# Action Recognition by Fusing Features from Skeleton Sequence

Suolan Liu, Lizhi Kong

Changzhou University, Changzhou 213164, China

lan-liu@163.com

## Abstract

**Skeleton sequences are useful for action recognition since they provide informative clues. However, the inherent drawbacks such as being sensitive to occlusion and the variety among the same actions performed by distinct subjects, which limits the application of skeleton-based method. In this paper, we propose a new framework by fusing spatial feature and temporal difference. In spatial feature extraction, we divide the joint configuration according human physical structure from center to the extremities and extract four independent features from three projective maps insteading of the traditional one whole feature. Multi-layer extreme learning machine with the objective to improve the classification accuracy is used as classifer. We evaluate our method on two public datasets of MSRAction3D and UTD-MHAD. As a result, the proposed scheme outperforms some existing methods. Especially, it can effectively suppress the influence of occlusion.**

## Keywords

**Action recognition, skeleton sequence, fuse features, classification.**

## 1. Introduction

Action recognition has raised considerable research interest during the last two decades. It can facilitate a variety of applications including intelligent surveillance, human-computer interaction, video games and sign language analysis [1-5]. However, how to accurately recognize different actions is still a challenging task. For many years, action recognition has involved using video sequences captured by color cameras. The inherent limitation of this sensing device such as variations of actiors in appearance, illumination changes, complex backgrounds, occlusion and perspective of camera views seriously influences the recognition accuracy and practical application [6-9]. As an alternative to RGB cameras, depth sensors have been popularized in recent severl years by the low-cost and providing abundant information for learning and recognizing. An example of the depth sensor is Microsoft Kinect, which allows capturing RGB sequence and depth sequence. Moreover, skeletion joints can be generated from depth maps in real-time. Skeleton joints are a high discriminant representation that allows efficient extraction of informative clues for action classification. This advance promotes the development of a range of skeleton-based action recognition approaches [10].

Roughly, the existing skeleton-based approaches may be divided into two categories: joint-based approaches and body part-based approaches [11]. Joint-based approaches consider the human skeleton simply as a set of points. 3D points positions are often used as features; either the x, y, z coordinates are used directly without any post processing [12], or they are normalized to be invariant to orientation and scale [13,14]. 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 [11,15].

In summary, although skeleton-based methods have been popular, they cannot perform reliably in practical applications where exist large intra-class variations, such as the action-speed difference, occlusion or variations in the same action performed by different subjects. An action recognition scheme should be independent of the identity of the subjects of different actions and the speed or habits of the performance. Moreover, the approach should be able to support a large number of actions with high classification accuracy and especially with fast classification performance for being used online.

In this paper, we use skeleton-based approach for human action recognition. Our idea is to extract simple yet efficient spatial features and temporal features by using the relationship of body skeletal joints so as to addressing theses issues such as occlusion and action- speed difference. Specially, extreme learning machine(ELM) based approach is applied for the classification, which provides very high recognition accuracy. Moreover, the training and the recognizing are very fast. It makes online classification and application possible for our proposed method.

## 2. Related Work

Spatio-temporal based approaches and sequential based approaches are widely ustilized for human action recognition. Spatio-temporal based approaches consider input image sequences as 3D volume. Features can be extracted from the whole frame or trajectories. In [16], the depth cuboid similarity feature was present to describe the local 3D depth cuboid around the spatio-temporal interest points. Furthermore, a histogram of the cuboid prototypes was used as the action descriptor and SVM for classification. In [17], depth data of an action instance was regarded as a spatio-temporal volume of depth values. Small cuboids were extracted from the volume with selected reference points as centers. Comparative Coding Descriptor (CCD) was proposed to represent the depth information for action analysis. This approach was proved to be robust to viewpoint variations. Dense trajectories within the volume were produed in [4] by sampling dense points from each frame and tracking them based on displacement information from a dense optical flow field. A novel descriptor based on motion boundary histograms was employed for recognition. Histogram of oriented displacements (HOD) was proposed as a new descriptor for 3D trajectories in [18] and a linear SVM classifier was used to achieve human action recognition.

On the other hand, sequential based approaches regard frame sequences as a series of observations, and classify differnet actions according to the degree of similarity between frames in the sequences. Inspired by DNA sequence alignment method used in bioinformatics, authors in [19] present a framework called enhanced sequence matching (ESM), which modeled the new scoring function to measure the similarity between two action sequences. In [20] a hierarchical sequence summarization approach for action recognition was proposed which learned multiple layers of discriminative feature representations at different temporal granularities. Obviously, recognition results based on sequential approaches are seriously affected by the selectiong of similarity measures and classifying criterions and suffered from the position and speed variation.

Extreme learning machine (ELM) proposed by Huang et al. in 2004, which belongs to the class of single-hidden layer feed-forward neural networks [21]. In ELM, the input weights and first hidden layer biases do not need to be learned but are assigned randomly, which makes the learning extremely fast. ELM has been successfully used for solving many classification problems. In [22], ELM is used to recognize human activities from video data based on multiple types of features including spatio-temporal and local static features. Chen et al [12] use ELM to classify different activities by using human skeleton joints position and temporal difference features. Tests on multiple databases of Kinect, mocap and even accelerometer data show high classification accuracies with a few milliseconds required to classify a single motion sequence.

Regardless of the progress mention above, there is still not a promising representation that can be effectively applied for action recognition. In response to this, we propose a novel framework that elegantly fuses features from 3D skeletal data and extreme learning machine (ELM) classifier. Spatial clues and temporal difference are extracted and concatenated to generate final feature from three projective views of skeleton sequence. Through extensive experiments we demonstrate the proposed approach achieves good performance and suppresses the influence of intra-class variations, such as the action-speed difference or variations in the same action performed by different subjects.

## 3. Features Description

Effective and efficient use of the skeletal information is a key to a computationally efficient algorithm for action recognition based on the sequences of skeleton. In this paper, we propose a novel feature representation based on skeleton motion map for recognition. In order to reduce computational complexity and extract the more effective information for classification, we project the skeleton in three projective views defined as front view map ($VM_f$), side view map ($VM_s$), and top view map ($VM_t$).

To a given skeleton sequence with $F$ frames, the $n$th joint on the $f$th frame is formulated as $p_n^f = (x_n^f, y_n^f, z_n^f)^T$, where $f \in (1, \cdots, F), n \in (1, \cdots, N)$, $N$ denotes the total number of skeleton joints in each skeleton. The value of $N$ is determined by some skeleton estimation algorithms. As reported in [14], there are two common different layouts of human body representation. One is composed by 15 joints and another is 20 joints. In this section, we use the joint configuration in the UTD-MHAD dataset [23], where $N$ equals to 20. FIGURE.1 shows the definition of 3D skeleton with 20 tracked skeleton joints. FIGURE.2 shows a skeleton example of the action 'draw circle (clockwise)' from UTD-MHAD dataset [23] and the three projective maps.

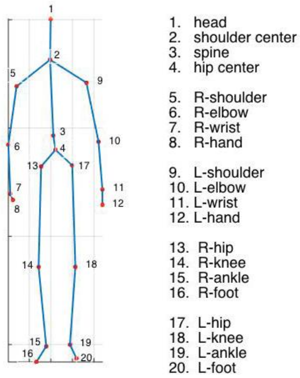
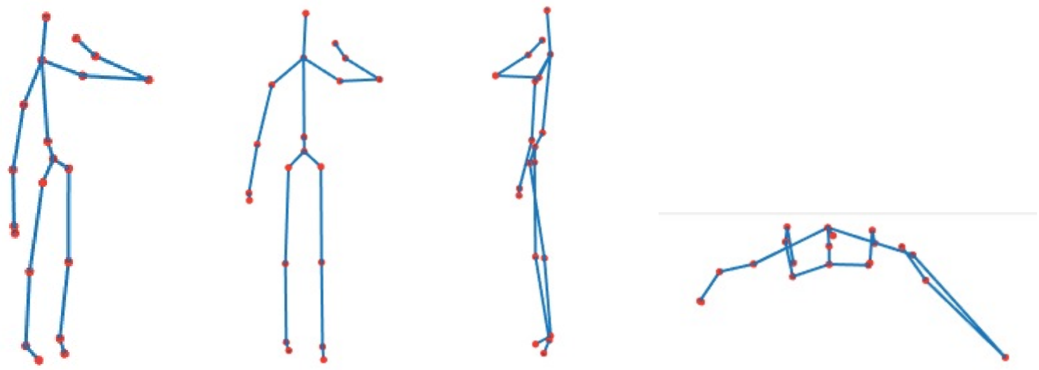



FIGURE.1 Skeleton with 20 joints



(a) 3D skeleton (b) $VM_f$ (c) $VM_s$ (d) $VM_t$

FIGURE.2 Skeleton example and three projective views of the action 'draw circle' (clockwise)

### 3.1 Human Skeleton Normalization and Subgroups Setup

In human action recognition, we need to construct a stand coordinate system since same action done by different subjects can also have different coordinates e.g. due to the different body sizes. 3D coordinates of joints are sensible to different human on the constructed coordinate system. We thus have to build a stable coordinate system and normalization skeleton size. In this paper, the hip center joint is selected as the orignin of the coordinate system and method report in [12] is applied to normalize human 3D joint positions. By this step, the coordinate becomes invariant to different size of perfomers.

Compared to all joints producing one single feature, we divide skeletons into four subgroups to obtaining smaller and more informative features. Occlusion is one of main reasons that affect the recognition accuracy of skeleton-based methods. Therefore, by this processing we expect the independent features can effectively suppress the influence of occlusion. This assumption is verfied by our experiments, as presented in Section 4. In our method, we divide the joints according human physical structure from center to the extremities. Subgroups are shown in TABLE.1.

TABLE.1 Four subgroups from 20 joints

| Subgroup | Skeletal Joints |
|---|---|
| S1 | head, shoulder center, spine, hip center, R-shoulder, R-elbow,R-wrist, R-hand |
| S2 | head, shoulder center, spine, hip center, L-shoulder, L-elbow, L-wrist, L-hand |
| S3 | hip center, R-hip, R-knee,R-ankle,R-foot |
| S4 | hip center, L-hip, L-knee,L-ankle,L-foot |

### 3.2 Features Extraction

It is obvious that the distinctiveness of features significantly influences the effectiveness of the proposed recognition scheme and its practical application. Simple features make feature extraction become easy by reducing computation complexity. On this premise, we extract features that do not require any complex calculations.

In each subgroup, spatial features are extracted from each view of the skeleton joints separately. To simplify calculation, the pair-wise joints' distance of skeleton positions is computed. The joint positions of $N$ joints can be defined as:

$$F_{S,f}^{ij} = \{ J_f^i - J_f^j | i,j = 1,2 \cdots N^S; i \neq j; S = 1,2,3,4\} \tag{1}$$

Where $J_f^i$ is the coordinate of the joint $i$ in the sequence index $f$. $N^S$ denotes the total joints in the subgroup $S$. Then for every subgroup, we concatenate the calculated 3 features from 3 views to generate a subgroup spatial feature. Furthermore, 4 subgroups spatial feature are concatenated to produce the final spatial feature of a frame.

By analysis, we find that some distinct actions may be very similar to each other on skeletons. For example, two different actions of 'sit to stand' and 'stand to sit '. The high similarity of skeleton maps will lead to serious possibility of failure classification. Actually, they contain almost identical frames but reversed in time. Therefore, we need to calculate time difference of skeleton sequences as temporal constraint, extract it as a kind of temporal feature, which can effectively help to distinguish different actions [12].

Let us denote the final spatial feature of the $f$th frame as $F_f$. The temporal feature $F'_f$ can be definedd as follows:

$$F'_f = \begin{cases} F_f & 1 \leq f < f' \\ \dfrac{F_f - F_{f-f'-1}}{\| F_f - F_{f-f'-1} \|} & f' \leq f \leq F \end{cases} \tag{2}$$

Where $f'$ is the temporal offset parameter, $1 < f' < F$.

The feature of a frame is the concatenation of the final spatial feature and temporal feature:

$$F_f = [(F_{1,f})\ (F_{2,f})(F_{3,f})\ \ (F_{4,f})\ \ (F'_f)]^T \tag{3}$$

## 4. Experimental Results and Analysis

We evaluated the accuracy of our method on two publicly available datasets: MSRAction3D [25] and UTD-MHAD [23]. Both datasets are challenging due to some pairs of actions very similar. Base on the work in [21,24], an improved multi-hidden layers extreme learning machine proposed in our previous work [26] is utilized as classifier in this paper. Our method is then compared with some existing methods.

### 4.1 MSRAction3D Dataset

The MSRAction3D is the most common dataset for 3D human action recognition and is composed by 20 actions. Each action was performed by 10 subjects for two or three times.

**Setting1**- The same experimental setting reported in [25] is followed. TABLE.2 lists the three action subsets. For each subset, three different tests are performed. In test one, 1/3 of the samples are used for training and the rest for testing; in test two, 2/3 of the samples are used for training and the rest for testing; in the cross subject test, one half of the subjects (1, 3, 5, 7, 9) are used for training and the rest for testing.

Our method is compared with some existing methods. The comparison results are illustrated in TABLE.3. It can be seen that our method outperforms the method reported in [12], [10] and [11] in test one. In test two, our method produces 99.1% average recognition rate, which is best to all compared methods. For the challenging cross subject test, our method is slightly lower than the method reported in [11] about 2%. However, it should be noted that our method does not need complex matching calculation as described in [11] and thus it is computationally much more effcient.

TABLE.2 Three subsets of actions from the MSRAction3D dataset

| **Action set1 (AS1)** | **Action set2 (AS1)** | **Action set (AS1)** |
|---|---|---|
| Horizontal wave (2) | High wave (1) | High throw (6) |
| Hammer (3) | Hand catch (4) | Forward kick (14) |
| Forward punch (5) | Draw x (7) | Side kick (15) |
| High throw (6) | Draw tick (8) | Jogging (16) |
| Hand clap (10) | Draw circle (9) | Tennis swing (17) |
| Bend (13) | Two hand wave (11) | Tennis serve (18) |
| Tennis serve (18) | Side boxing (12) | Golf swing (19) |
| Pickup throw (20) | Forward kick (14) | Pickup throw (20) |

TABLE.3 Recognition rates (%) comparison of three Tests on MSRAction3D dataset

| Method | test one | | | | test two | | | | cross subject | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AS1 | AS2 | AS3 | Average | AS1 | AS2 | AS3 | Average | AS1 | AS2 | AS3 | Average |
| Chen et al.[12] | 92.5 | 91.2 | 95.4 | 93.0 | 94.7 | 99.7 | 96.0 | 96.8 | 77.2 | 78.1 | 82.4 | 79.2 |
| Vemulapalli et al. [10] | 87.6 | 92.2 | 94.7 | 91.5 | 91.8 | 97.3 | 98.2 | 95.8 | 86.4 | 90.6 | 91.3 | 89.4 |
| Huang et al.[11] | 91.9 | 92.1 | 95.1 | 93.0 | 95.5 | 98.0 | 98.7 | 97.4 | 87.2 | 89.7 | 91.9 | 89.6 |
| Jung et al. [19] | 93.7 | 91.9 | 97.8 | 94.5 | 99.1 | 94.2 | 96.6 | 96.6 | 85.1 | 90.2 | 90.6 | 88.6 |
| Ours | 94.2 | 95.6 | 97.3 | 95.7 | 98.2 | 99.1 | 100 | 99.1 | 83.0 | 87.7 | 92.5 | 87.7 |

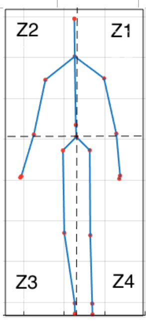



FIGURE.3 Simulation of occlusion

**Setting2**- Occlusion tests. In this setting, we divide human body into 4 parts and randomly occlude some parts to test the robustness of our method. The simulation of occlusion is shown in FIGURE.3.

Futhermore, 1/2 of the samples are used for training and the rest for testing. TABLE.4 shows the recognition rates for the different simulated occlusions. By analysis, we may find that the occlusion of Z1 have significance affects the classification accuracy, while influences from other occlusions or their combinations such as Z3 and Z4 are relatively slight from 5% to 1%.

TABLE 4. Recognition rates (%) for the simulated occlusion

| occlusion | AS1 | AS2 | AS3 |
|---|---|---|---|
| Z1 | 83.6 | 79.4 | 86.8 |
| Z2 | 92.7 | 81.7 | 89.2 |
| Z3 | 94.5 | 90.5 | 98.6 |
| Z4 | 96.1 | 93.8 | 96.3 |
| Z1+Z2 | 73.7 | 76.2 | 82.5 |
| Z1+Z4 | 88.3 | 92.6 | 94.3 |
| Z2+Z3 | 92.7 | 88.4 | 92.7 |
| Z3+Z4 | 95.6 | 93.8 | 98.6 |
| none | 96.1 | 98.5 | 100 |

### 4.2 UTD-MHAD Dataset

UTD-MHAD dataset contains 27 actions performed by 8 subjects (4 females and 4 males). Each subject repeated each action 4 times. The subjects were required to face the camera during the performance. In this dataset, we do the challenging cross subject test, one half of the subjects (1, 2, 3, 4) are used for training and the rest for testing. TABLE.5 shows the average recognition results of our method as well as other 4 methods. As can be seen that our mehtod can produce the most excellent recognition accuracy of 98.6%. It indicates that the proposed feactures can be effecitvely used to distinguish different actions on this dataset.

TABLE.5 Recognition rates (%) comparison on UTD-MHAD dataset

| Method | Recognition rates |
|---|---|
| Chen et al. [12] | 87.5 |
| Vemulapalli et al. [10] | 94.2 |
| Huang et al. [11] | 89.7 |
| Jung et al. [19] | 95.3 |
| **Ours** | **98.6** |

The real-time efficiency of the proposed scheme is further discussed and reported. The average computational time required for extracting a frame and projecting it is 3.8 ms on a PC equipped with Intel Xeon 3.4 GHz CPU with 16 GB RAM. Average time for features extraction and fusion is 9.4 ms. The average classification testing time is 11.6 ms. As a result, the total time needed in our method is about 24.8 ms/frame. The frame rate of the used datasets is 30 frames/s. It means that the processing time should not exceed 33.3ms/frame. Obviously, our method can meet the requirement and be used online.

## 5. Conclusion

In this paper, we propose an effective method to extract discriminative features from skeletal sequence. This feature descriptor combines spatial feature from skeleton maps and temporal feature from time difference. In action recogniton, multi-hidden layers ELM is utilized as the classifer, which can not only improve the learning and classify speed but also enhance the recognition accuracy. The experimental results on two public datasets demonstrate that the proposed method outperforms some existing methods and can reduce the influence of occlusion.

## Acknowledgements

This work is supported by NationalNatural Science Foundations of China (61572085, 61502058), Jiangsu Joint Research Project of Industry, Education and Research (BY2016029-15) and Changzhou Science and Technology Support Program (Social Development) Project (CE20155044).

## References

[1] Y Hsu, C Liu, T Chen, L Fu. Online view-invariant human action recognition using rgb-d spatio-temporal matrix. Pattern recognition, 2016,60:215-226

[2] B Zhang, Y Yang, C Chen. Action recognition using 3d histograms of texture and a multi-class boosting classifier. Image processing, 2017,26(10): 4648-4660

[3] H Chen, G Wang, J Xue. A novel hierarchical framework for human action recognition. Pattern recognition, 2016,55:148-159

[4] H Wang, A Klaser, C Schmid, C Liu. Action recognition by dense trajectories, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2011: 3169–3176.

[5] Z Gao, H Zhang, G Xu, Y Xue. Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition. Neurocomputing, 2015,151:554–564

[6] S Fanello, I Gori, G Metta. Keep it simple and sparse: real-time action recognition. Journal of machine learning research, 2013,14(1): 2617-2640

[7] D Wu, F Zhu, L Shao. One shot learning gesture recognition from rgbd images. Computer Vision and Pattern Recognition Workshops, 2012.

[8] I Junejo, E Dexter, I Laptev, P Perez. View-independent action recognition form temporal self-similarities, IEEE Trans. Pattern Anal. Mach. Intell. 2011,33(1):172-185

[9] Y YI, M Lin. Human action recognition with graph-based multiple-instance learning, Pattern recognition, 2016,53:148-162

[10] R Vemulapalli, F Arrate, R Chellappa. Human action recognition by representing 3D skeletons as points in a Lie group. in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014:588–595.

[11] Z Huang, C Wan. Deep learning on lie groups for skeleton-based action recognition. Computer Vision and Pattern Recognition, 2016

[12] X Chen, M Koskela. Skeleton-based action recognition with extreme learning machines, Neurocomputing, 2015(149): 387-396.

[13] G Zhu, L Cao. Human motion recognition based on skeletal information of Kinect Sensor, Computer Simulation, 2014,12(31): 329-345

[14] C Diogo, T Hedi, P David. Learning features combination for human action recognition from skeleton sequences, Pattern recognition letters, 2017, accepted.

[15] F Ofli, R Chaudhry, G Kurillo. Sequence of the Most Informative Joints (SMIJ): A New Representation for Human Skeletal Action Recognition. In CVPRW, 2012.

[16] L Xia, J Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. Computer vision & pattern recognition, 2013,9(4): 2834-2841

[17] Z Cheng, L Qin, Y Ye.Tian.Human daily action analysis with multi-view and color-depth data. In Computer Vision ECCV. Workshops and Demonstrations, 2012: 52–61

[18] M Gowayyed, M Torki. Histogram of oriented displacements(HOD): describing trajectories of human joints for action recognition. International Joint Conference on Artificial Intelligence, 2013 :1351-1357

[19] H Jung, K Hong. Enhanced Sequence Matching for Action Recognition from 3D Skeletal Data, Computer vision, ACCV 2014:226-240

[20] Y Song, L Morency, R Davis. Action recognition by hierarchical sequence summarization. CVPR, 2013,9(4): 3562-3569

[21] G Huang, Q Zhu, C Siew. Extreme learning machine: a new learning scheme of feedforward neural networks. Proceedings of international joint conference on neural networks (IJCNN 2004), 25-29 July 2004.

[22] R Minhas, A Baradarani, S Seifzadeh, Q Wu. Human action recognition using extreme learning machine based on visual vocabularies, Neurocomputing 73 (10) (2010) 1906–1917.
[23] 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 Proc. IEEE Int. Conf. Image Process., Sep. 2015: 168–172.
[24] G Huang, Q Zhu, C Siew. Extreme learning machine: Theory and applications. Neurocomputing, 70(2006): 489-501.
[25] W Li, Z Zhang, Z Liu. Action recognition based on a bag of 3D points, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010: 9–14.
[26] S Liu, H Wang. Action Recognition Using Key-frame Features of Depth Sequence and ELM, International journal of advanced computer science and applications. 2017,8(10): 52-56.