed approach is approximated by $\mathcal{O}\left(F^3\right)$.

5 Experimental Studies

We evaluate the discriminative ability of the proposed approach on three public available benchmarks, including the KARD dataset [24], the UTKinect dataset [25], and the Florence3D dataset [26], each of them provides 15 or 20 3D joint locations for the participating person. We compare our approach with state-of-the-art approaches on these three datasets, and carry out parameters analysis of the proposed approach on the KARD dataset. To remove center symmetry of eigenvectors, we take the absolute value of all MTFs, and then normalize the MPF and all MTFs to have unit L_2 norm before concatenation for training and testing. In all experiments, a linear SVM classifier is used for classification.

The hyperparameter settings for the baseline methods here follow settings in the original papers. In particular, the number of clusters k and the number of states of HMM N in [24] are set by grid search approach, and the search scope are $k \in \{15, 16, \cdots, 51\}$ and $N \in \{3, 4, \cdots, 17\}$ respectively. The search scope of the number of clusters k in [46] is $k \in \{3, 5, 10, 15, \cdots, 35\}$. The set of clusters and number of HMM states in [25] are taken $k = 125$ and $N = 6$, respectively. The number of levels of the descriptor in [20] is set to $L = 3$. The search scope for parameters of each method is summarized in Table 1.

5.1 KARD Dataset

Gaglio et al. [24] collected the KARD dataset that contains 18 activities, which are divided into ten gestures and eight actions as listed in Table 2 with different fonts. Ten gestures are simple sequences related to specific parts of a body,

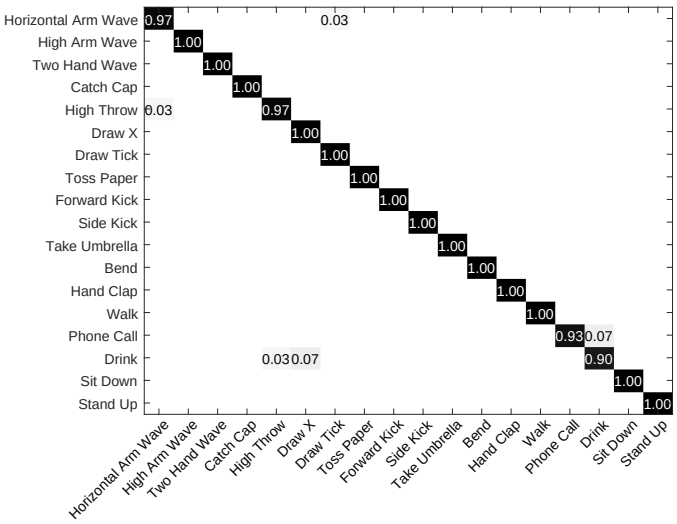

Fig. 5. Confusion matrix of the proposed approach using the KARD dataset under "new-person" setting. The x and y axis denote the actual label and predicted label of the sequences, respectively.

while eight actions involving interaction between different parts of the body are complex. Each activity is performed 3 times by 10 different performers, therefore, there are 540 ($18 \times 10 \times 3$) video sequences in total.

Previous work [24] uses three different experimental setups and two modalities of dataset splitting. The three experimental setups are:

1) *One-Third Setup* (A): One-third of the samples of each person are used for training and the rest is for testing.
2) *Two-Third Setup* (B): Two-thirds of the samples of each person are used for training and the rest is for testing.
3) *Half Setup* (C): Half of the samples of each person are used for training and the rest for is testing.

The activities constituting the dataset are divided into following groups:

1) Gestures and Actions.
2) Three subsets: Activity Set 1, Activity Set 2, and Activity Set 3 as listed in Table 2. From subset 1 to 3, the similarity of activities increases gradually.

We perform the experiments on three different experimental setups and five dataset splittings. All results are obtained under the parameter setting: $T = 2$, $k = 6$, $L = 3$ unless otherwise specified. Due to the randomness in the procedure of splitting training and testing data, each experimental setup is run 10 times. Table 3 presents the average classification accuracy. We can find that the proposed approach beats comparative methods on three activity subset and the gestures subset under all three experimental setups. In particular, the accuracy of the proposed approach on gestures subset is much higher than the other methods. Our approach is only inferior to the method proposed in [46] with a small disadvantage on the actions subset. The difference between performances on the gestures and actions subset is caused by following the reasons. The PTD is a linear descriptor, which has limited capacity to capture

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2803838, IEEE Transactions on Big Data

TABLE 3
Accuracy (%) of the proposed approach compared with other methods using the KARD dataset for different experimental setups with different Activity Sets and split in Gestures/Actions.

Methods \ Setups	Activity Set 1			Activity Set 2			Activity Set 3			Gestures			Actions		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Gaglio et al. [24]	95.1	99.1	93.0	89.9	94.9	90.1	84.2	89.5	81.7	86.5	93.0	86.7	92.5	95.0	90.1
Cippitelli et al. [46]	98.0	99.0	97.7	99.8	100	99.6	91.6	95.8	93.3	89.9	95.9	93.7	**99.0**	**99.9**	**99.1**
The Proposed Approach	**98.2**	**99.3**	**98.8**	**99.9**	**100**	**100**	**94.5**	**97.9**	**95.9**	**97.9**	**98.6**	**98.2**	98.1	99.4	97.9

TABLE 4
Accuracy (%) of the proposed approach compared with state-of-the-art methods, using the KARD dataset under "new-person" setting.

	Methods	Accuracy
Reported Results	Gaglio et al. [24]	84.8
	Cippitelli et al. [46]	95.1
	Hussein et al. [20]	96.8
Our Results	MPF + SVM	92.5
	MPF + HTDA + SVM	93.2
	MTF + SVM	95.7
	MTF + HTDA + SVM	96.1
	PTD + HTDA + SVM	**98.7**

TABLE 5
Accuracy (%) of the proposed approach compared with the state-of-the-art methods on UTKinect dataset.

Methods	Accuracy
Xia et al. [25]	90.92
Devanne et al. [47]	91.46
Wang et al. [48]	93.47
The Proposed Approach	**94.97**
Liu et al. [31]	95.00
Cippitelli et al. [46]	95.10

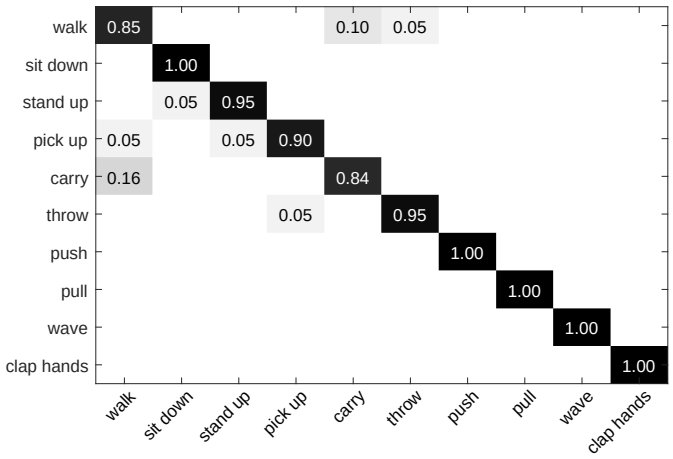

Fig. 6. Confusion matrix of the proposed approach using the UTKinect dataset under "leave-one-sequence-out" setting. The x and y axis denote the actual label and predicted label of the sequences, respectively.

some features of complex activities. On the other hand, there might be very similar activities in the actions subset.

In addition, "new-person" scenario, also known as leave-one-person-out setting is evaluated. The test results of this setting are in line with those of [46]. We implemented the code of Cov3DJ according to the paper [20], and report the best result by varying their parameters across all the possible range in this scenario. The results of the proposed approach compared with state-of-the-art approaches are listed in Table 4. Our approach achieves the best performance under the methods combination of "PTD + HTDA + SVM", where the abbreviation "HTDA" means the proposed hierarchical temporal dividing algorithm under nonlinear degree. The corresponding confusion matrix is illustrated in Fig. 5, from which we can observe that most activities obtain a high recognition accuracy except for a very small level of confusion among several activities. In particular, the confusion between activities *phone call* and *drink* is the main reason that leads to the lower performance in the experimental setup of "actions" in Table 3. Note that all misclassified activities belong to Activity Set 3, the main reason causing this confusion is that these activities are very similar from the perspective of 3D joint locations, and there is too little information available to avoid confusion between classes. On the whole, from the strong diagonal it is evident that we achieve encouraging recognition performance for the different human activities in this dataset.

5.2 UTKinect Dataset

UTKinect dataset [25] contains 10 kinds of human actions in indoor settings: *walk*, *sit down*, *stand up*, *pick up*, *carry*, *throw*, *push*, *pull*, *wave*, and *clap hands*. 10 different subjects perform each action twice, and there are 199 recorded human action sequences in total.

In this dataset, we adopt the leave-one-sequence-out cross validation experimental setup as employed in [25]. Since some recorded action sequences are too short to implement the hierarchical temporal dividing algorithm, we expand the action sequence by linear interpolation when the sequence length is less than 32. The use of linear interpolation is acceptable since it holds the intrinsic structure of the initial posture sequence. We compare the proposed approach with state-of-the-arts as listed in Table 5, which shows that our approach is superior to most of the comparative methods and inferior to [46] and [31] with a small disadvantage. The possible reasons are as follows. There are no much geometric structures preserved in very short

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2803838, IEEE Transactions on Big Data

TABLE 6
Accuracy (%) of the proposed approach compared with the state-of-the-art methods on Florence3D dataset.

Methods	Accuracy
Cippitelli et al. [46]	86.10
Devanne et al. [47]	87.04
Vemulapalli et al. [9]	90.88
The Proposed Approach	91.59
Ma et al. [17]	91.72
Wang et al. [48]	**92.25**

sequence, which limits the performance of hierarchical temporal dividing. However, the postures in [46] are selected with clustering and do not suffer from short sequence. Since spatio-temporal LSTM [31] considers the recurrent analysis in spatial domain and discovers the spatial dependency patterns between different joints in each frame, it could capture more discriminative dynamics and motion patterns than the proposed approach that simply concatenate the joints information. The attendant shortcoming of this RNN-based method is the high computational complexity. On the other hand, the linear interpolation would include some synthetic noise to some extent. Therefore, one possible limitation of the proposed method is that it might not be applicable to very short sequences. The best result of 94.97% is achieved under the parameters setting: $T = 2$, $k = 6$, $L = 3$. The corresponding confusion matrix is illustrated in Fig. 6. It is obvious that the proposed approach recognizes most of the actions in UTKinect dataset correctly.

5.3 Florence3D Dataset

Florence dataset [26] is captured using a Kinect camera and collected at the University of Florence. It includes 9 human actions: $wave$, $drink\ from\ a\ bottle$, $answer\ phone$, $clap$, $tight\ lace$, $sit\ down$, $stand\ up$, $read\ watch$, and bow. 10 different subjects perform each action 2 or 3 times, and there are 215 action samples in total. The high intra-class variations (the same action is performed using the left hand in some samples and right hand in some others) make this dataset more challenging.

In this dataset, we adopt the leave-one-subject-out cross validation experimental setup as employed in previous work [47], which means that the person for testing cannot appear in the training. Since there are some actions performed by left-hand in some instances, we modify the dataset by exchanging the y-coordinates of left and right hands (ankles and shoulders) when the y-coordinate of left hand is larger than right hand, this heuristic effectively transforms left-handed actions to right-handed actions. Moreover, we also expand the action sequence when the sequence length is less than 32 as we did in UTKinect dataset. The classification accuracy is presented in Table 6. The proposed approach obtains the accuracy of 91.59% under the parameters setting: $T = 2$, $k = 6$, $L = 2$, which is just lower than the best result by 0.66%. Due to the high intra-class variations, the proposed descriptor would not characterize the meaningful posture compared with key-pose-motifs obtained in [48]. On the other hand, taking into account of multiscale features in RNN with convolutional operation, Ma et al. [17] beats the proposed approach narrowly. Fig. 7 shows the corresponding confusion matrix, we can find that most of the actions in Florence dataset are recognized correctly by the proposed approach.

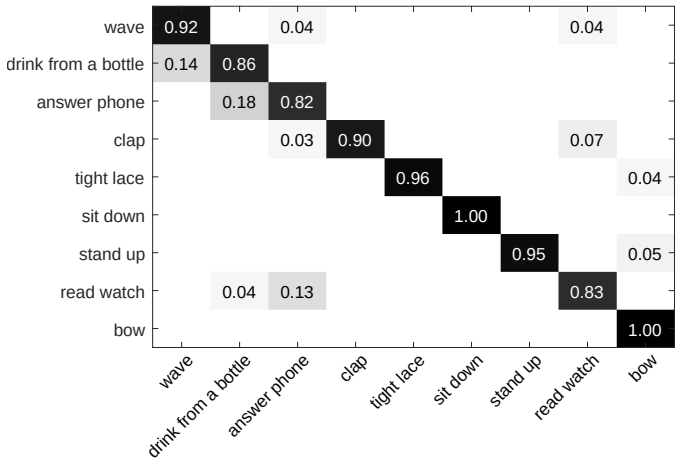

Fig. 7. Confusion matrix of the proposed approach using the Florence dataset under "leave-one-subject-out" setting. The x and y axis denote the actual label and predicted label of the sequences, respectively.

5.4 Parameter Analysis

Finally, we investigate the three key parameters in the proposed approach: MTF number T, hierarchical level number L and, sequential neighbors number k. We run all parameters settings on KARD dataset with leave-one-out scenario.

5.4.1 Analysis of MTF Number

To evaluate the influence of the MTF number T, we vary the sequential neighbors number k ranging from 2 to 13 with step 1 and fix the hierarchical level number $L = 3$. The MTF number T is specified to be 0, 1, 2, and 3, respectively (where $T = 0$ means only MPF included). Fig. 8 shows the corresponding result. We can observe that the performance is significantly improved by supplementing MTF to MPF. The main reason is that MTF is able to model the evolution of an action. Thus MPF and MTF complement each other and obtain the best performance. Moreover, compared with $T = 1$ and $T = 3$, $T = 2$ performs best in most cases, we attribute this phenomenon to two aspects: 1) $T = 1$ is not capable of capturing sufficient dynamical tendency information, while 2) $T = 3$ may lead to slight over-fitting.

5.4.2 Analysis of the Hierarchical Level Number

To evaluate the influence of the hierarchical level number L, we fix the MTF number $T = 2$ and vary the sequential neighbors number k ranging from 2 to 13 with step 1. The hierarchical level number L is specified to be 1, 2, 3, and 4, respectively. The result is presented in Fig. 9. It can be observed that there is a remarkable performance improvement when we divide the entire human action sequence into two sub-sequences (from $L = 1$ to $L = 2$). The main reason is that the hierarchical levels encode the temporal relationship, which is an important characteristic of a human action. Moreover, compared with $L = 2$ and $L = 4$, $L = 3$ performs

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2803838, IEEE Transactions on Big Data

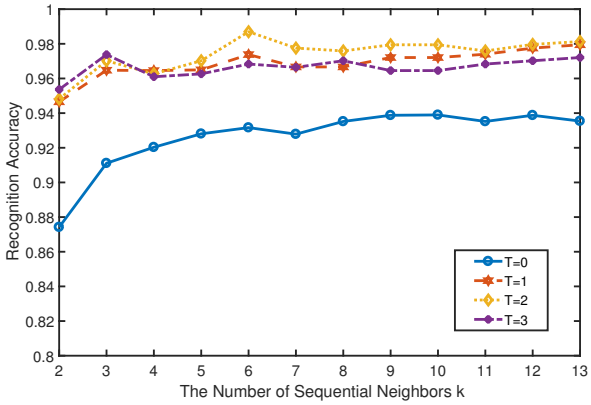

Fig. 8. The influence of the MTF number T with the hierarchical level number $L = 3$ on activity recognition accuracies using the KARD dataset under "new-person" setting.

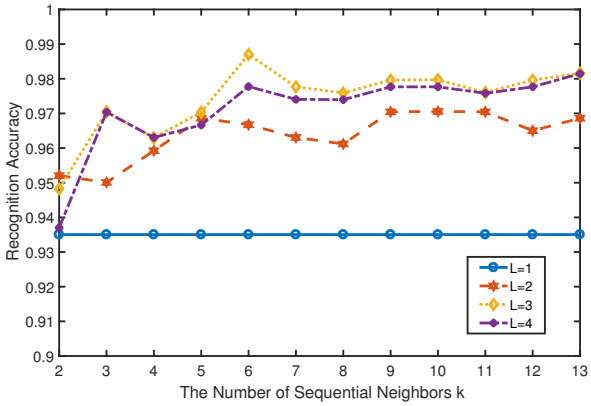

Fig. 9. The influence of the hierarchical level number L with the MTF number $T = 2$ on activity recognition accuracies using the KARD dataset under "new-person" setting.

best in all situations, this confirms our conjectures: 1) $L = 2$ is not able to capture enough temporal information, while 2) $L = 4$ may destroy the structural integrity of action snippets.

5.4.3 Analysis of the Sequential Neighbors Number

In both Fig. 8 and Fig. 9, the sequential neighbors number k is adjusted from 2 to 13 with step 1. For a given setting of parameters (such as $T = 2$, $L = 3$ in Fig. 8 or Fig. 9), the performance presents a noticeable upward tendency with the increase of sequential neighbors k, and when $k \geq 6$, the performance tends to be stable. It is expected since the proposed hierarchical temporal dividing algorithm is based on the computation of nonlinearity degree, which is heavily dependent on the k sequential neighbors graph. Larger neighbors number k holds the intrinsic structure of human action sequence to a higher degree, which leads to more reasonable graph representation. However, larger is not always better: if k is too large, it may break the basic assumption that continuous postures tend to have closer features. On the other hand, larger k means greater computational complexity. As we have seen, $k = 6$ is a good compromise in the proposed approach.

6 Conclusion

In this paper, a novel descriptor PTD is constructed from 3D joint locations sequence for human action recognition. To encode the temporal information, we design a hierarchical temporal dividing algorithm to divide a sequence into two compact sub-sequences (action snippets) based on the manifold assumption on action sequence, and multiple PTDs are deployed over an entire action sequence and its sub-sequences. The final representation of a human action is a hierarchical temporal description of the initial sequence. Our experimental studies show the advantage of the proposed approach. The success of the proposed approach results from two aspects. Albeit simple, the major posture feature and main dynamical tendency feature can capture most discriminative features of the human posture represented by joint locations. On the other hand, the natural and intuitive dividing algorithm can encode the temporal dynamics and preserve the geometric structures of action snippets within an activity.

There are some directions to improve the proposed approach further in future work. First, for an action sequence with highly nonlinear degree, the major posture would result in meaningless posture, in which case, the major posture can be replaced with representative posture by extracting the key frame in the action snippet. Besides, in addition to focus on the global dynamical tendency by computing the covariance of the whole body joint locations, considering covariance of body parts that relevant to action evolution would obtain discriminative local dynamics. Finally, the manifold dividing algorithm can be performed based on the entire distance ratio matrix, which holds distance ratio between pairwise postures in the whole action sequence. However, dividing sequence with global information could increase the computation complexity to some extent. On the other hand, application of the proposed approach in video surveillance system for abnormal activities detection is an interesting direction, which also comprise our future work.

Acknowledgment

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2016YFB1000905), the National Natural Science Foundation of China (Grant Nos. 91546116, 61673363).

References

[1] J. Aggarwal and M. Ryoo., "Human activity analysis: A review." *ACM Computing Surveys (CSUR)*, pp. 16:(1–43), 2011.
[2] J. Aggarwal and L. Xia., "Human activity recognition from 3d data: A review." *Pattern Recognition Letters*, pp. 70–80, 2014.
[3] M. Vrigkas, C. Nikou, and I. Kakadiaris., "A review of human activity recognition methods." *Frontiers in Robotics and AI*, 2015.
[4] J. Shotton, T. Sharp, A. Kipman, and et al., "Real-time human pose recognition in parts from single depth images." *Communications of the ACM*, pp. 116–124, 2013.
[5] O. Oreifej and Z. Liu., "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences." *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 716–723, 2013.
[6] L. Xia and J. Aggarwal., "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera." *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2834–2841, 2013.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2803838, IEEE Transactions on Big Data

[7] H. Rahmani, A. Mahmood, D. Huynh, and et al., "Histogram of oriented principal components for cross-view action recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 2430–2443, 2016.

[8] J. Wang, Z. Liu, Y. Wu, and J. Yuan., "Learning actionlet ensemble for 3d human action recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 11–40, 2014.

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

[10] Y. Du, W. Wang, and L. Wang., "Hierarchical recurrent neural network for skeleton based action recognition." *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1110–1118, 2015.

[11] A. Yao, J. Gall, F. Fanelli, and L. Gool., "Does human action recognition benefit from pose estimation?" *British Machine Vision Conference*, 2011.

[12] V. Zarsiorsky., "Kinetics of human motion." *Human kinects*, 2002.

[13] D. Gong, G. Medioni, and X. Zhao, "Structured time series analysis for human action segmentation and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1414–1427, 2014.

[14] F. Lv and R. Nevatia, "Recognition and segmentation of 3-d human action using hmm and multi-class adaboost," *Computer Vision–ECCV 2006*, pp. 359–372, 2006.

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

[16] G. W. Taylor, G. E. Hinton, and S. T. Roweis, "Modeling human motion using binary latent variables," in *Advances in neural information processing systems*, 2007, pp. 1345–1352.

[17] Q. Ma, L. Shen, E. Chen, S. Tian, J. Wang, and G. W. Cottrell, "Walking walking walking: Action recognition from action echoes," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 2457–2463.

[18] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, vol. 2. IEEE, 2006, pp. 2169–2178.

[19] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1290–1297.

[20] M. Hussein, M. Torki, M. Gowayyed, and M. El-Saban., "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations." *International Joint Conference on Artificial Intelligence*, pp. 2466–2472, 2013.

[21] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," in *Computer vision–ECCV 2012*. Springer, 2012, pp. 872–885.

[22] A. Elgammal and C.-S. Lee, "Inferring 3d body pose from silhouettes using activity manifold learning," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. IEEE, 2004, pp. II–II.

[23] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[24] S. Gaglio, G. Re, and M. Morana., "Human activity recognition process using 3-d posture data." *IEEE Transactions on Human-Machine Systems*, pp. 586–597, 2015.

[25] L. Xia, C. Chen, and J. Aggarwal., "View invariant human action recognition using histograms of 3d joints." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 22–27, 2012.

[26] L. Seidenari, V. Varano, S. Berretti, and et al., "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses." *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 479–485, 2013.

[27] G. Johansson, "Visual motion perception." *Scientific American*, 1975.

[28] D. Wu and L. Shao, "Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 724–731.

[29] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4041–4049.

[30] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie *et al.*, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks." in *AAAI*, vol. 2, 2016, p. 8.

[31] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 816–833.

[32] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data." in *AAAI*, 2017, pp. 4263–4270.

[33] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 7, 2017.

[34] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *International Workshop on Human Behavior Understanding*. Springer, 2011, pp. 29–39.

[35] J. Luo, W. Wang, and H. Qi, "Group sparsity and geometry constrained dictionary learning for action recognition from depth maps," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1809–1816.

[36] R. Vemulapalli and R. Chellapa, "Rolling rotations for recognizing human actions from 3d skeletal data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4471–4479.

[37] Z. Huang, C. Wan, T. Probst, and L. Van Gool, "Deep learning on lie groups for skeleton-based action recognition," vol. 7, 2017.

[38] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[39] L. Wang and D. Suter, "Learning and matching of dynamic shape manifolds for human action recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 6, pp. 1646–1661, 2007.

[40] D. Gong and G. Medioni, "Dynamic manifold warping for view invariant action recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 571–578.

[41] X. Zhang, Y. Yang, L. Jiao, and F. Dong, "Manifold-constrained coding and sparse representation for human action recognition," *Pattern Recognition*, vol. 46, no. 7, pp. 1819–1831, 2013.

[42] R. Slama, H. Wannous, M. Daoudi, and A. Srivastava, "Accurate 3d action recognition using learning on the grassmann manifold," *Pattern Recognition*, vol. 48, no. 2, pp. 556–567, 2015.

[43] I. Bülthoff, H. Bülthoff, and P. Sinha., "Top-down influences on stereoscopic depth-perception." *Nature neuroscience*, pp. 254–257, 1998.

[44] R. Wang, S. Shan, X. Chen, and et al., "Manifold-manifold distance and its application to face recognition with image sets." *IEEE Transactions on Image Processing*, pp. 4466–4479, 2012.

[45] M. Gowayyed, M. Torki, M. Hussein, and M. El-Saban., "Histogram of oriented displacements (hod): Describing trajectories of human joints for action recognition." *International Joint Conference on Artificial Intelligence*, pp. 1351–1357, 2013.

[46] E. Cippitelli, S. Gasparrini, E. Gambi, and et al., "A human activity recognition system using skeleton data from rgbd sensors." *Computional Intelligence and Neuroscience*, 2016.

[47] M. Devanne, H. Wannous, S. Berretti, and et al., "3-d human action recognition by shape analysis of motion trajectories on riemannian manifold." *IEEE Transactions on Cybernetics*, pp. 1340–1352, 2015.

[48] C. Wang, Y. Wang, and A. Yuille., "Mining 3d key-pose-motifs for action recognition." *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2639–2647, 2016.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2803838, IEEE Transactions on Big Data

Yaqiang Yao received the BSc degree in computer science and technology from University of Science and Technology of China (USTC), Hefei, China, in 2013. He is currently pursing the PhD degree at USTC–Birmingham Joint Research Institute in Intelligent Computation and Its Applications (UBRI), School of Computer Science and Technology, University of Science and Technology of China (USTC), Hefei, China. His research interests include sequential data modeling and sequential decision making.

Yan Liu received the BSc degree and Master degree from University of Science and Technology of China (USTC), Hefei, China, in 2014 and 2017, respectively. Currently, he is with the Hikvision Digital Technologies Co., Ltd as a engineer. His research interests include machine learning and its application on computer vision.

Zhenyu Liu is an Associate Professor in the School of School of Information Management for Law at the China University of Political Science and Law. He obtained his undergraduate degree from Beihang University in China and his PhD degree from the University of Birmingham in UK. Prof. Zhenyu's research interests include reinforcement learning, nature inspired computation, machine learning and the real-world applications of AI techniques in the fields of legal practice.

Huanhuan Chen received the BSc degree from the University of Science and Technology of China (USTC), Hefei, China, in 2004 and the PhD degree in computer science from the University of Birmingham, Birmingham, UK, in 2008. He is currently a Full Professor at UBRI, School of Computer Science and Technology, USTC. His research interests include neural networks, Bayesian inference and evolutionary computation. Dr. Chen received the 2015 International Neural Network Society Young Investigator Award, the 2012 IEEE Computational Intelligence Society Outstanding Ph.D. Dissertation Award, and the 2009 British Computer Society Distinguished Dissertations Award. He is an associate editor of the *IEEE Transactions on Neural Networks and Learning Systems*. He is a senior member of the IEEE.