JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 12, No. 1, March 2018, pp. 1-11

Human Activities Recognition Based on Skeleton Information via Sparse Representation

Suolan Liu
School of Information Science and Engineering, Changzhou University, Changzhou, China;
University of Texas at Dallas, Richardson, TX, USA
lan-liu@163.com

Lizhi Kong
School of Materials Science and Engineering, Changzhou University, Changzhou, China
lzkong@cczu.edu.cn

Hongyuan Wang*
School of Information Science and Engineering, Changzhou University, Changzhou, China
hywang@cczu.edu.cn

Abstract
Human activities recognition is a challenging task due to its complexity of human movements and the variety performed by different subjects for the same action. This paper presents a recognition algorithm by using skeleton information generated from depth maps. Concatenating motion features and temporal constraint feature produces feature vector. Reducing dictionary scale proposes an improved fast classifier based on sparse representation. The developed method is shown to be effective by recognizing different activities on the UTD-MHAD dataset. Comparison results indicate superior performance of our method over some existing methods.

Category: Human-Computer Interaction

Keywords: Activity recognition; Skeleton feature; Temporal feature; Sparse representation

I. INTRODUCTION

Recognition of human activities has raised considerable interest in the area of computer vision. The main goal of this research is to achieve automatic analysis and classification. In the past few decades, activities recognition has involved using video sequences captured by color cameras. However, the inherent limitation of this sensing device seriously influences the recognition accuracy and restricts previous methods [1-3]. For example, spatiotemporal-based approaches are widely used in human activities recognition by using traditional RGB sequences.

These approaches rely on the detection and representation of spatiotemporal volumes. Based on Laplacian pyramid, Shao and Zhen [4] decompose videos into a series of sub-band feature 3D volumes and present a novel descriptor,

Open Access http://dx.doi.org/10.5626/JCSE.2018.12.1.1 http://jcse.kiise.org

Received 25 April 2017; **Revised** 06 July 2017; **Accepted** 08 January 2018
*Corresponding Author

 pISSN: 1976-4677 eISSN: 2093-8020

called spatiotemporal Laplacian pyramid coding, for holistic human action representation. Laptev et al. [5] do action recognition by extracting spatiotemporal interesting points (STIPs) based on the Harris and Stephens' interest point operators. They extend the Harris corner function defined for the 2D spatial domain into the 3D spatio-temporal domain and employ distinct spatial scale σ and temporal scale τ. Although vision-based activities recognition approaches continue to progress [6-9], their performance is still limited because RGB data is highly sensitive to various factors like illumination changes, variations in viewpoint, occlusions and background clutter [10].

As imaging technology advances and with the release of a low-cost sensor of Microsoft Kinect, it has become possible to capture human skeletal information in real-time from depth sequences by using Kinect for Windows SDK 2.0 [11]. We can represent the human body as an articulated system of rigid segments connected by joints. Therefore, we can consider this activity as a continuous evolution of the spatial configuration of these rigid segments [12].

In this paper, we present an algorithm for human activities recognition from skeleton sequences. We test the effectiveness of our method from the perspective of recognition accuracy. Kinect captures skeletal data. We normalize the skeletal joints coordinates in orientation and scale and then produce the skeleton motion map (SMM) generated by accumulating skeleton trajectories. We project skeleton motion map in three projective views, including front view, side view and top view, to reduce computational complexity and improve computational efficiency. By experimental testing and analysis, the most effective projective information is selected and used as feature descriptor. Sparse representation is a hot research topic in recent years [13-15]. Motivated by its success in face classification, we design a fast sparse representation classifier with regularized least square for human activities recognition. We can summarize our contributions as follows: first, we present an algorithm to produce skeleton motion map. Second, we present using a histogram of gradient (HOG) of SMM as a descriptor for skeletal structure information and use temporal difference feature as sequence constraint. Third, the improved sparse representation classification algorithm cannot only compress the dictionary and reduce the complexity, but also ensure the recognition performance at the same time.

We have organized the rest of the paper as follows: in Section II, we review previous work related to human activities recognition from a skeleton. We introduce our feature extraction algorithm in Section III. In Section IV, we introduce an improved classifier with dictionary optimization based on sparse representation. In Section V, the experimental results demonstrate the effectiveness of our proposed framework with comparisons to other published results. Finally, we conclude the paper in Section VI.

II. RELATED WORKS

Skeleton-based approaches have become popular thanks to the work of Shotton, who proposed an algorithm to accurately estimate the 3D locations of skeletal body joints in super real-time from single depth images [16, 17], and the release of Microsoft Kinect. Roughly, the existing skeleton-based methods may be divided into two categories: joint-based approaches and body part-based approaches [18]. Joint-based approaches consider the human skeleton simply as a set of points. The 3D point positions are often used as features; either the x, y, z coordinates are used directly without any post-processing [19], or they are normalized to be invariant to orientation and scale [20, 21].

On the other hand, body part-based approaches consider the human skeleton as a connected set of rigid segments (body parts). These approaches either model the temporal evolution of individual body parts or focus on (directly) connected pairs of body parts and model the temporal evolution of joint angles [18, 22].

Based on skeletons extracted from depth maps, many methods are proposed to classify. In [23] the bag of words approach is used to model the action sequence and extract features from the entire sequence and quantize them using k-means.

The distance of each feature is computed and classification is done by computing (dis)similarity between two sequences. Furthermore, Ofli et al. [22, 23] also used multiple kernels learning (MKL) to compute the optimal linear combination of multi-view similarities for the task of action recognition. In [11], the authors propose a representation of human pose by the histogram of 3D joints (HOJ3D). They project the sequence of histograms using linear discriminant analysis (LDA) and label each posture using the k-means algorithm. Each posture label from a sequence is fed into a discrete hidden Markov model (HMM) that gives the matching probability for each action [21].

Qiao et al. [24] define the trajectorylet as a novel local descriptor that captures static and dynamic information in a short interval of joint trajectories, and exemplar-SVM (ESVM) is used to learn a large set of detectors for a large number of sampled trajectorylets, one for each sampled trajectorylet.

Then for each action instance, they select a few discriminative trajectorylet detectors as candidate detectors of discriminative trajectorylet. Evaluating on standard datasets demonstrated that this method obtains superior results over existing approaches under various experimental setups.

Sparse representation (sparse coding) has been successfully used for solving many classification problems [25-27].

For example, in [28], sparse linear representation for dictionary-based classification is applied to the log-covariance matrices to recognize human activities from video data. Yuan et al. [29] propose a novel Multi-Task

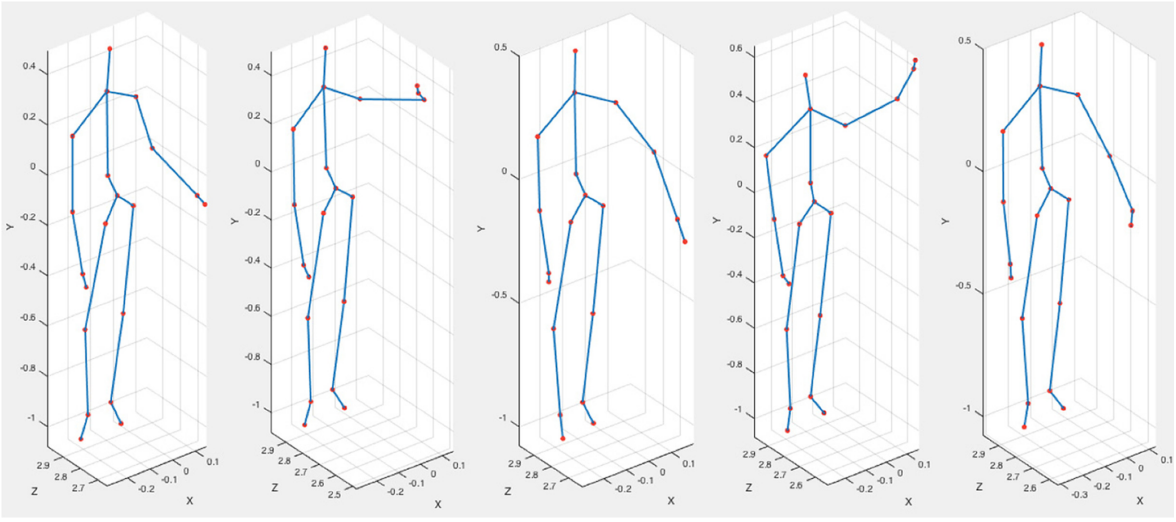


Fig. 1. Skeleton examples of the action 'draw circle' (clockwise).

Sparse Learning (MTSL) model combined with beta process prior to human action recognition. The MTSL model enforces the robustness in coefficient estimation compared with performing each task independently. Besides, the sparseness is achieved via the beta process formulation. In our work, we applied a sparse coding-based approach to human activities recognition from skeletal data. The experiments show that our classifier performs very favorably against other commonly used classifiers on recognition accuracy.

III. SKELETON FEATURES EXTRACTION

A. Human Body Skeletal Data Acquisition

A Kinect sensor can capture both RGB video and depth information, and extract skeletal data in real-time from depth sequences by using Kinect SDK. Fig. 1 shows some skeleton examples of the action 'draw circle (clockwise)' from UTD Multimodal Human Action Dataset (UTD-MHAD) dataset [30]. Fig. 2 shows the definition of a 3D skeleton with 20 tracked skeleton joints. A skeletal image can provide human body structural information as depth maps. Furthermore, skeleton joints position and angle values can provide more abundant and effective information for action classification.

In the process of movement, parts of human body are in a constantly changing state. Therefore, the selection of benchmark coordinate system is crucial for features extraction based on skeletal data. Inspired by the work in [19], we select the first frame with movements as the standard skeleton, and use its coordinates as the common basis, then transform all the other motion skeletons in the sequence to it. Note that the first frame with movement may not be the first frame in the sequence for there are

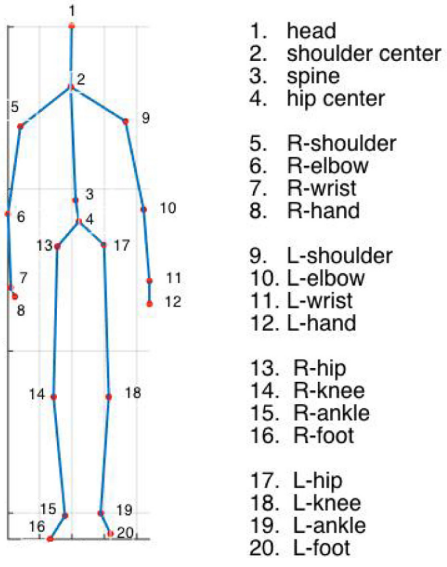


Fig. 2. Skeleton with 20 tracked joints.

some static frames at the beginning. On the other hand, we only selected motion frames for two reasons. First, the static skeletal data does not contribute to the motion characteristics of the video sequence and second, for the consideration of reducing calculation cost.

B. Skeleton Motion Map

Skeleton sequences can provide abundant information for activities recognition, such as shape, structure, position and, angles. In this paper, we propose a novel feature extraction method based on skeleton motion changing for recognition. We accumulate skeletons in an action sequence, and define it as skeleton motion map (*SMM*).

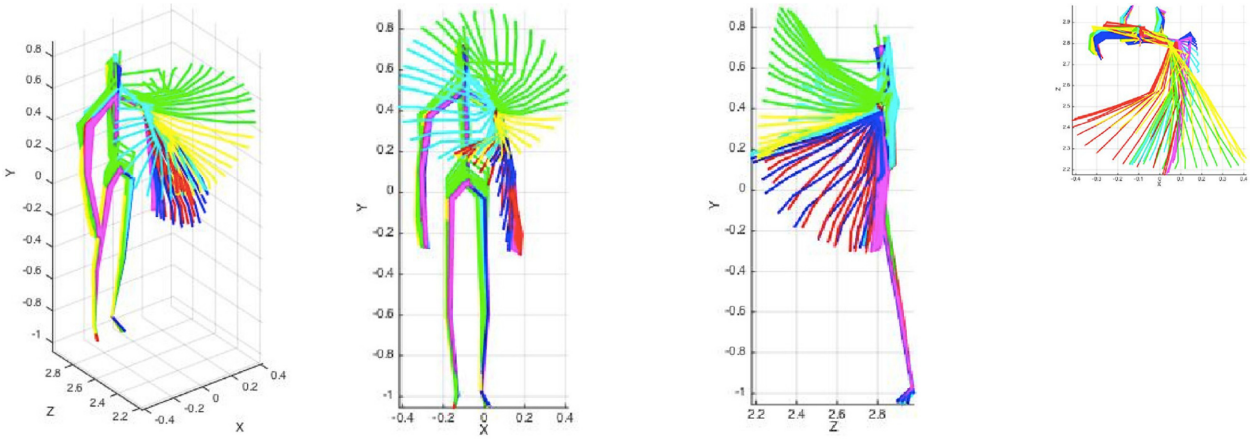


Fig. 3. Skeleton motion maps of 'draw circle'. (a) *SMM*, (b) SMM_f, (c) SMM_s, and (d) SMM_t.

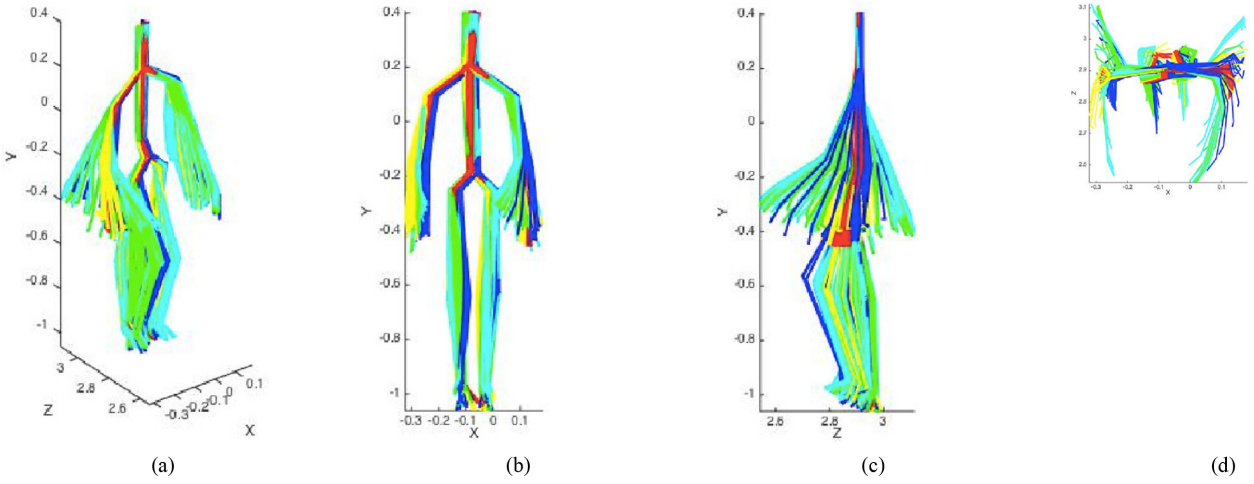


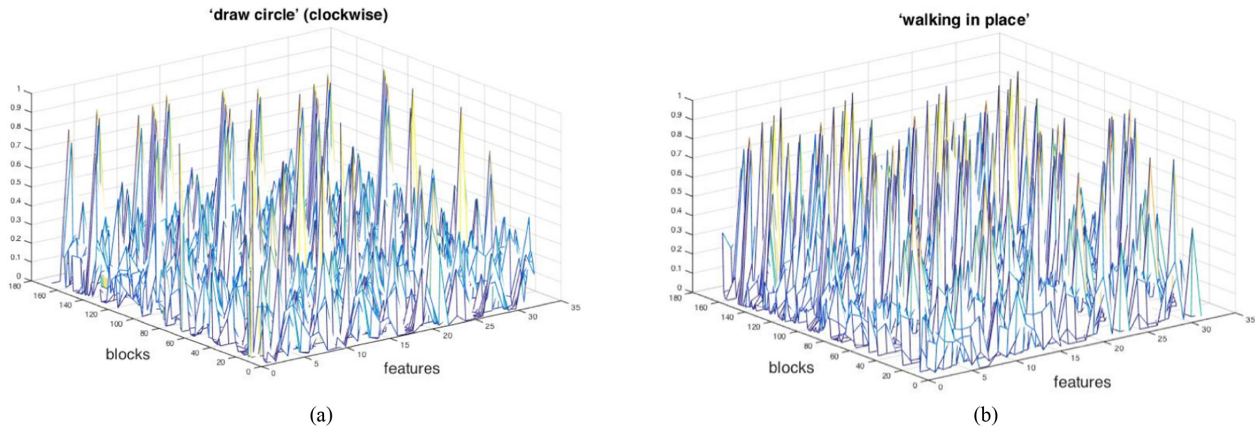
Fig. 4. Skeleton motion maps of 'walking in place'. (a) *SMM*, (b) SMM_f, (c) SMM_s, and (d) SMM_t.

SMM can be seen as a record of action state, so it won't be influenced by the speed.

We projected a skeleton motion map in three projective views defined as front view map (SMM_f), side view map (SMM_s), and top view map (SMM_t) to reduce computational complexity and extract the more effective information for classification. Two examples of *SMM* and its three projected maps are shown in Fig. 3 and Fig. 4 produced from two different activities of 'draw circle (clockwise)' and 'walking in place'. We render these map images over time by setting line width as two and colors as blue, blue-green, green, yellow, red, and fuchsia for every 15 frames from the first frame to the last one to facilitate observation. Here we set 6 different colors because the average number of skeleton frames in action from the used dataset is about 70. Compared with visual results, we can find that SMM_t has the lowest impact on classification, and SMM_f can partly capture body movement but also lost some important information. For example, it can express upper limbs movement of draw circle (clockwise) and walking in place, but for walking in place we can hardly extract the lower limbs movement. On the contrary, SMM_s can effectively capture the global motion information. Distinguishable characters between different activities are also very distinct.

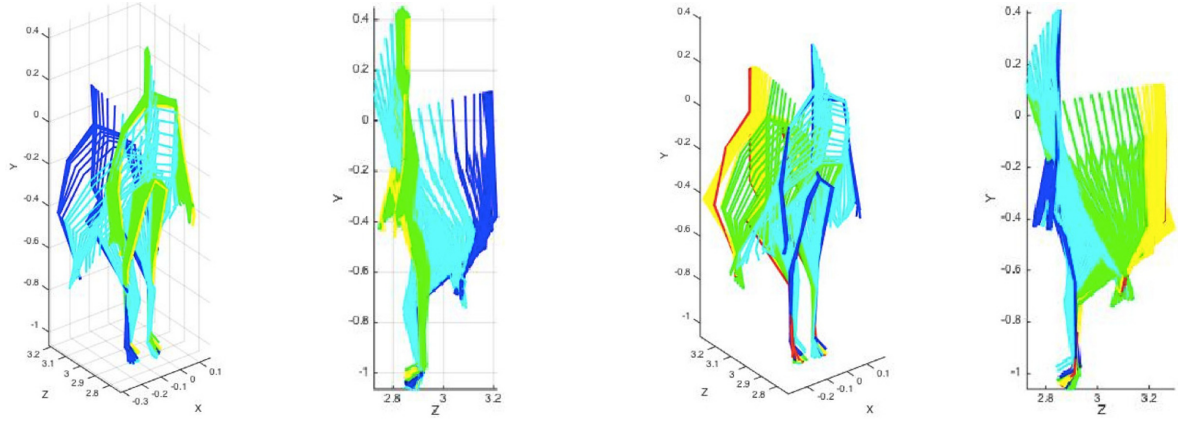
C. SMM_s–HOG Descriptor

HOG introduced by Dalal and Triggs [31] are currently one of the most effective and widely used methods for feature expression [32]. We introduce HOG under skeleton motion map as a novel feature extraction method. It follows the same procedure as HOG on intensity image. Since the feature extraction is based on SMM_s, we define



(a) (b)

Fig. 5. Features of SMM_s–HOG descriptor: (a) 'draw circle' (clockwise) and (b) 'walking in place'.



(a) (b) (c) (d)

Fig. 6. Examples of *SMM* and SMM_s. (a) *SMM* of 'sit to stand', (b) SMM_s of 'sit to stand', (c) *SMM* of 'stand to sit', and (d) SMM_s of 'stand to sit'.

it as an SMM_s–*HOG* descriptor.

It is known that different action sequences may have different sizes.

Although lower resolutions can reduce computational cost in calculating HOG, we extract HOG descriptors only from SMM_s, instead of each frame. So for each sequence, the difference of computational time between different sizes is limited. Due to its computational simplicity, the size of 200×100 suggested in [33] is adopted in this work. The resized SMM_s is denoted by $\overline{SMM_s}$. For each $\overline{SMM_s}$, we define 20×10 non-overlapping cells and 8 gradient orientation bins. The block is composed of 2×2 cells.

Therefore, each SMM_s generates a descriptor HOG with the dimension of (20–1)×(10–1) ×2×2×8=5472. Fig. 5 shows features of SMM_s–*HOG* produced from 'draw circle' and 'walking in place'.

D. Feature of Temporal Constraint

By analysis, we find that some distinct activities may be very similar to each other on *SMM* and SMM_s. For example, Fig. 6 shows *SMM* and SMM_s of two different activities of 'sit to stand' and 'stand to sit'. The high similarity of skeleton maps will result in a serious possibility of failure classification. They contain almost identical frames but reversed in time. Therefore, we need to calculate the time difference of skeleton sequences as temporal constraint and extract it as a kind of feature, which can effectively help to distinguish different activities. For the feature of temporal constraint extraction, we employ the method described in [19] due to its low computational complexity. Lastly, we produce the final feature vector by concatenating this feature with SMM_s–*HOG*.

IV. CLASSIFICATION BASED ON SPARSE REPRESENTATION

The basic idea of classification is to discriminate its category of a new test sample on the condition that C classes of training samples are already labeled. Training samples of the i^{th} object class can be expressed as follows

to classify based on sparse representation:

$$A_i = [v_{i,1}, v_{i,2}, \cdots, v_{i,n_i}] \in R^{m \times n_i} \quad (1)$$

We can describe all training samples of C classes as a matric A:

$$A = [A_1, A_2, \cdots A_i, \cdots A_C] = [v_{1,1}, v_{1,2}, \cdots v_{C,n_C}] \quad (2)$$

Any new test sample $y \in R^m$ from the same class will approximately lie in the linear span of the training samples. We can write the linear representation of y in terms of all training samples as:

$$y = Ax_0 \in R^m \quad (3)$$

where $x_0 = [0, \cdots, 0, \alpha_{i,1}, \alpha_{i,2} \cdots, \alpha_{i,n_i}, 0, \cdots, 0]^T \in R^n$ is a coefficient vector whose entries are zero except for those associated with the i^{th} class.

We can obtain the vector x_0 by solving the linear system of equations $y = Ax$. However, in activities recognition the system $y = Ax$ is underdetermined since the condition of $m>n$ is always not satisfied. We can express it as the following l_1-norm minimization problem to reach a solution:

$$\hat{x} = argmin\|x\|_1 \quad \text{subject to } Ax = y \quad (4)$$

In sparse representation based classification (SRC), the l_1-norm sparsity constraint is imposed on x to make the solution stable.

However, as described in [34], it is the collaborative representation but not the l_1-norm sparsity constraint that improves the classification accuracy. The l_2-regularized approach can generate similar classification results but with a significantly lower computational complexity. Therefore, we can express it as:

$$\hat{x} = argmin\{\|y - Ax\|_2^2 + \lambda|x\|_2^2\} \quad (5)$$

where λ is the regularization parameter. Furthermore, we can derive the Eq. (5) as:

$$\hat{x} = (A^TA + \lambda I)^{-1}A^Ty \quad (6)$$

Collaborative representation can effectively reduce computational cost while keeping high recognition rate [34, 35]. In this work, A can be seen as a redundant dictionary composed by a feature vector of skeleton sequences of all training samples.

So, if we can significantly downscale the redundant dictionary, we can further reduce computational complexity. Here, we propose an improved algorithm based on inter-class dispersion matrix and intra-class dispersion matrix to downscale dictionary by reducing atoms. Intra-class dispersion matrix and inter-class dispersion matrix are defined as follows:

$$S_B = \sum_{i=1}^{C} n_i(u_i - u)(u_i - u)^T,$$

$$S_W = \sum_{i=1}^{C} \sum_{v_{i,j} \in A_i} (v_{i,j} - u_i)(v_{i,j} - u_i)^T \quad (7)$$

where u_i is the mean of i^{th} class training sample A_i. n_i is the statistical probability of i^{th} class activities to all training samples. u is the total mean of all samples. An objective function is constructed by using S_B and S_W.

$$J(A) = \frac{|A^TS_BA|}{|A^TS_WA|} \quad (8)$$

Lagrange multiplier can be used to solve the optimal problem of $J(A)$, and the optimal solution is the downscaled feature vector, expressed as A^{opt}.

For a query activity sample, the class label of y can be obtained by following formula:

$$class(y) = argmin\{err_i\} \quad (9)$$

where $err_i = \|A^{opt}x - A_i\hat{x}_i\|_2$ is the reconstruction error. It is used as discriminant factor, which means that the smallest error is favored.

V. EXPERIMENTAL RESULTS

We conduct experiments on the public UTD-MHAD dataset with the skeletal data captured by a Kinect sensor to evaluate the performance of our approach. Our method is then compared with existing methods in the literature. All the experiments are conducted on a PC equipped with Intel Xeon 3.4 GHz CPU and 16 GB RAM.

A. UTD-MHAD Dataset

We choose UTD-MHAD to test the proposed activities recognition approach. We collected the UTD-MHAD dataset using a Microsoft Kinect sensor and a wearable inertial sensor in an indoor environment. The dataset contains 27 actions performed by 8 subjects (4 females and 4 males). Each subject repeated each action four times. The subjects were required to face the camera during the performance. After removing three corrupted sequences, the dataset includes 861 data sequences. The 27 actions are: right arm swipe to the left, right arm swipe to the right, right hand wave, two hand front clap, right arm throw, cross arms in the chest, basketball shoot, right hand draw x, right hand draw circle (clockwise), right hand draw circle (counter clockwise), draw triangle, bowling (right hand), front boxing, baseball swing from right, tennis right hand forehand swing, arm curl (two arms), tennis serve, two hand push, right hand knock on a door, right hand catch an object, right hand pick up and throw, jogging in a place, walking in a place, sit to stand, stand

Table 1. Two activities subsets from UTD-MHAD dataset

Subset1	"1" right arm swipe to the left, "2" right arm swipe to the right, "3" right hand wave, "4" two hand front clap, "5" right arm throw, "6" cross arms in the chest, "7" basketball shoot, "8" right hand draw x, "9" right hand draw circle (clockwise), "10" right hand draw circle (counter clockwise), "11" draw triangle, "13" front boxing, "14" baseball swing from right, "15" tennis right hand forehand swing, "16" arm curl (two arms), "17" tennis serve, "18" two hand push, "19" right hand knock on a door, "20" right hand catch an object
Subset2	"12" bowling (right hand), "21" right hand pick up and throw, "22" jogging in place, "23" walking in place, "24" sit to stand, "25" stand to sit, "26" forward lunge (left foot forward), "27" squat (two arms stretch out)

to sit, forward lunge (left foot forward), and squat (two arms stretch out). We recorded four data modalities of RGB videos, depth videos, skeleton joint positions, and the inertial sensor signals. In this paper, we only use the data of skeleton points. Some examples of this dataset are shown in Fig. 1.

B. Subsets Setup

In our test, we divided the activities in UTD-MHAD into two subsets as shown in Table 1: stationary and locomotive. In the stationary action subset, we did not need the subject to make moves, such as right hand wave, right hand knock on door. On the other hand, locomotive subset requires the subjects to have a distinct movement of lower limbs; typical examples include jogging in a place and a forward lunge. For each action subset, we performed two different tests. In test one, for each action and each subject, we used the first two action sequences as training samples and the rest as test samples. In test two, subjects #1, 2, 3 and 4 are used for training and subjects #5, 6, 7 and 8 are used for testing.

C. Preprocessing and Parameter Selection

For each skeleton sequence, we first need to preprocess it for neglecting static frames. We consider two consecutive frames at a time and remove the first one if total distance shift between all corresponding skeletal joint points is lower than a set threshold. The purpose of this processing is that the approximate static skeletal data does not contribute to any motion characteristics of the action sequence. Then, we accumulated the preserved frames and extracted the features as described in Section III.

In l_2-regularized approaches, the parameter λ affects the activities recognition accuracy [34, 36]. To find the best value of λ, we did tests on both Subset1 and Subset2 by using methods of test one and test two. The accuracies of different values of λ are shown in Fig. 7. From this figure, one can see that with the increase of sparsity (λ varies from 10^{-6} to 0.1), we obtain a quite stable and good recognition rate. While $\lambda>0.1$, recognition accuracy shows great changes in test two.

Based on this analysis, we set $\lambda = 0.001$ to all experiments in this research.

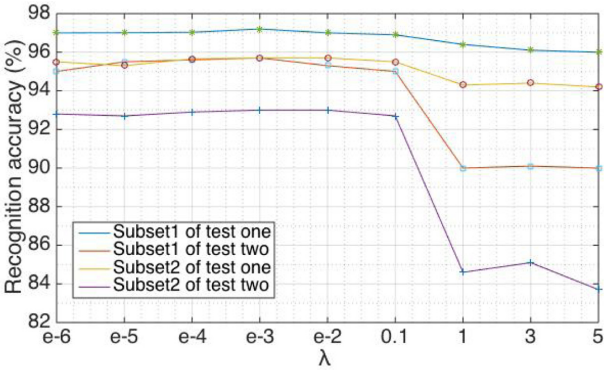


Fig. 7. Recognition accuracy of different values of λ.

D. Comparison with Other Classifiers

We also experimented with other classification algorithms and recorded the obtained classification accuracies to assess the effectiveness of the approach proposed in this paper. We selected three algorithms: first, an algorithm from literature [19] where we used an extreme learning machine to classify, which belongs to the class of single-hidden layer feed-forward neural networks. We extracted skeleton joints position and temporal information from skeletal data. Second, an algorithm from literature [20], where 15 direction cosine angle values are concatenated as a feature vector and multi-class support vector machines are used to classify different activities. Third, an algorithm from literature [12], where skeletal data is expressed as lie group representation and lie group network is used to do action recognition.

The classification accuracies of the developed approach and compared algorithms are shown in Table 2. The best recognition results are highlighted in bold. By comparison, it can be seen that our approach achieves higher accurate rates than other methods reported in [20] and [12] in all tests. However, in Test one and Subset1, our algorithm produces 97.2% accuracy, which is slightly lower than Chen's method of 97.9%. In test two our method outperforms the other three methods.

E. Recognition Rates

The confusion matrix of our proposed scheme for the

Table 2. Comparison of recognition accuracies (%)

	Chen and Koskela [19]	Zhu and Cao [20]	Vemulapalli et al. [12]	Our
Test One				
Subset1	97.9	88.5	95.0	97.2
Subset2	92.5	80.7	82.4	95.7
Average	95.2	84.6	88.7	96.5
Test Two				
Subset1	94.5	88.1	89.7	95.7
Subset2	91.3	73.7	84.1	93.0
Average	92.9	80.9	86.9	94.4

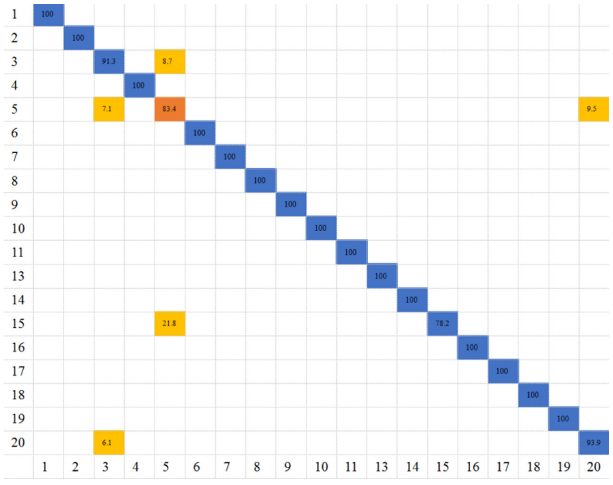


Fig. 8. Confusion matrix of our method for Subset1 of Test One.

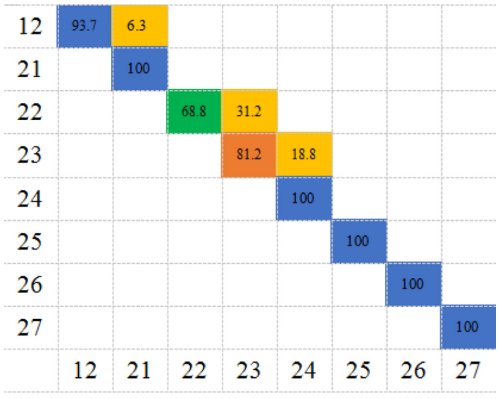


Fig. 9. Confusion matrix of our method for Subset2 of Test Two.

UTD-MHAD dataset is shown in Fig. 8 for Subset1 of test one and Fig. 9 for Subset2 of test two. As shown in Fig. 8 some activities have a high recognition rate of 100%, for example, cross arms on the chest and front boxing. On the contrary, misclassifications occurred on activities such as the right hand wave, tennis right hand forehand swing and right hand catch an object. The main reason of the latter phenomenon may be analyzed as follows: intra-class variations existed in the same activities since they were performed by different subjects. In Fig. 9, low accuracies of 68.8% occurred in jogging in place and 81.2% walking in place activities. In the process of activities collection, some subjects confused these two activities as jogging in low speed and walking in high speed, and they performed the walking activity as running; therefore, the difference was nearly eliminated.

For a better view, we use numbers to represent each activity category, i.e., "1" right arm swipe to the left, "2" right arm swipe to the right, "3" right hand wave, "4" two hand front clap, "5" right arm throw, "6" cross arms on the chest, "7" basketball shoot, "8" right hand draw x, "9" right hand draw circle (clockwise), "10" right hand draw circle (counter clockwise), "11" draw triangle, "13" front boxing, "14" baseball swing from right, "15" tennis right hand forehand swing, "16" arm curl (two arms), "17" tennis serve, "18" two hand push, "19" right hand knock on a door, "20" right hand catch an object, "12" bowling (right hand), "21" right hand pick up and throw, "22" jogging in place, "23" walking in place, "24" sit to stand, "25" stand to sit, "26" forward lunge (left foot forward), and "27" squat (two arms stretch out).

F. Running Time

There are four main components in our method: the SMM_s generation for each skeleton sequence, the computation of temporal constraint for each skeleton sequence, SMM_s–HOG feature generation, and action recognition (using classifier with dictionary optimization). To illustrate the effectiveness of the proposed dictionary optimization approach, we compare it with the normal l_2-regularized algorithm without optimization. The statistical average running time for every component is displayed in Table 3. The total time needed for our method is 22.8 ms. The frame rate in the dataset is 30 frames per second. This means that the processing time for each frame should not

Table 3. Processing time associated with main components

Activities recognition	Processing time (ms/frame)
SMM_s generation	1.8
SMM_s–HOG	4.5
Temporal constraint	7.2
Classifier	
With dictionary optimization	9.3
Without dictionary optimization	11.4

exceed 33.3 ms. Therefore, our method meets the requirement. Although dictionary optimization needs some computational cost, the speed of classification has been improved to about 18.4% compared to a normal classifier. It implies that when more classes and samples are needed to discriminate, our approach will show better performance in the processing time from one-time optimization which will be benefiting to all tests on the subset.

VI. CONCLUSION

In this work, we presented a new framework for human activities recognition using only skeleton joints extracted from depth maps. We extracted features from skeleton motion maps and temporal sequence information. We also introduced a fast classification method based on sparse representation, which improved the structure of redundant dictionary by reducing its scale and enhanced the sparsity. An average recognition rate of about 95.5% on the UTD-MHAD dataset was achieved. The comparison outcomes of the experimentation indicated the superior performance of our method over the compared algorithms.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundations of China (No. 61572085, 61502058), Jiangsu Joint Research Project of Industry, Education and Research (No. BY2016029-15) and Changzhou Science and Technology Support Program (Social Development) Project (No. CE20155044).

REFERENCES

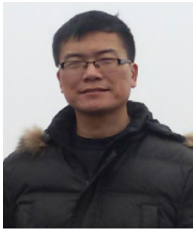
1. R. Slama, H. Wannous, M. Daoudi, and A. Srivastava, "Accurate 3D action recognition using learning on the Grassmann manifold," *Pattern Recognition*, vol. 48, no. 2, pp. 556-567, 2014.
2. J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: a review," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1318-1334, 2013.
3. R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision*, Kyoto, Japan, 2009, pp. 104-111.
4. L. Shao and X Zhen, "Spatio-temporal Laplacian pyramid coding for action recognition," *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 817-827, 2014.
5. I. Laptev, B. Caputo, C. Schuldt, and T. Lindeberg, "Local velocity-adapted motion events for spatio-temporal recognition," *Computer Vision and Image Understanding*, vol. 108, no. 3, pp. 207-229, 2007.
6. R. Qiao, L. Liu, C. Shen, and A. van den Hengel, "Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition," *Pattern Recognition*, vol. 66, pp. 202-212, 2017.
7. J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "RGB-D-based action recognition datasets: a survey," *Pattern Recognition*, vol. 60, pp. 86-105, 2016.
8. D. K. Vishwakarma, R. Kapoor, and A. Dhiman, "A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics," *Robotics and Autonomous Systems*, vol. 77, pp. 25-38, 2016.
9. N. Ashraf, C. Sun, and H Foroosh, "View invariant action recognition using projective depth," *Computer Vision and Image Understanding*, vol. 123, pp. 41-52, 2014.
10. X. Ji, J. Cheng, D. Tao, X. Wu, and W. Feng, "The spatial Laplacian and temporal energy pyramid representation for human action recognition using depth sequences," *Knowledge-Based Systems*, vol. 122, pp. 64-74, 2017.
11. L. Xia, C. C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Proceedings of 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* Providence, RI, 2012, pp. 20-27.
12. R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 588-595.
13. I. Lillo, J. C. Niebles, and A. Soto, "Sparse composition of body poses and atomic actions for human activity recognition in RGB-D videos," *Image and Vision Computing*, vol. 59, pp. 63-75, 2017.
14. P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Anchorage, AK, 2008, pp. 1-8.
15. S. Gao, I. W. H. Tsang, and L. T. Chia, "Kernel sparse representation for image classification and face recognition," in *Proceedings of the 11th European Conference on Computer Vision*, Crete, Greece, 2010, pp. 1-14.
16. M. Devanne, H., Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "Space-time pose representation for 3D human action recognition," in *International Conference on Image Analysis and Processing: New Trends in Image Analysis and Processing*, Naples, Italy, 2013, pp. 456-464.
17. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio,

R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011, pp. 1297-1304.

18. Z. Huang, C. Wan, T. Probst, and L. Van Gool, "Deep learning on lie groups for skeleton-based action recognition," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 6099-6108.
19. X. Chen and M. Koskela, "Skeleton-based action recognition with extreme learning machines," *Neurocomputing*, vol. 149, pp. 387-396, 2015.
20. G. Zhu and L. Cao, "Human motion recognition based on skeletal information of Kinect Sensor," *Computer Simulation*, vol. 31, no. 12, pp. 329-345, 2014.
21. D. C. Luvizon, H. Tabia, and D. Picard, "Learning features combination for human action recognition from skeleton sequences," *Pattern Recognition Letters*, vol. 99, pp. 13-20, 2017.
22. F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (smij): a new representation for human skeletal action recognition," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 24-38, 2014.
23. F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: a comprehensive multimodal human action database," in *Proceedings of 2013 IEEE Workshop on Applications of Computer Vision*, Tampa, FL, 2013, pp. 53-60.
24. R. Qiao, L. Liu, C. Shen, and A. van den Hengel, "Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition," *Pattern Recognition*, vol. 66, pp. 202-212, 2017.
25. J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210-227, 2009.
26. H. Zhang, J. Yang, J. Xie, J. Qian, and B. Zhang, "Weighted sparse coding regularized nonconvex matrix regression for robust face recognition," *Information Sciences*, vol. 394, pp. 1-17, 2017.
27. J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031-1044, 2010.
28. K. Guo, P. Ishwar, and J. Konrad, "Action recognition in video by sparse representation on covariance manifolds of silhouette tunnels," in *Recognizing Patterns in Signals, Speech, Images and Videos*. Heidelberg: Springer, 2010, pp. 294-305.
29. C. Yuan, W. Hu, G. Tian, S. Yang, and H. Wang, "Multi-task sparse learning with beta process prior for action recognition," in *Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, 2013, pp. 423-429.
30. C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proceedings of 2015 IEEE International Conference on Image Processing*, Quebec City, Canada, 2015, pp. 168-172.
31. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, 2005, pp. 886-893.
32. Y. P. Hsu, C. Liu, T. Y. Chen, and L. C. Fu, "Online view-invariant human action recognition using RGB-D spatio-temporal matrix," *Pattern Recognition*, vol. 60, pp. 215-226, 2016.
33. X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proceedings of the 20th ACM International Conference on Multimedia*, Nara, Japan, 2012, pp. 1057-1060.
34. L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: which helps face recognition?," in *Proceedings of 2011 IEEE International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 471-478.
35. C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps," *Journal of Real-Time Image Processing*, vol. 12, no. 1, pp. 155-163, 2016.
36. T. Kerola, N. Inoue, and K. Shinoda, "Graph regularized implicit pose for 3D human action recognition," in *Proceedings of 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Jeju, Korea, 2016, pp. 1-4.

Suolan Liu

She is an associate professor in School of Information Science and Engineering, Changzhou University, China. Her research interests are in pattern recognition and computer vision. She has published more than 30 journal articles for these researches. She received her Ph.D. in computer application technology from Nanjing University of Science and Technology in 2008.



Lizhi Kong

He received the B.S. and Ph.D. degrees in physics and chemistry from University of Science and Technology of China in 2001 and 2008, respectively. Currently, he works as a lecturer in School of Materials Science and Engineering, Changzhou University. His research interests include physical model analysis and numerical analysis.

Hongyuan Wang

He received the Ph.D. degree in computer science from Nanjing University of Science and Technology. He is currently a professor at Changzhou University. His general research interest is in pattern recognition and intelligence system.