Action Recognition using Key-Frame Features of Depth Sequence and ELM

Suolan Liu
Department of computer science
Changzhou University
Changzhou, China

Hongyuan Wang
Department of computer science
Changzhou University
Changzhou, China

***Abstract*—Recently, the rapid development of inexpensive RGB-D sensor, like Microsoft Kinect, provides adequate information for human action recognition. In this paper, a recognition algorithm is presented in which feature representation is generated by concatenating spatial features from human contour of key frames and temporal features from time difference information of a sequence. Then, an improved multi-hidden layers extreme learning machine is introduced as classifier. At last, we test our scheme on the public UTD-MHAD dataset from recognition accuracy and time consumption.**

***Keywords*—Action recognition; features; key frame; temporal; extreme learning machine**

I. Introduction

Action recognition has been a hot research topic due to its wild range of applications in many areas, such as intelligent video surveillance, smart living and human-computer interaction [1]-[3]. Although quite a lot of achievements have been reported in the latest several years, human action recognition still has great difficulties [4], [5]. Challenge mainly includes the high intra-class variability, e.g. one same action performed by different subjects and inter-class similarity of actions, e.g. different actions captured from one same person. These difficult issues constraint the further progress of vide technology based on RGB sequences [6], [7]. However, the release of Kinect sensor presents a new idea and more information to resolve these problems [8]-[10]. The Kinect sensor can provide high-resolution RGB images, depth maps and skeleton at same time. Compared with traditional color sequence, depth sequence is invariant and stable to the illumination and body appearance. Besides, it also provides body structure and shape information for action classification. Based on these advantages, many methods were proposed these years. In [11], Chen et al. projected depth images onto three orthogonal planes and a depth motion maps (DMMs) was produced by stacking these projected maps. Histogram of oriented gradients (HOG) [12] was then used as feature descriptor. Xia et al [13] detected STIPs from depth maps directly (called DSTIP) and then used a correction function to remove interest points resulting from noise. Further, for every DSTIP, they extract a depth cuboid similarity feature (DCSF). This feature is applied to describe the local 3D depth cuboid, which size is setting adaptable. Unlike using information only from depth sequences, there are some methods combing multiple information to do action recognition, such as color data, skeleton data and depth maps. Ni et al. [14] proposed two multimodality fusion methods, which is simply based on the concatenation of color and depth sequences. Moreover, two feature representation methods are introduced for action classification. Zhang et al. [15] extracted coarse depth-skeleton (DS) feature by utilizing gradient information from depth sequence and distance information from skeletal joints. To refine the coarse DS feature, they combine the sparse coding approach and max pooling method. Then, the Random Decision Forests (RDF) was used to classify different actions. Hsu et al. [16] introduced a new scheme by producing Spatio-Temporal Matrix Intensity (STMI) from raw RGB and Spatio-Temporal Matrix Depth (STMD) images from depth images respectively. This method was demonstrated to be view-invariant. HoG and HoF features were generated by constructing BoW-Pyramids, which made the classification of reversed actions become possible, such as from sit to stand and from stand to sit. Finally, the presented representation was applied to train a support-vector-machine (SVM) for recognizing different actions. Theoretically, the combination of different attributive data can effectively improve the recognition rate. However, the difficulties and disadvantages are negligible, such as features selection, different dimensional features fusion, training and testing times consumption, which has great relationship to judge the algorithm whether can be used on-line or not.

Inspired by the effectiveness of depth-based action recognition, in this paper we propose a novel algorithm for recognition using depth maps. To reduce calculating burden, key frames are produced from skeleton sequence by using joints as spatial-temporal interest points (STIPs) and mapped into depth sequence to represent an action sequence. Human contour is extracted from each key frame. Then feature representation is introduced including features obtained from human contour and temporal difference. Finally, an improved multi-hidden layers extreme learning machine is utilized as classifier for action recognition. The rest of the paper is organized as follows. In Section 2, we introduce key-frame extraction technique. Section 3 describes the proposed feature representation method. In Section 4, an improved method of multi-hidden layers extreme learning machine is presented for performing action recognition. In Section 5, the experimental results demonstrate the effectiveness of our framework from recognition accuracy and time consumption. Finally, we conclude our work in Section 6.

II. Key Frames Extraction

Key frames are usually used as the most informative frames because they can capture the major elements of a sequence. Key frame extraction approaches can be roughly divided into two categories: one is based on the interframe difference and the other is based on clustering [17], [18]. In the approaches of interframe difference, a new key-frame will be extracted if the interframe difference exceeds a setting threshold. Clustering-based approaches try to look for similar low-level features from frames and group them. Then a frame is selected as the key-frame, which locates closely to the cluster center [19], [20]. In this paper, considering skeleton provides detail body joint positions, so key frame extraction method based on distance difference accumulation is proposed. Define a joint position as $P_{i,j} = \{x_{i,j}, y_{i,j}, z_{i,j}\}$, i is frame index and j is joint index. The accumulated difference of the ith frame can be calculated as follows:

$$D_i = \sum_{j=1}^{n} \left\| P_{i,j} - P_{i-1,j} \right\|^2 \quad (1)$$

Where, $\|*\|$ and n denote the Euclidean distance and the number of skeletal joints, respectively.

Usually, the key frames are defined as the motion with maximum or minimum D_i within a sliding window. However, in most cases, D_i has low value in the first or last several frames or shows extremes in intervening time. As a result, the extracted key frames will be more centralized, and the sequence cannot be accurately and comprehensively expressed. Here we propose the following steps to address these issues:

1) For an action video with N frames, accumulate the total differences from the second frame to Nth frame and express as:

$$D_N = \sum_{i=2}^{N} D_i \ .$$

2) Set key frames number as K and calculate the average differences increment:

$$D_{avg} = D_N / K .$$

3) From the second frame to Lth frame, we calculate the difference:

$$W_L = D_L - k * D_{avg},\ k \in K .$$

We gain a set $\{W_L\}$ and the minimum value of this set on sth frame. So, the sth frame is the key frame.

The improved algorithm can effectively extract key frames to express the whole sequence. Key frame numbers are mapped to depth sequence and then human contour is extracted. A complete overview of the involved stages can be seen in Fig. 1. We select an action of 'draw circle (clockwise)' from UTD Multimodal Human Action Dataset (UTD-MHAD) as an example [21]. The first row of Fig. 1 shows the extracted six key frames from the skeleton sequence. The second row shows the corresponding depth images. In the third row, we list human contours based on the second row images. To facilitate the next step's feature representation, a treatment method with data smoothing and curve fitting is applied in the processing of contour extraction.

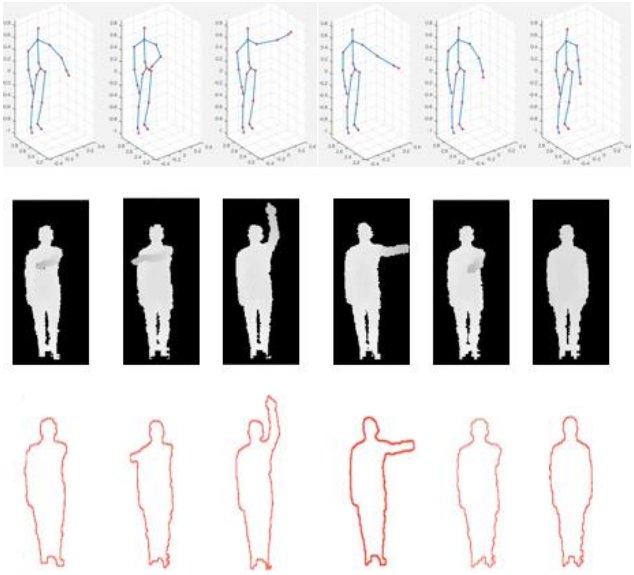

Fig. 1. Human contour based on key-frame extraction.

III. Feature Representation

To a contour with m points $CP = \{cp_1, cp_2, \cdots, cp_m\}$, the contour center $CP_c = (x_c, y_c)$ can be calculated with respect to the m number of points:

$$x_c = \frac{\sum_{i=1}^{m} x_m}{m} \ , \quad y_c = \frac{\sum_{i=1}^{m} y_m}{m} \quad (2)$$

Take the center we can divide the contour into Q radial bins of the same angle. Then, based on the work in [22], [23] we extract features from contour. The point wise Euclidean distances between each contour point and the center of mass are calculated and recorded as cp_i, $\forall\, i \in \{1, \cdots, m\}$. Considering contour points should be in the same order, the corresponding bin q_i of each contour point CP_i is assigned as follows:

$$q_i = \begin{cases} \arccos\left(\frac{y_i - y_c}{cp_i}\right) \times \frac{180}{\pi} & , \text{if } x_i \geq 0 \\ 180 + \arccos\left(\frac{y_i - y_c}{cp_i}\right) \times \frac{180}{\pi} & , \text{else} \end{cases} \quad (3)$$

$$q_i = \left\lfloor \frac{Q \times q_i}{360} \right\rfloor \quad (4)$$

So, the feature vector of each bin can be described as:

$$\bar{v}_i = \frac{f(cp_k, \cdots cp_l)}{q_k, \cdots q_l} \quad (5)$$

$f(cp_k,\cdots cp_l)=\sum_{j=k}^{l}(cp_j - u_i)$ and u_i is the average distance of the contour points of bin q_i. Next, we concatenate Q bins features to form a feature V_k^d for the kth key frame image based on its contour, d is feature dimension.

However, by analyzing we find that some distinct activities may be very similar to each other on key frames. For example, two different activities of 'sit to stand' and 'stand to sit', the high similarity of key frames will lead to serious possibility of failure classification. Actually, they contain almost identical frames but different in time. So, we have to calculate time difference of key frames as temporal features, which can effectively help to distinguish different actions.

Assume kth key frame's original number in depth sequence is k'. The feature of kth key frame in previous work is V_k^d. The temporal difference of feature vector V_k^{td} can be defined as:

$$V_k^{td}=\begin{cases} V_{k'}^d & ,1\le k'<k_o \\ \dfrac{V_{k'}^d - V_{k'-k_o+1}^d}{\left\|V_{k'}^d - V_{k'-k_o+1}^d\right\|} & ,k_o\le k'\le N \end{cases} \quad (6)$$

k_o is the temporal offset parameter $1<k_o<N$.

The final features V of a key frame are concatenation of the spatial feature V_k^d and the temporal feature V_k^{td}, $V=[(V^d)^T(V^{td})^T]^T$.

IV. Action Recognition

Extreme learning machine (ELM) was proposed by Huang et al. [24], [25] as a novel learning algorithm, which is based on the single hidden layer feed forward neural networks (SLFNs). In ELM, the input weights and first hidden layer biases can be assigned randomly instead of learning. This advantage guaranties the learning and classification extremely fast and particularly suitable for online applications.

For N training samples (x_i,t_i), where $x_i=[x_{i1},x_{i2},\cdots,x_{in}]^T\in R^n$ and $t_i=[t_{i1},t_{i2},\cdots,t_{im}]^T\in R^m$. The standard SLFNs have Z hidden neurons. Then the activation function can be formulated as follows:

$$\sum_{i=1}^{Z}\beta_i g(w_i\cdot x_j+b_i)=o_j,\quad j=1,\cdots,N \quad (7)$$

$w_i=[w_{i1},w_{i2},\cdots,w_{in}]^T$ is used as weight vector, which connects the ith hidden neuron and the input. $\beta_i=[\beta_{i1},\beta_{i2},\cdots,\beta_{im}]^T$ is defined as the weight vector connecting the ith hidden node to the output nodes. b_i is applied as the bias term of the ith hidden neuron. If SLFNs can approximate the N samples with zero error, which means that $\sum_{j=1}^{Z}\|o_j - t_j\|=0,\quad j=1,\cdots,N$. There exist β_i, w_i and b_i such that $\sum_{i=1}^{Z}\beta_i g(w_i\cdot x_j+b_i)=t_j$.

It can be expressed as the following matrix calculation:

$$H\beta=T \quad (8)$$

Where, $H=\begin{bmatrix} h(x_1) \\ \vdots \\ h(x_n) \end{bmatrix}=\begin{bmatrix} g(x_1;w_1,b_1)\cdots g(x_1;w_Z,b_Z) \\ \vdots \quad \ddots \quad \vdots \\ g(x_n;w_1,b_1)\cdots g(x_n;w_Z,b_Z) \end{bmatrix}_{N\times Z}$.

There exists $\beta=\begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_Z^T \end{bmatrix}_{Z\times m}$ and $T=\begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N\times m}$.

In ELM the input weight and bias are initialized and valued randomly, the output weight can be generated by solving the least square of β. In the condition where the number of hidden nodes is same with the number of input samples, the resulting H matrix will be square and invertible. But in actual applications, the number of hidden nodes is always not equal to the input samples, which makes H non-invertible. As a result, β can be formulated as finding a least squares solution $\hat{\beta}$.

$$\begin{aligned} &\left\|H(w_1\cdots w_Z,b_1\cdots b_Z)\hat{\beta} - T\right\| \\ &=\min_{w_i,b_i,\beta_i}\left\|H(w_1\cdots w_Z,b_1\cdots b_Z)\beta - T\right\| \end{aligned} \quad (9)$$

To enhance the stability of the numerical solution of SLFNs, a regularization coefficient γ is given by considering the application of ridge regression method and Tikhonov regularization. The least-squares solution of (8) can be expressed as follows:

$$\hat{\beta}=H^T(HH^T+\gamma I)^{-1}T \quad (10)$$

Therefore, the output function of ELM can be expressed as $o(x)=g(x)\hat{\beta}$.

Let P_k^s denote the probability to an input sample x_s, whose output is o_k. $d_j, j=1,\cdots,Z$ is variable quantity of activation function of hidden layer nodes. Then, we construct a matrix W composed by w_i and $b_i, i=1,\cdots,Z$:

$$W=\begin{bmatrix} b_1 & b_2 & \cdots & b_Z \\ w_{11} & w_{12} & \cdots & w_{1z} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nZ} \end{bmatrix}_{n'z} \quad (11)$$

So, P_k^s can be formulated as follows:

$$P_k^s = [w_{11}, \cdots, w_{nZ}, b_1, \cdots, b_Z, c_1, \cdots, c_n, d_1, \cdots, d_Z, j\ ]^T \quad (12)$$

$c_q \in \{0,1\}, q = 1, \cdots, n$ is a binary variable. It is used to express an input reserved or not.

Multi-hidden layers ELM is a multilayer neural network based on extreme learning machine. It not only can approximate complicated function but also does not need iteration during the training process. It has much better generalization performance and processing rate.

For single hidden layer feed forward neural network, the activation function usually defined as a sigmoidal function. But for multi-hidden layers ELM, we define the activation functions as follows:

$$g(x_i; w_j; b_j) = \begin{cases} 0 & , d_j = 0 \\ 1 / 1 + \exp(-(w_j \times x_i + b_j)) & , d_j = 1 \\ w_j \times x_i + b_j & , \text{else} \end{cases} \quad (13)$$

Once P_k^s is calculated, the next formulation is used to discriminate the final classification of an input:

$$C_k(x_s) = \max_k \{P_k^s\} \quad (14)$$

V. Experimental Results and Analysis

In this section, we test our proposed action recognition scheme on the public UTD-MHAD [21] dataset that consists of depth sequences and skeleton data. Our method is then compared with some existing methods.

A. UTD-MHAD Dataset and Tests Setting

The dataset records 27 different actions performed by 8 persons (4 females and 4 males). Each subject repeated each action 4 times. The subjects were required to face a Kinect during the performance. The same experimental settings as reported in [26] are followed in our tests. 20 actions are divided into three subsets as illustrated in Table 1. In test one, half of the action samples are utilized for training and the rest for testing; in test two, 3/4 action samples are applied as training samples; and in the cross subject test, half of the subjects including 1,3,5,7 are applied as training samples and the rest subjects are used for testing.

B. Comparison with Other Methods

In order to evaluate the effectiveness of approach proposed in this paper, our method is compared with the existing methods and the obtained classification accuracies are recorded. Three algorithms are selected: first, algorithm from literature [13]. In this method spatio-temporal information and depth cuboid similarity feature (DCSF) are used. Then bag-of-words is presented for classification. Second, algorithm reported in [27]. A depth motion maps (DMM)-based human action recognition method using l2-regularized collaborative representation classifier is introduced. Third, method in [28] skeleton joint position information with temporal difference is produced as final feature, and extreme learning machine is used for action recognition. The comparison results are listed in Table 2. The best recognition results are highlighted in bold. By comparison, it can be seen that our scheme outperforms the approaches published in [13] in all three test cases. For the challenging cross subject test, algorithm in [28] produces better results on AS2 and AS3. The most probable reason for this may be that actions in the two subsets are more complicated and the proposed accurate joint position information can effectively solve the problems of high intra-class variability and inter-class similarity. In test one and test two, only on action set 1 our results are slightly lower than C Chen's method [27] from 0.3% to 0.6%, while our method shows highest recognition rate in the overall results.

TABLE I. Three Subsets of UTD-MHAD Dataset

Action Set 1 (AS1)	Action Set 2 (AS2)	Action Set 3 (AS3)
1. right arm swipe to the left 2. right arm swipe to the right 3. right hand wave 5. right arm throw 8.right hand draw x 9. right hand draw circle (clockwise) 10. right hand draw circle (counter clockwise) 11. draw triangle 19. right hand knock on door 20. right hand catch an object	4. two hand front clap 6. cross arms in the chest 7. basketball shoot 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	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)

TABLE II. Recognition accuracies(%) of Different tests

		X Lu et al. [13]	C Chen et al. [27]	X Chen et al. [28]	Our method
Test One	AS1	91.1	**95.0**	87.4	94.7
	AS2	87.3	91.4	86.2	**92.8**
	AS3	90.5	93.6	89.1	**94.1**
	Average	89.6	93.3	87.6	**93.9**
Test Two	AS1	91.6	**98.7**	89.1	98.1
	AS2	91.3	93.4	90.5	**95.2**
	AS3	92.5	99.3	91.0	**100**
	Average	91.8	97.1	90.2	**97.8**
Cross Subject Test	AS1	86.4	90.6	73.8	**91.9**
	AS2	72.3	82.7	**87.6**	81.5
	AS3	78.1	83.4	**86.2**	79.3
	Average	78.9	**85.6**	82.5	84.2

The real-time efficiency of the proposed scheme is further discussed and reported. There are three major processing components including key frames computation, features extraction and fusion, and classification. In Table 3, we list the average time needs of each component for the UTD-MHAD dataset. All the experiments are carried out using MATLAB on a PC equipped with Intel Xeon 3.4 GHz CPU with 16 GB

RAM [29]. From the report in Table 3 we can find that the proposed scheme can be applied on a real-time depth video processing which requires the processing rate to be not less than 30 frames per second.

TABLE III. PROCESSING TIMES ASSOCIATED WITH THE COMPONENTS OF OUR METHOD

Action classification	Average processing time (ms/frame)
Key frames extraction	2.3
Features extraction and fusion	6.4
Classification	13.7

VI. CONCLUSION

In this work, we present an action recognition scheme for Kinect captured data. We extract features from human contour of key frame from depth sequence and calculating temporal difference as constraint. We use an improved multi-hidden layers extreme learning machines as the classifier for its high classification accuracy and low time consumption. Experimental results indicate that the proposed features not only can be easily obtained but also provide distinctive information for classification. To further expand our work, we plan to conduct some experiments involved human-human interactions by using method proposed in this paper.

ACKNOWLEDGMENT

This work is supported by National Natural 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] E Paul. K Mohan. Human action recognition using genetic algorithms and convolutional neural networks. Pattern recognition, 2016,59:199-212.

[2] S Megrhi, M Jmal, W Souidene. Spatio-temporal action localization and detection for human action recognition in big dataset. Journal of visual communication and image representation, 2016,41: 375-390

[3] C Yuan, W Hu. Multi-task sparse learning with Beta process prior for action recognition, CVPR, 2013, 9(4): 423-429

[4] N Harbi, Y Gotoh. Aunified spatio-temporal human body region tracking approach to action recognition. Neurocomputing, 2015,161: 56-64

[5] Z Gao, H Zhang. G Xu. Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition. Signal processing, 2015,112:83-97

[6] L Fademerecht, I Bulthoff. Action recognition is viewpoint-dependent in the visual periphery, Vision Research, 2017,135:10-15

[7] J Wang, Z Liu, Y Wu. Mining action let ensemble for action recognition with depth cameras, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1290–1297.

[8] J Zhang, W Li, P Ogunbona. RGB-D-based action recognition datasets: A survey. Pattern recognition, 60(2016):86-105

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

[10] H Liu, M Yuan. RGB-D action recognition using linear coding, Neurocomputing, 2015, 149:79-85

[11] C Chen, R Jafari, N Kehtarnavaz. Action recognition from depth sequences using depth motion maps-based local binary patterns. In: WACV, 2015: 1092–1099

[12] N Dalal, B Triggs. Histograms of oriented gradients for human detection. CVPR2015: 886–893

[13] L Xia, J Aggarwal. Spatio-Temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera. CVPR, 2013.

[14] B Ni, G Wang. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In: ICCV Workshops. (2011) 1147–1153

[15] H Zhang, P Zhong. Combining depth-skeleton feature with sparse coding for action recognition, neurocomputing, 2017, 230: 417-426

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

[17] N Ejaz, T Tariq, S Baik. Adaptive key frame extraction for video summarization using an aggregation mechanism. Journal of Visual Communication and Image Representation, 2012,23:1031–1040

[18] S Lei, G Xie, G Yan. A novel key-frame extraction approach for both video summary and video index, the scientific world journal, 2014.

[19] S Avila, A Lopes. VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recognition Letters, 2011,32: 56–68

[20] S Kuanar, R Panda, A Chowdhur. Video key frame extraction through dynamic Delaunay clustering with a structural constraint. Journal of Visual Communication and Image Representation, 2013, 24: 1212–1227

[21] C Chen, R Jafari, N Kehtarnavaz. Fusion of depth, skeleton and inertial data for human action recognition, Proc. of acoustics, speech and signal processing, 2016.

[22] K Yun, J Honorio. Two-person interaction detection using body-pose features and multiple instance learning.

[23] A Andre, F florez. A low-dimensional radial silhouette-based feature for fast human action recognition fusing multiple views. CVPR2012: 28-35.

[24] 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,2004: 25-29

[25] G Huang, Q Zhu, C Siew. Extreme learning machine: Theory and applications. Neurocomputing, 70(2006): 489-501.

[26] W Li, Z Zhang, Z Liu. Action recognition based on a bag of 3D points. CVPRW, 2010: 9–14

[27] C Chen, K Liu, N Kehtarnavaz. Real-time human action recognition based on depth motion maps. Journal of real-time image processing, 2016,12(1):155-163.

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

[29] S Liu, C Chen, N Kehtarnavaz. A computationally efficient denoising and hole-filling method for depth image enhancement. Real-Time Image and Video Processing , 2016(9897:1-8