Learning Complex Spatio-Temporal Configurations of Body Joints for Online Activity Recognition

Jin Qi, Zhangjing Wang, Xiancheng Lin, and Chunming Li

***Abstract*—Geometric dynamic configurations of body joints play an essential role in distinguishing different human activities. However, many existing human activity recognition approaches lack the capability of automatically learning these configurations from sequences of joints in four-dimensional space (spatio and temporal). In this paper, the authors propose an automatic joint configuration learning method, based on dictionary learning and sparse representation. The proposed method achieves the following features: 1) it automatically learns dynamic spatio-temporal geometric configurations of body joints, involved in activities, in a simple way; 2) it dispenses with the hand crafted feature designing process and provides a new method to organize joint coordinate data as fixed length column vectors, which are suitable for dictionary learning; 3) it replaces the conventional bag of words model with sparse coding method; words in learned dictionary capture subactivity features, and the frequencies of different words appearing in different activities characterize the categories of global activity; 4) it is robust to time misalignment and can classify any length of video sequence (online classification) in real time; 5) it is easy to combine this method with other forms of data for better performance, because of its data driven nature and flexible framework. The proposed method is tested with three state-of-the-art public human activity recognition datasets and the results are found to be better than those of CAD-60 dataset, and comparable to those of both MSR Action 3D and MSR Daily Activity datasets (source codes are publicly available at https://github.com/jinqijinqi/SparseCodingDictionaryLearningHumanActivityRecognition).**

***Index Terms*—Bag-of-words (BoW) model, body joints, dictionary learning, human activity recognition, sparse coding.**

I. Introduction and Motivation

SMART environment is a small world where different categories of smart systems work continuously to make inhabitants' lives more comfortable [1]. The extant environment of ubiquitous computing systems, supported by cost-effective depth image/body joints acquiring camera (for instance, Microsoft Kinect), provides the right opportunity to build a smart place, such as health-assistive home/environment, at low cost [2]. To continuously track individuals'/patients' functional health and to initiate timely medical intervention, the daily living activities, such as feeling painful, falling down, drinking, eating, washing face, brushing teeth, cooking, dressing, and taking medicine have to be monitored [3], [4]. Automatic recognition of human activity, thus, forms the basis for building a smart place. Some algorithms of body-sensor-based activity recognition have been proposed earlier in [2], [5], and [6]. Recently, body joint-based method has evoked great interest in activity-recognition research community because of the availability of low cost, joint-acquiring Kinect cameras. Fig. 1 shows body joint trajectories (x–y views for best observation) of some

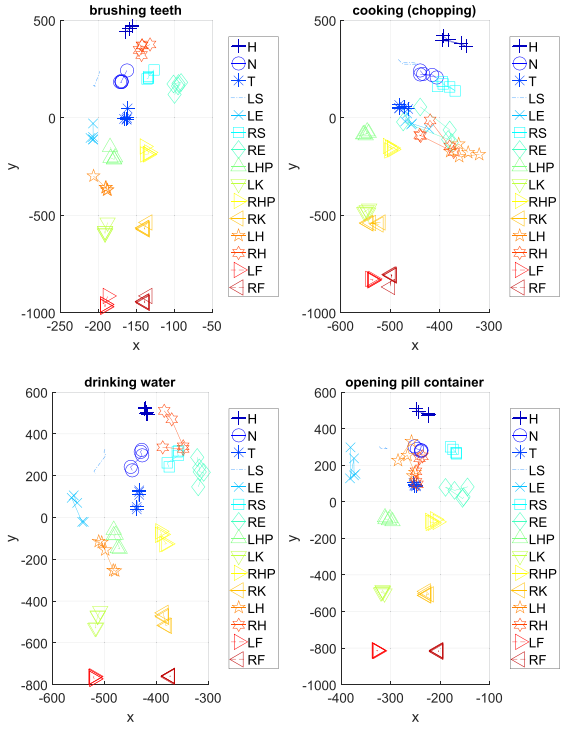


Fig. 1. Joint trajectories (to avoid cluttering, only joints from five randomly chosen frames are shown; different joints with different colors and markers, best viewed in the web version of this paper) of four example activities from CAD-60 dataset: brushing teeth, cooking-chopping, drinking water and opening pill container. H: head; N: neck; T: torso; LS: left shoulder; LE: left elbow; RS: right shoulder; RE: right elbow; LHP: left hip; LK: left knee; RHP: right hip; RK: right knee; LH: left hand; RH: right hand; LF: left foot; RF: right foot.

Manuscript received February 12, 2016; revised November 27, 2016, June 2, 2017, and February 9, 2018; accepted May 7, 2018. Date of publication July 24, 2018; date of current version November 13, 2018. This work was supported in part by the State Key Laboratory Open Project from Science and Technology on Electronic Information Control Laboratory (M162017510007000049), 29th Institute of China Electronics Technology Group Corporation, and in part by the Fundamental Research Fund for the Central Universities from the Center for Information in Biomedicine, University of Electronic Science and Technology of China (A03013023801137). This paper was recommended by Associate Editor G. Fortino. *(Corresponding author: Jin Qi.)*

J. Qi, Z. Wang, and C. Li are with the Department of Electrical Engineering, Center for Information in Biomedicine, Center for Digital Media and Culture, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: jqi@uestc.edu.cn; wangzhangjing@hotmail.com; li_chunming@hotmail.com).

X. Lin is with the Department of Geography, Sichuan Normal University, Chengdu 610068, China (e-mail: xianchenglin@163.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2018.2850301

activities, from public dataset CAD-60 [7], such as "brushing teeth," "cooking-chopping," "drinking water," and "opening pill container." From Fig. 1, it can be seen that different activities have different joint trajectory distributions, whose geometrical configurations carry important information that helps in identifying those activities. It has been shown that a multitude of human activities can be recognized by using only joint positions [8]. Therefore, body-joint-based human activity recognition method is emerging to be the most popular one.

Conventional human activity recognition methods normally use hand-designed local features and complex temporal dynamic models (TDM). For example, [9], [10] use interest point detectors (Harris3D, Cuboid, Hessian) to find key points and compute local feature descriptors [histogram of oriented gradient (HOG), histogram of optical flow (HOF), cuboid, and speeded up robust feature (SURF)] for each key point to describe the local pattern around each interest point. Then the histogram of frequencies of local features that appear in the activity is used as TDM. In [7], [11]–[13], humans or objects (such as mugs, bowls) are segmented out and local feature HOGs computed from the data around the centroids of segmented objects. The sequences of local features in activity [13], Markov model [7], [11], and latent support vector machine (LSVM) [12] are used as TDMs. Some other methods, such as [14], [15], use volume local binary pattern (VLBP) [14] and spacial time occupancy pattern (STOP) feature [15] as local features of instant posture in a video. Others use histogram of frequency of local features in the activity [14] and action graph [15] as TDM.

Although the recent methods [12], [16]–[30] take advantage of human body joint points, they still use engineered features and complex TDM. In most of these methods, relative Euclidean distance and direction between body joints are computed as local features of instant posture in a video. The TDMs in vogue include, besides others, Markov model [16], [19], [26], [30], graph model [22], dynamic Bayes mixture model [21], bag of words (BOW) model [17], Naive Bayes nearest neighbor [18], latent structure model[12], voting model [22], and temporal pyramid matching [29].

Most of the aforementioned models lack the ability to automatically learn four-dimensional (4-D) spatio-temporal features from pure body joint coordinate data. The purpose of this paper is to make up for this deficiency by providing an automatic feature-learning method for joint-based human activity recognition, using only joint data. Although the focus is only on joint data, the proposed method can be easily extended to other forms of data too.

In this study, an automatic 4-D spatio-temporal feature learning method is proposed by using $L0$ norm constrained dictionary learning and sparse coding. The proposed method achieves the following features: 1) it automatically learns, in a simple way, the dynamic spatio-temporal geometric configurations of body joints that are involved in activities; 2) it dispenses with the hand-crafted feature designing process and provides a new method to organize joint coordinate data as fixed length column vectors, which are suitable for dictionary learning; 3) it replaces the conventional BoW model with sparse coding method; words in learned dictionary capture subactivity features, and the frequencies of different words appearing in different activities characterize the categories of global activity; 4) it is robust to time misalignment and can classify any length of video sequence (online classification) in real time; and 5) it is easy to combine this method with other forms of data for better performance, by virtue of its data driven nature and flexible framework.

Elaborate experiments were carried out across three public databases. The experimental results show that the proposed algorithm outperforms or compares well with the state-of-the-art algorithms. The following are the chief contributions of this paper to human activity recognition research:

1) provides a new method to organize joint-coordinate data as fixed length column vectors, which are suitable for dictionary learning;
2) proposes a new sparse coding-based BoW model to learn word frequency histogram as feature vector;
3) proposes a simple dictionary learning-based method to automatically learn the complex 4-D spatio-temporal geometrical configurations of all body joints involved in activity, from joint coordinates alone, without going through the handcrafted feature designing step;
4) proposes a flexible framework that can combine $L0$ norm-constrained dictionary learning with the sparse coding-based BoW model for joint-based human activity recognition. Besides, it can be combined with extra red, green, blue and depth (RGBD) data for performance improvement;
5) provides source codes to reproduce all the results presented in this paper.

To the best of the authors' knowledge, this is the first attempt in using $L0$ norm-constrained dictionary learning method to jointly learn 4-D complex spatio-temporal geometrical configurations of joints, directly from the original 4-D joint coordinate volume in human activity recognition research community. It is also the first to propose a new sparse coding-based BoW model, and also a new method to organize joint coordinate data as fixed length column vectors, which are suitable for machine learning.

For more information on human activity recognition, the readers are referred to several survey papers on conventional 2-D image/video-based human activity recognition algorithms [31]–[33] and recent surveys on depth/3-D data-based human activity recognition [34], [35].

II. Related Work

In this paper, a new method is proposed to automatically learn 4-D spatio-temporal features from body joint coordinate data by using $L0$ norm constrained dictionary learning and sparse coding.

Several automatic feature-learning methods have been developed earlier for human activity recognition. Ni *et al.* [36] use dictionary learning-based approach to learn features from original depth image sequences. In [29], a 3-D spatial dictionary is learned for each joint from the difference in position (coordinate difference) between that joint and each of the other joints within the same frame. Different joints have different dictionaries and these dictionaries are learned separately. In [37], conventional independent component analysis (ICA) method with $L1$ norm constraint is used to learn the dictionary from joint coordinates.

These methods treat joints independently [29] or use conventional ICA [37] and $L1$ norm [29], [37] for dictionary learning and sparse coding. The performance of these methods can be further improved by treating the joints jointly and using $L0$ norm constrained sparse coding.

Sparse constrained dictionary learning method is generally considered superior to the traditional "BoW" model-based feature learning method [38]. In BoW method, the so called codebook is constructed by k-means or fuzzy k-means clustering method, and a local feature vector is quantized by one or several codewords (representative words) through k-nearest neighbor finding method. The traditional k-means clustering step and k-nearest neighbor finding step can be replaced by modern sparse constrained dictionary learning method and sparse coding method, which can learn more accurate dictionary and avoid quantization error. BoW method, combined with sparse constrained dictionary learning, is expected to improve the performance of BoW-based recognition method.

Inspired by the works mentioned above and to obtain better performance of human activity recognition, the authors propose to use L_0 norm constrained sparse coding to learn 4-D feature jointly and to replace conventional BoW method. The proposed method is designed, based on L_0 norm constrained dictionary learning and sparse coding. In the proposed method, a 4-D spatio-temporal dictionary is learned for each activity (rather than a 3-D spatial dictionary for each joint in [29]) from the original coordinates (rather than the position difference between joints in [29]) of all joints within the same subvolume consisting of several adjacent frames (rather than joints within one frame in [29]), using a recent L_0 norm-based dictionary learning method (rather than using conventional ICA in [37]). Then , sparse coefficients from sparse coding are used, instead of conventional BoW model, to build word-frequency histograms as feature vectors. Finally, SVM is used to perform the classification task.

In the method, proposed here, the 4-D atoms in the learned dictionary naturally capture information about spatio-temporal dynamics of subactivities. Therefore, atomic activities can be efficiently represented by the proposed dictionary atoms. Information on temporal dynamics can be easily captured by tuning the length of temporal dimension of each subvolume sample. Sparse coding inference is performed in lower dimensional space by using principal component analysis (PCA) for dimension reduction. The proposed algorithm achieves real time performance. It performs more experiments in this paper than in [37]. The experimental results show that the method proposed in this paper achieves better results than those in [37] with CAD-60 dataset.

The remainder of this paper is organized as follows: the method proposed here is described in detail in Section III; the experimental results are presented in Section IV, and the same are discussed in Section V; and the conclusions drawn from this study are presented in Section VI.

III. Proposed Feature Learning Method

The block diagram of the proposed method is shown in Fig. 2. For better view, the combination of every three consecutive frames, fewer than the number of frames in the algorithm, is considered a subvolume. The joint coordinates of each subvolume are reorganized into a column vector in joint category order and denoted by a thin vertical rectangular bar, where "x," "y," "z" indicate the x, y, and z components of each joint. Then, these column vectors are used to learn a dictionary. Finally, the histogram of word frequency is learned as input feature vector to SVM classifier.

A. Data Organization and Preprocessing

For this paper, the number of body joints in each frame is assumed to be N. Each joint is a point in 3-D space, consisting of x, y, and z coordinate components, which form the output of Kinect camera. The ith joint at tth frame is denoted by a coordinate component vector, as shown below

$$\boldsymbol{p}_i(t) = [x_i(t)\ y_i(t)\ z_i(t)],\ i = 1, 2, \ldots, N \tag{1}$$

where $x_i(t)$, $y_i(t)$, $z_i(t)$ denote, respectively, the x, y, and z coordinate components of joint point $p_i(t)$. $\boldsymbol{p}_i(t)$ is normalized by subtracting the mean of all vectors $\boldsymbol{p}_i(t)$, $i = 1, \ldots, N$, in the tth frame, to make it independent of camera coordinate system

$$\hat{\boldsymbol{p}}_i(t) = \boldsymbol{p}_i(t) - \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{p}_i(t). \tag{2}$$

Each normalized joint vector $\hat{\boldsymbol{p}}_i(t)$, $i = 1, \ldots, N$, in the tth frame, is arranged as a row vector of the 2-D coordinate matrix $P(t)$ at instantaneous moment t, i.e.,

$$P(t) = \begin{bmatrix} \hat{\boldsymbol{p}}_1(t) \\ \hat{\boldsymbol{p}}_2(t) \\ \cdots \\ \hat{\boldsymbol{p}}_N(t) \end{bmatrix} = \begin{bmatrix} \hat{x}_1(t) & \hat{y}_1(t) & \hat{z}_1(t) \\ \hat{x}_2(t) & \hat{y}_2(t) & \hat{z}_2(t) \\ \cdots & & \\ \hat{x}_N(t) & \hat{y}_N(t) & \hat{z}_N(t) \end{bmatrix}. \tag{3}$$

A final 3-D coordinate matrix $\boldsymbol{P}$ is built by concatenating all the 2-D coordinate matrices $P(t)$ in the third dimension (t dimension), in a chronological order.

The 3-D matrix P is densely sampled along time dimension (the third dimension) to obtain subvolumes as samples. For this study, the subvolume sample size t_s, along time axis (t axis), is set to 11, 13, and 23 for CAD-60 database, MSR Action3D dataset, and MSR Daily Activity dataset, respectively. Each subvolume sample is collapsed into a column vector, called sample, by concatenating all the column vectors in the subvolume sample. The length of each sample vector is $15 \times 3 \times 11$ for CAD dataset with 15 joints per body, $20 \times 3 \times 13$ for MSR Action3D dataset with 20 joints per body, and $20 \times 3 \times 23$ for MSR Daily Activity dataset with 20 joints per body.

Each sample vector is normalized by subtracting its mean vector. Each normalized sample vector is treated as a column vector of matrix $\boldsymbol{X}$. The normalized sample vectors in matrix $\boldsymbol{X}$ are whitened and their dimensions reduced by using PCA [39] as follows:

$$XX^T = WDW^T,\ \hat{X} = W_k D_k^{-\frac{1}{2}} W_k^T X \tag{4}$$

where D, W, D_k, and W_k denote, respectively, full diagonal eigenvalue matrix, full eigenvector matrix, k largest eigenvalue

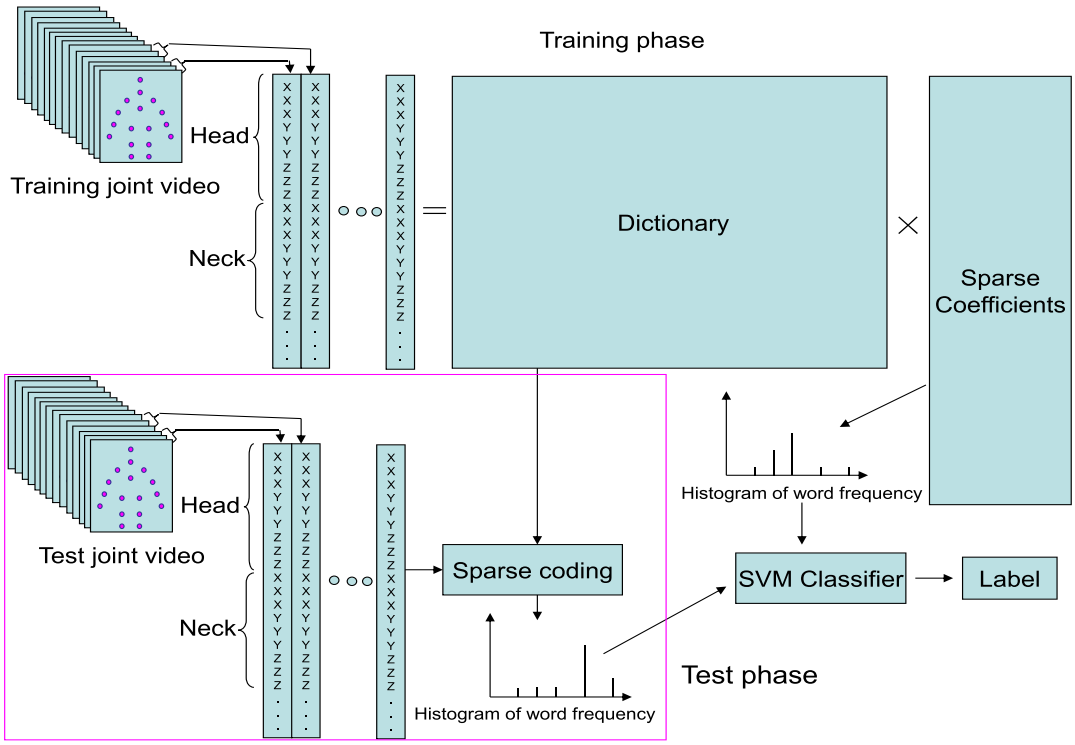


Fig. 2. Block diagram of the proposed system: Each combination of three consecutive frames (fewer than the frames in the proposed algorithm, for better view) is considered a subvolume; the joint coordinates of each sub-volume are reorganized into a column vector, in joint category order, denoted by a thin vertical rectangular bar whose " x," " y," "z" indicate the x, y, and z components of each joint; these column vectors are then used to learn a dictionary; Finally the histogram of word frequency is learned as the input feature vector to SVM classifier.

matrix and its corresponding eigenvalue matrix. k is the smallest number of eigenvalues, whose sum is equal to or more than some proportion (99% in this paper) of the sum of all eigenvalues.

B. Spatio-Temporal Dictionary Learning

With the whitened and dimension reduced sample vectors as columns of matrix $\hat{X}$, the l_0 norm-based dictionary learning method, recently presented in [40], is used to learn 4-D dictionary by solving the following optimization problem:

$$\min_{D,C} \frac{1}{2}||\hat{X} - DC||_2^2 + \lambda||C||_0$$
$$\text{s.t. } ||d_i||_2 = 1, 1 \le i \le m \tag{5}$$

where d_i is the ith column (word) in dictionary D, $||C||_0$ is the number of nonzero entries in sparse coefficient matrix C, and m is the total number of words in dictionary D ($m = 400$ in this paper). The balancing parameter λ in (5) is set to 6500, as recommended by [40], and validated by cross validation of this study. A fast proximal method is proposed in [40] to solve this optimization problem and obtain the dictionary D.

It is well known that different kinds of human activities have different features (for instance, different joint configurations and their evolution along time axis). To capture the class-specific features of each kind of activity of this study, a class-specific dictionary is learned for each activity category, using samples of the same class. Thus, the words in the class-specific dictionary are very good for representing the samples of the same class, but not for the samples of other class. To obtain a dictionary, which can work for any sample from any activity class, a final large dictionary $\bar{D}$ is built by simply combining all the words from the learned class-specific dictionaries.

C. Sparse Coding Based BoW Model for Feature Vector Generation

Once the final large dictionary $\bar{D}$ is built, the sparse coding vector s of a sample vector x can be obtained by solving the following sparse coding model:

$$\min_{s} ||\bar{D}s - x||_2^2 + \lambda||s||_1 \tag{6}$$

where l_1 norm $||s||_1$ is the sum of the absolute values of entries in vector s. The balancing parameter λ in (6) is the same as the one in (5), and the same is validated by cross validation in this study. A large number of algorithms were proposed to solve the above sparse representation problem [41]. In this study, sparse coefficient vector s is obtained by using "orthogonal matching pursuit" (OMP) method [42].

The ith nonzero entry in sparse vector s of sample vector x indicates that the ith word appears in sample vector x. Therefore, by checking all the sparse coding vectors from an activity video, the words that appear in the video, as also how many times they appear, are known. Based on the "BoW" model [38], and the words appearing in the activity video, a histogram of word frequency is built as a feature vector of the video.

D. SVM-Based Classification

Each feature vector is expanded by using the explicit γ homogeneous kernel expansion [43] with "χ squared" kernel ($\gamma = 0.01$ in this paper). In [43], the explicit expression of kernel mapping function is as follows:

$$\frac{\psi[x]_j}{\sqrt{xL}} = \begin{cases} \sqrt{k\,(0)}, & j = 0 \\ \sqrt{2k\left(\frac{j+1}{2}L\right)}\cos\left(\frac{j+1}{2}L\log x\right), & j = odd \\ \sqrt{2k\left(\frac{j}{2}L\right)}\sin\left(\frac{j}{2}L\log x\right), & j = even \end{cases} \quad (7)$$

where $\psi[x] \in R^{2n+1}$ and $j = 0, 1, \ldots, 2n$. This function can map a positive number x to a 2n + 1 dimension vector. For more information on homogeneous kernel, the readers are referred to [43]. The expanded histogram feature vectors are used in training class-specific linear support vector machines [44], following the one-versus-rest strategy. For each class, the model parameter w of one linear SVM classifier $f(x) = w^T x + b$ is obtained (with balancing parameter c set to 0.01) by solving the following quadratic optimization problem:

$$\min_{w,b} L(w, b) = \frac{1}{2}||w||_2^2 + c\sum_i \max\left(0, 1 - y_i(w^T x_i + b)\right) \quad (8)$$

where x_i, y_i denote the training sample vector and its label, respectively. After training, the test video can be classified by using the trained SVM classifier. The SVM classifiers are trained by using the source codes publicly available from http://www.vlfeat.org/.

IV. Experimental Results

A. Datasets for Performance Evaluation

The algorithm proposed here is tested exhaustively, with the available body joints, using three public datasets: Cornell Activity Dataset -60 (CAD-60) [7], MSR Action3D [45], and MSR Daily Activity 3D [28]. The quality of joint data varies between the datasets.

CAD-60 set is acquired by Microsoft Kinect camera with frame rate 30 Hz and 640 × 480 image resolution. The skeleton tracking algorithm [46] within the camera sensor outputs body joint coordinates in 3-D space. In the database, each person has 15 joints. Four subjects (two males and two females) perform 14 activities in indoor environment. Each activity video is, on an average, approximately 1000 frames-long. Of the four subjects, three are right-handed and one left handed. The quality of this dataset is good.

In MSR Action3D dataset, ten subjects individually perform 20 activities, repeating each activity two or three times. The activities are covered by 567 depth videos, with image resolution of 320 × 240 and frame rate of 16 Hz. The 3-D coordinates of 20 classes of joints are available. Of the 567 videos, only 557 videos are used, just as in [28], because of missing or erroneous joints in the remaining ten videos. The quality of this dataset is poor.

MSR Daily Activity 3D dataset, acquired by Microsoft Kinect camera, has 960 RGBD videos, with 320 videos for each channel. For this dataset, each of the ten subjects performs each of the 16 activities twice. The 3-D coordinates of 20 classes of body joints for each person are available in this dataset. The joint positions are very noisy as the activities are performed in two different poses: "sitting on sofa" and "standing close to sofa". For the proposed method, only the joint information is used. The quality of this dataset is very poor, because of which it poses a big challenge to recognition algorithms.

TABLE I
Parameter Value Setting in the Proposed Method

Dataset	t_s	N_w	λ	γ	c
CAD-60	11	400	6500	0.01	0.01
MSRAction3D	13	400	6500	0.01	0.01
MSRDailyActivity3D	23	400	6500	0.01	0.01

t_s: sample size along time dimension; N_w: number of words in dictionary; λ: balancing parameter in (5) and (6); γ: kernel map parameter; c: balancing parameter in (8).

TABLE II
Evaluation of Time Performance of the Proposed Method, Using Dataset CAD-60, MSR Action3D, and MSR Daily Activity 3-D

Dataset	**Average Testing Time per video (seconds)**
CAD-60	0.9443
MSR Action 3D	0.1679
MSR Daily Activity 3D	0.1864

B. Parameter Setting and Time Performance

Five parameters are used in the proposed method: 1) sample size along time dimension t_s; 2) number of words in dictionary N_w; 3) balancing parameter λ in (5) and (6); 4) γ parameter with kernel map; and 5) balancing parameter c in SVM training. Cross validation is used to find the optimal parameter values, as shown in Table I. From this Table, it can be seen that only the parameter t_s (sample size along time dimension) changes (depending on action speed and frame rate in video) across different datasets.

The algorithm of the proposed method is implemented in MATLAB language, with a single thread, using a normal personal computer with Intel(R) Xeon(R) CPU E5-1603@2.80 GHz, under 64 bit Ubuntu 14.04.1 LTS. The average testing time for each video, across the three datasets, is shown in Table II.

From this Table, it can be seen that the maximum testing time per video in these three datasets is less than 1 s (0.9443 s).

C. Results From CAD-60 Dataset

In this dataset, the third subject is left-handed, while the other three are right-handed. No good recognition performance can be expected from a system, when it is trained with three right-handed subjects and tested with a left-handed subject (the so called "new person" setting in this paper). The joints from the left-handed subject are mirrored to make her behave like a right-handed subject, as was done in [7], [16]. Specifically, for each frame of the left-handed subject, one plane l is computed by fitting it to the four joint points: left-arm, right-arm, left hip, right hip. Then, mirror plane l_m is the plane, which is perpendicular

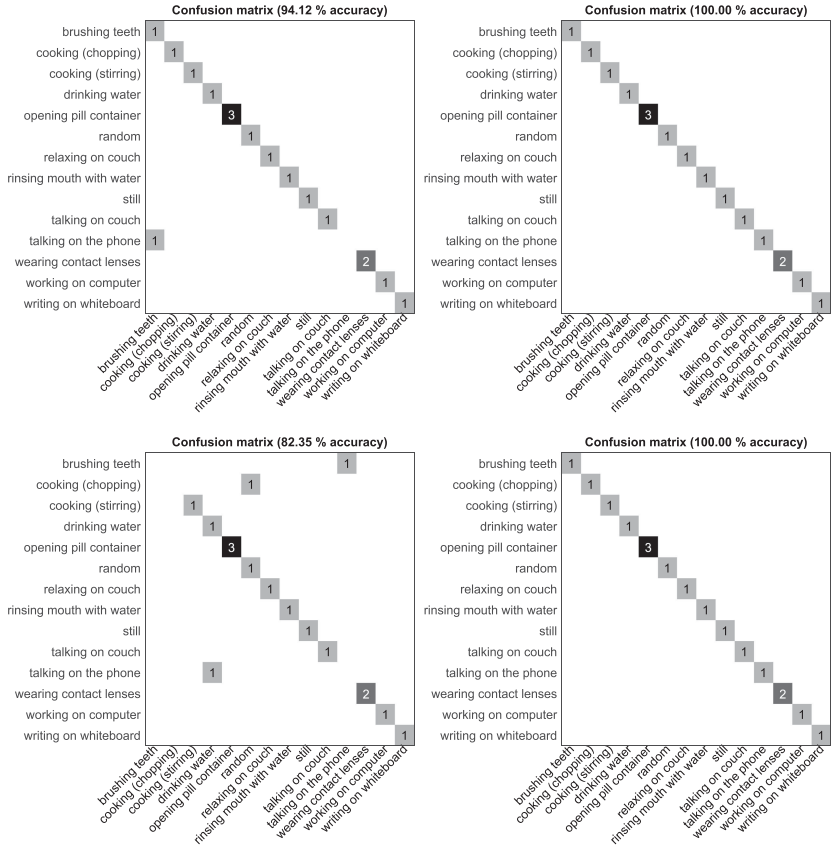


Fig. 3. Four confusion matrices (with accuracies of 94.12%, 100.00%, 82.35.12%, 100.00%) for four experimental settings on CAD60 dataset: each setting corresponds to the case in which one subject is for testing and the other three are for training.

to the computed plane P and passes through two midpoints, one between two arm joints and the other between two hip joints. Virtual joints, used in the proposed algorithm, are obtained by mirroring joints within this frame , with respect to the computed mirroring plane l_m.

Also, of the four subjects, three are chosen for training and one for testing, by strictly following the "new person" experimental setting in [7]. Therefore, four experimental settings are available, and anyone of the four subjects can be used for testing. The four confusion matrices, corresponding to the four experimental settings, are shown in Fig. 3. The proposed algorithm achieves accuracies of 94.12%, 100.00%, 82.35%, and 100.00%, respectively, with the first, second, third, and fourth subjects, as testing subjects. From the top left subfigure in Fig. 3, it can be seen that the "talking on the phone" action is wrongly classified as "brushing teeth" action, when the first subject is used as the testing subject. These two actions are quite similar to each other, because both of them have similar geometrical configurations of body joints at many moments. Therefore, information from body joints alone is not enough to differentiate one action from the other. The relatively low accuracy (82.35%) obtained with the third subject as testing subject could possibly be due to the error introduced by the extra mirror operation, which converts left-handed subject to right-handed subject.

The proposed method is compared with other state-of-the-art algorithms in terms of average accuracy, average precision, and average recall (see Tables III and IV). The results relating to the state-of-the-art algorithms are taken directly from the website http://pr.cs.cornell.edu/humanactivities/results.php.

TABLE III
AVERAGE ACCURACY VALUES (%) WITH CAD-60 DATASET

Algorithm	Feature	Average Accuracy (%)
Markov Model [16], [7].	Joint+RGBD	51.9
Markov Random Field [11]	Joint+RGBD	71.4
Order Preserve Sparse Coding [36]	RGBD	65.32
Actionlet [20]	Joint+RGBD	74.70
Interest Point [9]	RGBD	87.5
Pose Kinetic Energy [22]	Joint	91.9
Proposed algorithm	Joint	**94.12**

Proposed algorithm, with accuracy of 94.12%, ranks first among the seven algorithms.

From Table III, it can be seen that the proposed algorithm , which uses only joint information, outperforms some other joint [22] or RGBD [9] or RGBD + Joint [20]-based methods. The proposed method, with accuracy of 94.12%, ranks first among the seven algorithms. The second best algorithm, called "pose kinetic energy" [22] in Table III, uses only joint information

TABLE IV
AVERAGE RECALL AND PRECISION VALUES (%) ON CAD-60 DATASET

Algorithm	Feature	Average Precision(%)	Average Recall(%)
Markov Model [16], [7]	RGBD +Joint	67.9	55.5
Markov Random Field [11]	RGBD +Joint	80.8	71.4
Simple Joint Distance and Angle [17]	Joint	86	84
EigenJoints [18]	Joint	71.9	66.6
Hidden Markov Model [19]	Joint	70	78
Depth and Image Fusion [12]	RGBD +Joint	75.9	69.5
Pose Codewords [13]	RGBD	78.1	75.4
Posture Data [30]	Joint	77.3	76.7
Interest Point [9]	RGBD	**93.2**	84.6
Probabilistic Approach [21]	Joint	**91.1**	91.9
Pose Kinetic Energy [22]	Joint	**93.8**	**94.5**
Proposed algorithm	Joint	**90.18**	**92.86**

Proposed algorithm, with 92.86% recall and 90.18% precision, ranks, respectively, second and fourth among the 12 algorithms, in terms of recall and precision.

TABLE V
ACTION SETTINGS FOR SUBSETS AS1, AS2, AND AS3

AS1	AS2	AS3
Horizontal arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw x	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend two	hand wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pickup & throw	Side boxing	Pickup & throw

and gives lower accuracy of 91.9%. Table III also shows that the performance of the proposed method, which is a joint-based method, is better than that of even the nonjoint-based methods.

The performance of some algorithms is reported only in terms of their precision and recall values. The performance of the proposed algorithm is compared with such algorithms also in terms of the same metrics, as shown in Table IV. From this Table, it can be seen that the proposed algorithm, using only joint information, outperforms some other joint [17]–[19], [30] or RGBD [13] or RGBD + Joint [7], [11], [12], [16]-based methods. Further, the proposed method, with 92.86% recall and 90.18% precision, ranks second and fourth among the 12 algorithms, in terms of recall and precision, respectively. However, the performance gap between the proposed method and the other top ranking methods is very small.

TABLE VI
ACCURACY VALUES (%) OF THE STATE-OF-THE-ART METHODS ON MSR ACTION 3D DATASET, UNDER "OVERALL" SETTING

Method	Data	Accuracy(%)
Recurrent neural network [23]	Joint	42.5
Dynamic temporal warping [24]	Joint	54.0
Hidden Markov model [25]	Joint	63.0
Action graph on bag of 3D points [45]	RGBD	74.7
Histogram of 3D joints [26]	Joint	78.97
STOP feature [15]	RGBD	84.8
Eigenjoints [26]	Joint	82.3
Random occupy pattern [45]	RGBD	86.2
Actionlet ensemble [28]	Joint	**88.2**
sparse coding and pyramid matching [29]	Joint	**93.83**
Ours	Joint	**86.81**

Proposed algorithm, with accuracy of 86.81%, ranks third among the 11 algorithms.

D. Results From MSR Action3D Dataset

As in [29] and [45], this dataset also is divided into three subsets (see Table V): AS1, AS2, and AS3. AS1 and AS2 subsets include data of similar actions, and AS3 data of complex actions, which are combinations of simple activities. The proposed algorithm is also tested with AS1, AS2, AS3 subsets and whole dataset (hereafter called "overall"), using the same cross-subject experimental setting: subjects with numbers 1, 3, 5, 7, 9 worked for training and those with numbers 2, 4, 6, 8, 10 for testing.

The four confusion matrices of the proposed algorithm, corresponding to the four experimental settings: AS1, AS2, AS3, and overall, are shown in Fig. 4. The proposed algorithm achieves accuracy of 81.90%, 83.04%, 97.30%, and 86.81% in AS1, AS2, AS3, and overall settings, respectively. The third confusion matrix shows that the proposed algorithm performs very well, with accuracy of 97.30%, in recognizing complex activities of subset AS3. Only three "high throw" actions are incorrectly classified as "tennis serve", because both these actions are quite similar to each other, in terms of joint configuration. The proposed algorithm achieves lower accuracies (81.90%, 83.04%) in subset AS1 and AS2 than in subset AS3, because some actions in AS1 and AS2 are quite similar to each other. For example, the first confusion matrix (with accuracy 81.90%) from AS1 shows that "tennis serve" action is quite similar to "hammer," "forward punch," and "high throw" actions. The average performance of the proposed algorithm with CAD-60 dataset is better than that with MSR Action3D dataset. This is because the data in MSR Action3D dataset is more noisy.

Table VI compares the proposed algorithm, in terms of accuracy, with other state-of-the-art methods in overall setting. The accuracies of the state-of-the-art methods in Table VI are taken directly from [29]. From Table VI, it can be seen that the proposed algorithm, with accuracy of 86.81%, ranks third in overall setting. However, it is much simpler and more flexible than the two more accurate methods in the top [28], [29], both of which are based on pyramid strategy.

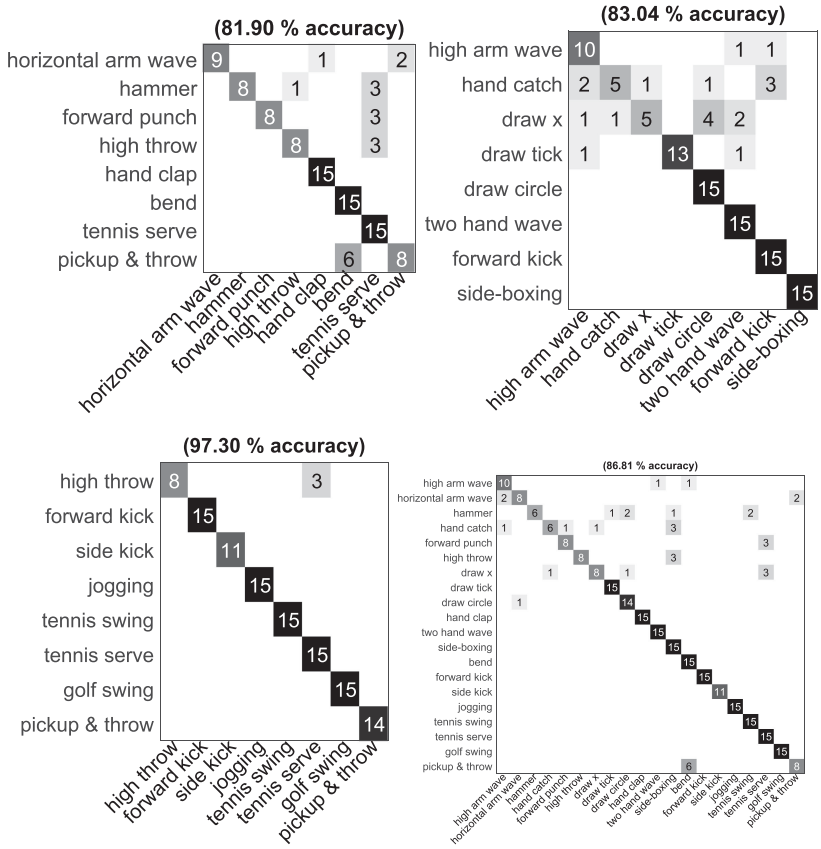


Fig. 4. Four confusion matrices from four subsets AS1, AS2, AS3, and "overall," using MSR Action 3D database, with accuracies of 81.90%, 83.04%, 97.3%, and 86.81%, respectively.

TABLE VII
ACTION SETTINGS FOR DATA SUBSETS AS1, AS2, AND AS3

AS1	AS2	AS3
eat	drink	use laptop
read book	call on cellphone	cheer up
write on a paper	use vacuum cleaner	play guitar
use laptop	sit still	stand up
toss paper	play game	sit down
walk	lie down on sofa	

E. Results From MSR Daily Activity 3D Dataset

This dataset is also divided into three subsets: AS1, AS2, and AS3 (see Table VII). The proposed algorithm is tested with these datasets and the overall dataset, using the same cross-subject experimental setting as the one used in [45] i.e., using the subjects with numbers 1, 3, 5, 7, 9 for training and those with numbers 2, 4, 6, 8, 10 for testing.

The four confusion matrices of the proposed algorithm, corresponding to the four experimental settings-AS1, AS2, AS3, and the overall one are shown in Fig. 5. The proposed algorithm achieves accuracies of 68.33%, 81.67%, 86.00%, and 68.75% in AS1, AS2, AS3, and the overall settings, respectively. The first confusion matrix shows rather low accuracy of 68.33%, because many pairs of activities in subset AS1 are similar to each other, such as "eat" versus "read book," "eat" versus "toss paper," and "read book" versus "use laptop." Subsets AS2 and AS3 show relatively higher performance, because such similar activity pairs are fewer in them. From the first and fourth confusion matrices (see Fig. 5), it can be seen that the accuracy in "overall" setting (68.75%) is a little bit higher (less than 0.5%) than that in AS1 (68.33%). Compared to AS1, the "overall" setting does not show much higher performance, because it has more activities that need to be distinguished, although many easier cases (relatively easy to distinguish) are added to it from AS2 and AS3.

Once again, it is seen that the performance of the proposed algorithm drops when used with MSR Daily Activity 3D dataset, because the noise in this dataset is much higher than that in the other two datasets.

Table VIII provides a comparison of the proposed algorithm with other state-of-the-art algorithms, in terms of accuracy, under the overall setting, used in MSR Daily Activity dataset. All the accuracies in Table VIII, except that of the algorithm proposed here, are taken directly from [29]. From this Table, it can be seen that the proposed algorithm, with accuracy of 68.75%, ranks third among the eight algorithms. However, the two more accurate methods in the top [28], [29] use more information, including RGBD and body joints. From the fifth row of Table VIII, it can be seen that the second best algorithm in [28] (with accuracy 68%) is a little worse than the proposed algorithm (with accuracy 68.75%), when only joint information are used. Furthermore, the proposed method is much simpler than the two methods in the top, which are pyramid strategy-based methods [28], [29].

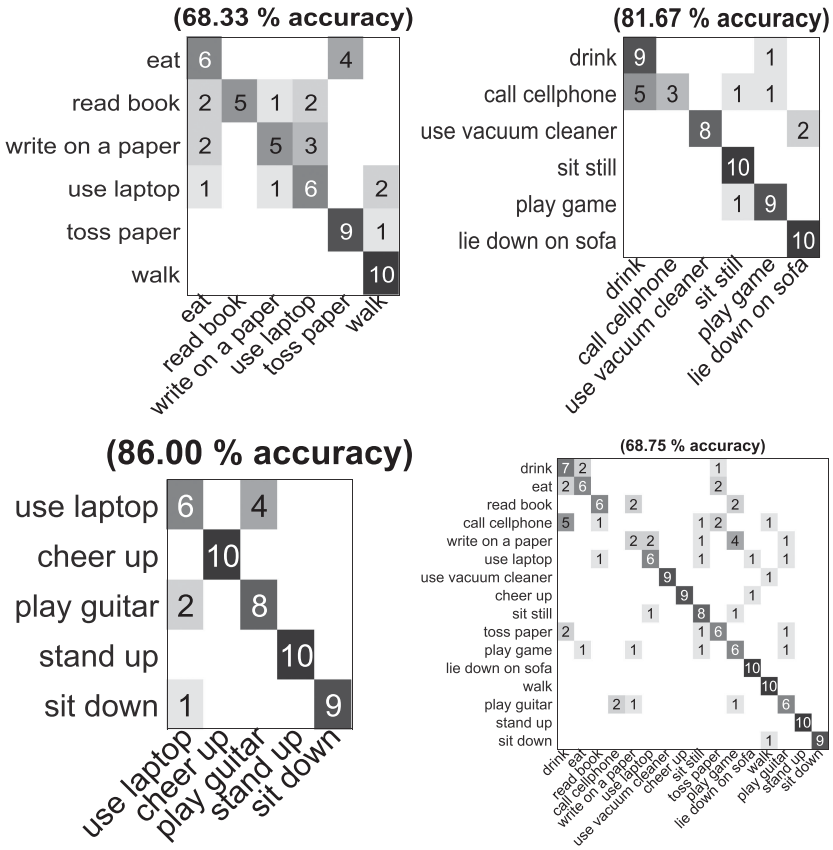


Fig. 5. Four confusion matrices from the four subsets AS1, AS2, AS3, and "overall", using MSR Daily Activity 3D dataset, with accuracies of 68.33%, 81.67%, 86.00%, and 68.75%, respectively.

TABLE VIII
COMPARISON BETWEEN THE PROPOSED METHOD AND THE STATE-OF-THE-ART METHODS, IN TERMS OF ACCURACY, USING MSR DAILY ACTIVITY 3-D DATASET UNDER "OVERALL" SETTING

Method	Data	Accuracy(%)
Cuboid+HoG (RGB)	RGBD	53.13
Dynamic Temporal Warping [24]	Joint	54
STIP+HOG/HOF [10]	RGBD	56.25
VLBP [14]	RGBD	59.38
Actionlet Ensemble [28]	Joint	68
Actionlet Ensemble [28]	RGBD +Joint	**85.75**
Sparse Coding and Pyramid Matching [29]	RGBD +Joint	**92.5**
Ours	Joint	**68.75**

Proposed algorithm ranks third among the eight algorithms, with accuracy of 68.75%.

The performance of the proposed method with MSR Daily Activity dataset (68.75% accuracy) is worse than that with MSR Action3D dataset (86.81% accuracy), possibly because some activities, such as "sitting on sofa" or "standing close to sofa," produce highly erroneous and noisy joint positions with MSR Daily Activity 3D dataset [28]. Furthermore, the average number of frames per video in that dataset is also small (60 frames per video), which may not provide enough information that can enable distinguishing between similar activities.

It can be seen that the performance of the proposed algorithm drops progressively, across the three datasets, in the following order: CAD-60 (accuracy 94.12%), MSR Action 3D (accuracy 86.81%), and MSR Daily Activity 3D (accuracy 68.75%). This is because these datasets have different properties, such as the length of the video, the number of activity classes and the quality of body joints. The quality of body joints in MSR Daily Activity 3D is particularly very poor or even erroneous [28].

F. Performance Comparison Across the Three Datasets

To the best of the authors' knowledge, there is no work, other than the one in [28], that reports on recognition performance across all the three datasets: CAD-60, MSR Action 3D, and MSR Daily Activity 3D. The recognition performance of the "actionlet ensemble" method [28] and that of the method proposed here, across these three datasets, are presented in Table IX. From this Table, it can be seen that, when these algorithms are used with CAD-60 dataset, the proposed algorithm, giving accuracy of 94.12%, performs much better (close to 20%) than the "actionlet ensemble" method [28], which gives accuracy of 74.70%. With MSR Action 3D dataset, "actionlet ensemble" method performs a little better (less than 1.4%) giving accuracy of 88.20%, compared to 86.81% accuracy given by the proposed method. On the other hand, with MSR Daily Activity 3D

TABLE IX
PERFORMANCE COMPARISON OF THE PROPOSED METHOD, IN TERMS OF ACCURACY (%), ACROSS THE THREE DATABASES, USING ACTIONLET ENSEMBLE METHOD [28]: "NEW PERSON" SETTING IN CAD-60; "OVERALL" SETTING IN BOTH MSR ACTION 3D AND MSR DAILY ACTIVITY 3D DATASETS; "JOINT" MEANS THAT ONLY JOINT INFORMATION WAS USED; "JOINT+DEPTH" MEANS THAT BOTH JOINT AND DEPTH INFORMATION WAS USED

Method\Data	CAD-60	MSR Action 3D	MSR Daily Activity 3D
Actionlet Ensemble [28]	74.70 (Joint+Depth)	88.20 (Joint)	68.00 (Joint)
Ours	94.12 (Joint)	86.81 (Joint)	68.75 (Joint)

dataset, the proposed algorithm gives slightly higher (less than 0.8%) accuracy (68.75%) than the "actionlet ensemble" method (68%). To sum up, both "actionlet ensemble" method [28] and the proposed algorithm give comparable accuracies with MSR Action 3D and MSR Daily Activity 3D datasets. Once again, it is observed that the performance of both of these methods deteriorates when they are used with MSR Daily Activity 3D dataset, because of the poor quality of body joints in that dataset.

V. DISCUSSION

The experimental results presented in Section IV show that the proposed method is simple and can perform very well across three public datasets; it even outperforms some RGBD + Joint based methods. The results also show that the quality of data has significant influence on the performance of the proposed algorithm. The performance of the algorithm is best with CAD-60 dataset, but poor with MSR Daily Activity 3D dataset, because of highly noisy or erroneous joints in the latter dataset. The performance of the proposed algorithm can be further improved by simply extending it to other forms of data, such as RGBD images. Once it is extended, the proposed algorithm will be tested with other popular public datasets, such as $ACT4^2$ Dataset [47], and compared with the trajectory-based method [48].

The proposed sparse coding-based feature learning method, which is simple and uses only joint information, is superior or comparable to the state-of-the-art methods that use joint or nonjoint information (RGBD) or both. The algorithm proposed here does not use any RGBD information, although it can be used to improve the performance of the proposed method.

VI. CONCLUSION

In this paper, the authors propose a new sparse coding and dictionary learning-based human activity recognition method, using only joint information. In this method, complex features with spatio-temporal geometric configurations of body joints from atomic subactivities are automatically learned. The proposed method achieves real time performance with PCA dimension reduction because of sparse coding in lower dimensional space.

The performance of the proposed algorithm is evaluated elaborately, using three datasets. The experimental results show that the proposed algorithm can achieve good performance with those datasets and it even outperforms some RGBD-based methods. The performance of the proposed method can be further improved by including other forms of data into its framework. Also source codes are provided here for reproducible research work and for encouraging other researchers to further improve the proposed method.

ACKNOWLEDGMENT

The authors are grateful for the insightful comments and suggestions offered by anonymous editors and reviewers of this paper.

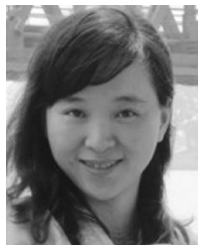
REFERENCES

[1] D. J. Cook and S. K. Das, *Smart Environments: Technologies, Protocols, and Applications*. New York, NY, USA: Wiley, 2005.
[2] Z. Wang, M. Jiang, Y. Hu, and H. Li, "An incremental learning method based on probabilistic neural networks and adjustable fuzzy clustering for human activity recognition by using wearable sensors," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, pp. 691–699, Jul. 2012.
[3] B. Das, D. J. Cook, M. Schmitter-Edgecombe, and A. M. Seelye, "Puck: An automated prompting system for smart environments: Toward achieving automated prompting–challenges involved," *Pers. Ubiquitous Comput.*, vol. 16, pp. 859–873, Oct. 2012.
[4] P. Kaushik, S. S. Intille, and K. Larson, "User-adaptive reminders for home-based medical tasks. A case study," *Methods Inf. Med.*, vol. 47, no. 3, pp. 203–207, 2008.
[5] G. Fortino, R. Giannantonio, R. Gravina, P. Kuryloski, and R. Jafari, "Enabling effective programming and flexible management of efficient body sensor network applications," *IEEE Trans. Human–Mach. Syst.*, vol. 43, no. 1, pp. 115–133, Jan. 2013.
[6] Z. Wang, D. Wu, R. Gravina, G. Fortino, Y. Jiang, and K. Tang, "Kernel fusion based extreme learning machine for cross-location activity recognition," *Inform. Fusion*, vol. 37, no. C, pp. 1–9, 2017.
[7] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from RGBD images," in *Proc. EEE Int. Conf. Robot. Autom.*, May 2012, pp. 842–849.
[8] G. Johansson, *Visual Motion Perception*. New York, NY, USA: Scientific American, 1975.
[9] Y. Zhu, W. Chen, and G. Guo, "Evaluating spatiotemporal interest point features for depth-based action recognition," *Image Vis. Comput.*, vol. 32, no. 8, pp. 453–464, 2014.
[10] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, no. 2/3, pp. 107–123, 2005.
[11] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *Int. J. Robot. Res.*, vol. 32, no. 8, pp. 951–970, 2012.
[12] B. Ni, Y. Pei, P. Moulin, and S. Yan, "Multilevel depth and image fusion for human activity detection," *IEEE Trans. Cybernet.*, vol. 43, pp. 1383–1394, Oct. 2013.
[13] R. Gupta, A. Y.-S. Chia, and D. Rajan, "Human activities recognition using depth images," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 283–292.
[14] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.
[15] A. Vieira, E. Nascimento, G. Oliveira, Z. Liu, and M. Campos, "Stop: Space-time occupancy patterns for 3D action recognition from depth map sequences," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (Lecture Notes in Computer Science), vol. 7441, L. Alvarez, M. Mejail, L. Gomez, and J. Jacobo, Eds. Berlin, Germany: Springer, 2012, pp. 252–259.
[16] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from RGBD images," in *Proc. AAAI Conf. Plan, Activity, Intent Recognit.*, 2011, pp. 47–55.
[17] C. Zhang and Y. Tian, "RGB-D camera-based daily living activity recognition," *J. Comput. Vis. Image Process.*, vol. 2, no. 4, 2012.
[18] X. Yang and Y. Tian, "Effective 3D action recognition using eigenjoints," *J. Visual Commun. Image Represent.*, vol. 25, no. 1, pp. 2–11, 2014.
[19] L. Piyathilaka and S. Kodagoda, "Gaussian mixture based hmm for human daily activity recognition using 3d skeleton features," in *Proc. 8th IEEE Conf. Ind. Electron. Appl.*, Jun. 2013, pp. 567–572.

[20] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 914–927, May 2014.
[21] D. Faria, C. Premebida, and U. Nunes, "A probabilistic approach for human everyday activities recognition using body motion from RGB-D images," in *Proc. RO-MAN: 23rd IEEE Int. Symp. Robot Human Interactive Commun.*, Aug. 2014, pp. 732–737.
[22] J. Shan and S. Akella, "3D human action segmentation and recognition using pose kinetic energy," in *Proc. IEEE Workshop Adv. Robot. Social Impacts*, Sep. 2014, pp. 69–75.
[23] J. Martens and I. Sutskever, "Learning recurrent neural networks with hessian-free optimization," in *Proc. 28th Int. Conf. Mach. Learn.*, New York, NY, USA, 2011, pp. 1033–1040.
[24] M. Muller and T. Roder, "Motion templates for automatic classification and retrieval of motion capture data," in *Proc. ACM SIGGRAPH/Eurographics Symp. Comput. Animation*, 2006, pp. 137–146.
[25] F. Lv and R. Nevatia, "Recognition and segmentation of 3-D human action using HMM and multi-class adaboost," in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 359–372.
[26] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. Workshop Human Activity Understanding 3D Data Conjunction CVPR*, Rhode Island, USA, 2012, pp. 20–27.
[27] X. Yang and Y. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *Proc. CVPR Workshops*, 2012, pp. 14–19.
[28] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1290–1297.
[29] J. Luo, W. Wang, and H. Qi, "Spatio-temporal feature extraction and representation for RGB-D human action recognition," *Pattern Recognit. Lett.*, vol. 50, pp. 139–148, 2014.
[30] S. Gaglio, G. Re, and M. Morana, "Human activity recognition process using 3-D posture data," *IEEE Trans. Human–Mach. Syst.*, vol. 45, pp. 586–597, Oct. 2015.
[31] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, 2010.
[32] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, pp. 16:1–16:43, Apr. 2011.
[33] T. B. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Underst.*, vol. 104, pp. 90–126, Nov. 2006.
[34] M. Ye, Q. Zhang, L. W. 0002, J. Zhu, R. Yang, and J. Gall, "A survey on human motion analysis from depth data.," in *Time-of-Flight and Depth Imaging* (Lecture Notes in Computer Science), vol. 8200, M. Grzegorzek, C. Theobalt, R. Koch, and A. Kolb, Eds. New York, NY, USA: Springer, 2013, pp. 149–187.
[35] J. Aggarwal and L. Xia, "Human activity recognition from 3D data: A review," *Pattern Recognit. Lett.*, vol. 48, pp. 70–80, 2014.
[36] B. Ni, P. Moulin, and S. Yan, "Order-preserving sparse coding for sequence classification," in *Computer Vision ECCV 2012* (Lecture Notes in Computer Science), A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Germany: Springer, 2012, pp. 173–187.
[37] J. Qi and Z. Yang, "Learning dictionaries of sparse codes of 3D movements of body joints for real-time human activity understanding," *Plos One*, vol. 9, no. 12, pp. e114–147, 2014.
[38] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, vol. 2, pp. 524–531.
[39] A. Hyvrinen, J. Hurri, and P. O. Hoyer, *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*, 1st ed. New York, NY, USA: Springer, 2009.
[40] C. Bao, H. Ji, Y. Quan, and Z. Shen, "L0 norm based dictionary learning by proximal methods with global convergence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3858–3865.
[41] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, pp. 1031–1044, Jun. 2010.
[42] T. T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Trans. Inf. Theory*, vol. 57, pp. 4680–4688, Jul. 2011.
[43] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, pp. 480–492, Mar. 2012.
[44] L. Bi, O. Tsimhoni, and Y. Liu, "Using the support vector regression approach to model human performance," *IEEE Trans. Syst., Man, Cybernet.-Part A, Syst. Humans*, vol. 41, pp. 410–417, May 2011.
[45] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 9–14.
[46] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," in *Proc. CVPR*, 2011, pp. 1297–1304.
[47] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian, "Human daily action analysis with multi-view and color-depth data," in *Computer Vision - ECCV 2012. Workshops and Demonstrations* (Lecture Notes in Computer Science), vol. 7584, A. Fusiello, V. Murino, and R. Cucchiara, Eds. Berlin, Germany: Springer, 2012, pp. 52–61.
[48] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, (Sydney, Australia), 2013, pp. 3551–3558.

Jin Qi received the M.Sc. degree in mathematics from Sichuan Normal University, Chengdu, China, in 2002, and the Ph.D. degree in computer science from the Institute of Automation, Chinese Academy of Science, Beijing, China, in 2005.

From 2005 to 2007, he was with the Department of Electrical Engineering, University of Electronic Science and Technology of China, as an Assistant Professor and from 2007 to 2009, as an Associate Professor. In 2010, he was a Visiting Scholar with the University of Chicago. From 2011 to Oct. 2013, he was a Postdoc with the Northwestern University. From Nov. 2013 to Oct. 2014, he was with the Medical College of Georgia. Since Nov. 2014, he has been with Children's National Medical Center as a Postdoc. His research interests include fingerprint recognition, biometrics, image processing, computer vision, pattern recognition, and machine learning.



Zhangjing Wang received the Bachelor's and Master's degrees in electrical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 1999 and 2008, respectively. Since 2010, she has been working toward the Ph.D. degree in the same university in electrical engineering.

Her research interests include signal processing, image processing, and radar system design.

Xiancheng Lin received the Bachelor's degree from Southwestern Normal University, Chongqing, China, in 1990, and the Master's degree from Sichuan Normal University, Chengdu, China, in 2002, both in geography. He received the Ph.D. degree in geography from Chengdu Institute of Technology, Chengdu, China, in 2008.

Since 2002, he has been with Sichuan Normal University. His research interests include signal processing and image processing.

Chunming Li received the Ph.D. degree in electrical engineering from the University of Connecticut, Storrs, CT, USA, in 2005.

He is currently a Professor of electrical engineering with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include image processing, computer vision, and medical imaging.

Dr. Li was the recipient of the IEEE Signal Processing Society Best Paper Awards in 2013 and 2015. He was an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, and as a referee and committee member for a number of international conferences and journals relating to image processing, computer vision, medical imaging, and applied mathematics.