IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 10, 2018, accepted November 27, 2018, date of publication December 12, 2018,
date of current version January 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2886362

Combining Adaptive Hierarchical Depth Motion Maps With Skeletal Joints for Human Action Recognition

RUNWEI DING[1], QINQIN HE[1], HONG LIU[1], (Member, IEEE), AND MENGYUAN LIU[2], (Member, IEEE)

[1]Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, Beijing 100871, China
[2]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798

Corresponding author: Hong Liu (hongliu@pku.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant U1613209, in part by the Scientific Research Project of Shenzhen City under Grant JCYJ20170306164738129, and in part by the Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality under Grant ZDSYS201703031405467.

ABSTRACT This paper presents a new framework for human action recognition by fusing human motion with skeletal joints. First, adaptive hierarchical depth motion maps (AH-DMMs) are proposed to capture the shape and motion cues of action sequences. Specifically, AH-DMMs are calculated over adaptive hierarchical windows and Gabor filters are used to encode the texture information of AH-DMMs. Then, spatial distances of skeletal joint positions are computed to characterize the structure information of the human body. Finally, two types of fusion methods including feature-level fusion and decision-level fusion are employed to combine the motion cues and structure information. The experimental results on public benchmark datasets, i.e., MSRAction3D and UTKinect-Action, show the effectiveness of the proposed method.

INDEX TERMS Action recognition, depth data, skeleton joint, depth motion maps.

I. INTRODUCTION

Human action recognition has wide potential applications in video analysis [1], [2], [4], Internet of Things [3], human-computer interaction [5]–[7] and smart surveillance [8], [9]. Previous studies have mainly focus on analyzing human action from videos captured by conventional RGB cameras. Although significant progress has been achieved, recognizing actions remains a challenging task due to some inherent limitations of RGB data, such as illumination changes, partial occlusions and clutter background. The availability of low-cost and easy-to-operate depth sensors like Microsoft Kinect enables easy access to depth data. Compared with RGB data, depth data is robust to changes of lighting conditions. Also, depth data simplifies tasks like foreground subtraction. In addition, depth data provides 3D information of human bodies, which facilities the capture of human actions in the depth direction [15].

The pipeline of human action recognition includes feature extraction, feature encoding and feature classification. In this paper, previous methods are mainly split into two categories: depth-based methods and skeleton-based methods.

Depth-based methods: with the emerge of depth maps, various feature representations have been proposed for human action recognition and three types of features have been widely used. 1) Hyper-surface normals. Oreifej *et al.* [10] proposed a histogram of oriented 4D surface normal (HON4D) to capture complex joint motion and geometry cues. Yang *et al.* [11] extended surface normals by grouping local hyper-surface normals into poly-normal and aggregate low-level poly-normals into Super Normal Vector (SNV). Surface normals are used as local features of actions, which show good robustness to partial occlusions. 2) Cloud points. Li *et al.* [12] sampled a bag of 3D points for depth map and employed an action graph to encode action evolution. Wang *et al.* [13] proposed a random occupancy pattern (ROP) features for 3D action recognition and used sparse coding to encode them. Xia *et al.* [14] extracted local spatio-temporal interest points (DSTIPs) from depth videos and constructed a depth cuboid similarity feature (DCSF) to describe the local 3D depth cuboid around DSTIPs. 3) Depth motion maps (DMMs). DMMs were proposed by Yang *et al.* [17]. They first projected depth maps onto three orthogonal planes and calculated global motion accumulations through entire video

sequences to generate DMMs, then histogram of oriented gradient (HOG) features were computed from DMMs as the representation of the sequences. Zhang *et al.* [18] proposed an edge enhanced depth motion map (E^2DMM) that balances the information weighing between shape and motion for action recognition. To realize real-time action recognition, Chen *et al.* [19] modified the DMMs by omitting the threshold, and then adopted local binary patterns (LBP) to encode texture information of DMMs in [20]. Bulbul *et al.* [21] used Contourlet Transform (CT) on three DMMs to enhance the smooth shape information of human actions. With the popularity of deep learning methods, some methods began to use deep learning for action recognition. For example, Wang and Li, [22] calculated DMMs in three projected planes and input each DMM to a Deep Convolutional Neural Network (ConvNet) for classification. Generally, surface normals and cloud points are local features which ignore global information of human body. DMMs are obtained by calculating accumulated motion of the entire action sequence, where the temporal information of an action is ignored. Therefore, it is difficult to recognize two actions with similar movements and reverse temporal orders, such as actions "*stand up*" and "*sit down*".

Skeleton-based methods: There have been many action recognition approaches based on skeletal joints [16]. Representations based on displacements of skeletal joints are widely used due to their simple structure. For example, Yang *et al.* [23] proposed EigenJoints based on position differences of joints to combine action information including static posture, motion, and offset. Then a Naive-Bayes-Nearest-Neighbor (NBNN) classifier was utilized to classify actions. Zanfir *et al.* [24] proposed moving pose descriptor to consider both pose information and differential quantities of human body joints for low-latency action recognition. Orientation-based methods are also popular since they are robust to body size and camera view variations. Sung *et al.* [25] computed the orientation matrix of each human joint as human posture features and a hierarchical maximum entropy Markov model (MEMM) was proposed for recognition. Zhang *et al.* [26] utilized pairwise features, including orientation between two joints and displacement of one joint between two adjacent frames. Despite the effectiveness of above skeleton-based methods, sole skeleton joints are usually insufficient for recognition task, since problems like partial occlusions will make the estimated skeleton joints rather noisy.

Depth and skeleton fusion methods: Benefiting from the depth maps and skeleton joints, fusion methods have been proposed in the last years. An actionlet ensemble model proposed by Wang *et al.* [41] is learnt to represent each action and capture the intra-class variance. Chen *et al.* [42] presented an approach that depth images, skeleton joint positions, and inertial signals are fused by utilizing three collaborative representation classifiers. Ji *et al.* [43] introduced a skeleton embedded motion body partition approach embedding the skeleton information into depth maps. The human is partitioned to a set of motion parts, which can capture the geometrical structure of human body. Furthermore, a simplified Fisher vector encoding method is used to aggregate local coarse features into a discriminative representation with unified form. Lu *et al.* [44] proposed a multi-feature human action recognition method based on depth images and skeleton data. Multi-feature included DMM (Depth motion map) feature and Quard (Quadruples skeletal) feature. Then, a strategy of multi-model probabilistic voting model was proposed on the classification. Jalal *et al.* [45] proposed an online HAR system which segments human depth silhouettes using temporal human motion information as well as it obtains human skeleton joints using spatiotemporal human body information. Then the spatiotemporal multi-fused features that concatenate four skeleton joint features and one body shape feature are extracted. Zhu *et al.* [46] discussed another feature-level fusion of these two features in by using random forests method. Furthermore, Chen *et al.* [47] designed a dataset, named UTD-MHAD, which consists of four temporally synchronized data modalities. These modalities include RGB videos, depth videos, skeleton positions, and inertial signals from a Kinect camera and a wearable inertial sensor for a comprehensive set of 27 human actions. The database can be used to study fusion approaches that involve using both depth camera data and inertial sensor data.

In this paper, we show the complementary property between depth and skeleton data for high accuracy human action recognition task. Specifically, we present Adaptive Hierarchical Depth Motion Maps (AH-DMMs) for capturing human actions from depth data and present Reference Joint based Distance Features (RDFs) for capturing human actions from skeleton data. Compare with existing DMMs, our proposed AH-DMMs preserves richer temporal and spatial information of human bodies through whole action sequences. The AH-DMMs is original proposed in our previous work [40]. In this work, we further present RDFs to characterize the structure of human skeleton joints, and boost the performance of AH-DMMs by fusing AH-DMMs and RDFs in both feature-level and decision-level. For the feature-level fusion, features of two modalities are merged before classification. For the decision-level fusion, a soft decision-fusion rule is used to combine predictions. The proposed method can not only depicts the global motion and shape information, but also takes local structure information of human actions into consideration. What's more, our method can encode temporal order of action sequences and is adaptive to action speed variations due to AH-DMMs. Our method is evaluated on two benchmark datasets and achieves superior performances over the state-of-the-art approaches. Main contributions of this paper are four-fold.

- We propose an effective human action recognition framework by combining motion features from depth data with structure features from skeleton data.
- Beyond existing Depth Motion Maps (DMMs), we propose Adaptive Hierarchical Depth Motion Maps (AH-DMMs) to capture richer temporal and spatial

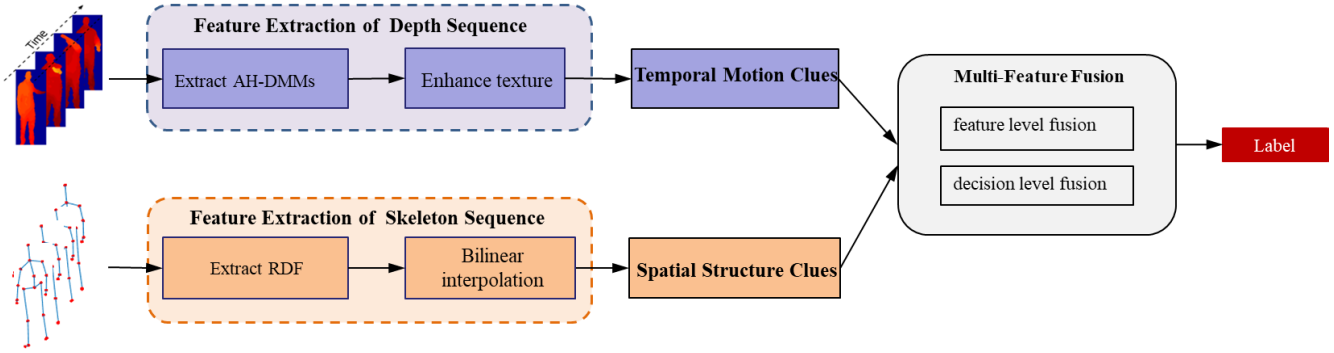

FIGURE 1. The pipeline of the proposed human action recognition framework.

information of human bodies through whole action sequences.

- We present a new skeleton feature called Reference Joint based Distance Features (RDFs), which boost the performances of AH-DMMs by feature fusion in both feature-level and decision-level.
- Our method is robust to noise, speed variations and achieves better performances than existing methods on MSRAction3D and UTKinect-Action datasets.

II. THE PROPOSED METHOD

A. OVERVIEW OF OUR FRAMEWORK

The pipeline of our framework is depicted in Figure 1, which starts with the pre-processing of depth map sequences and skeleton sequences. Then DMMs are calculated based on adaptive hierarchical model, generating AH-DMMs which can capture shape and motion cues. A Gabor filter is employed to enhance the texture information of AH-DMMs. Meanwhile, the spatial distances of skeletal joint positions RDFs are computed to characterize the structure information of human body. Finally, feature-level fusion and decision-level fusion based on logarithmic opinion pool (LOGP) rule are utilized to predict action type.

B. ADAPTIVE HIERARCHICAL DMMs

DMMs were first proposed by Yang *et al.* [17]. Given a depth video sequence with N frames, each frame in the sequence is projected onto three orthogonal planes and DMMs are generated by calculating the difference between two consecutive projected maps. The formulation [40] of DMMs can be expressed as:

$$DMM_{\{f,s,t\}} = \sum_{i=2}^{N} \left| map^{i}_{\{f,s,t\}} - map^{i-1}_{\{f,s,t\}} \right| \tag{1}$$

where $map^{i}_{\{f,s,t\}}$ refers to the projected map of the i-th depth map on front, side and top view.

DMMs can effectively capture the motion and shape information, but have the following drawbacks. First, DMMs lose the temporal order of video sequences which is very important for human action recognition. Two actions with the same movement but reverse temporal order will hardly be recognized using DMMs. The DMMs of such actions remain the same and are not sufficiently discriminative. Moreover, different people could have varied motion speed or frequency when they perform the same action, which may cause inter-class variations and have negative impact on the performance. To solve these problems, we propose AH-DMMs to preserve temporal information. AH-DMMs are calculated over a series of temporal hierarchical windows, and the adaptive windows are selected based on motion energy to make proposed descriptors robust to action speed. The motion energy $ME\ (i)$ of the i-th frame can be calculated using the following formula according to [11]. Here, we modify it by removing the threshold for better computational efficiency [40]:

$$ME\ (i) = \sum_{v=1}^{3} \sum_{j=1}^{i-1} \text{num}\left(\left| map_v^{j+1} - map_v^{j} \right| \right) \tag{2}$$

where the function num $(\cdot)$ returns the number of non-zero elements in the difference map, $v = \{1, 2, 3\}$ corresponds to front, side and top view, $i = 2, \ldots, N$. $ME\ (i)$ reflects the accumulated motion energy from the 1-st frame to the i-th frame, $ME\ (1)$ is defined as 0.

For an action sequence, we calculate the motion energy ME of each frame, denote it as $ME = [ME(1), ME(2), \ldots, ME(N)]$. Then the temporal segments are obtained by the adaptive hierarchical structure shown in Figure 2(a). In this structure, each window size W_l and step length S_l in level l can be computed as follows:

$$W_l = (\frac{1}{2})^{l-1} ME(N) S_l = \frac{1}{2} W_l \tag{3}$$

After dividing ME of the sequence, the frame indices of the segments are utilized to partition the action sequence. Figure 3 presents a specific example of generating AH-DMMs with three levels. The ME is normalized to [0, 1], and $ME(N)$=1. In level 1, DMMs are computed from entire sequence, according to formula(3), $W_1 = 1, S_1 = 0.5$. In level 2, we subdivide action sequence to three over-lapping windows, corresponding to $W_2 = 0.5, S_2 = 0.25$, and then we compute DMMs in each window. In level 3, the DMMs are

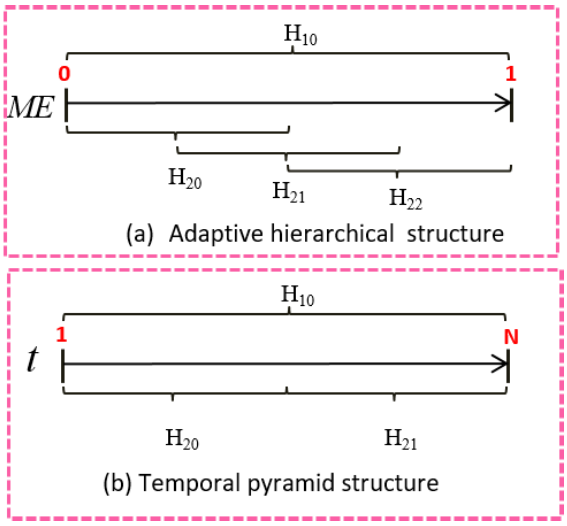

FIGURE 2. Comparison of (a) adaptive hierarchical structure and (b) temporal pyramid structure. H_{lm} denotes the m-th hierarchy in the l-th level.

calculated in seven overlapping segments according to $W_3 = 0.25, S_3 = 0.125$. Consequently, eleven groups of DMMs can be obtained form level 1 to level 3. Then DMM_f, DMM_s, DMM_t from all levels are normalized and concatenated to form our AH-DMM_f, AH-DMM_s, AH-DMM_t, respectively.

In contrast to the temporal pyramid that partitions sequences equally in the time axis without overlapping [27], as shown in Figure 2(b), the adaptive hierarchical structure partitions sequences based on the distribution of motion energy. Therefore, this structure is insensitive to speed variations. Compared with traditional DMMs, the proposed AH-DMMs can preserve temporal information of action sequences, more details of motion and more discriminative shape cues can be involved, as Figure 4 shows.

Because depth maps lacks of texture information, Gabor filters [28] are introduced to characterize the local appearance and shape on AH-DMMs. In this paper, 40 Gabor filters with five scales and eight orientations are generated, and then convolve with the AH-DMMs. For front, side and top views, a d-dimensional Gabor feature vector can be respectively extracted. Then, these three vectors are normalized to $[-1, 1]$ for more accurate action classification and faster convergence. Let $\mathbf{g}_{\mathrm{TA-DMM_f}}, \mathbf{g}_{\mathrm{TA-DMM_s}}, \mathbf{g}_{\mathrm{TA-DMM_t}}$ denote the normalized Gabor feature vectors extracted from three different views. The final feature representation $\mathbf{g}$ is the concatenation of three normalized feature vectors, defined as follows:

$$\mathbf{g} = [\mathbf{g}_{\mathrm{TA-DMM_f}}, \mathbf{g}_{\mathrm{TA-DMM_s}}, \mathbf{g}_{\mathrm{TA-DMM_t}}] \tag{4}$$

C. SKELETAL JOINT FEATURE EXTRACTION

The Kinect device provides 3D data. Figure 5 shows skeleton sequences of some actions from MSRAction3D dataset. To encode the inner structure information of human body, we unitize representations based on relative joint displacements, which compute spatial displacements of coordinates of human skeletal joints in the same frame at a time point. In this paper, we calculate two skeleton features : pairwise relative distance features (PDFs) and reference joint based distance features (RDFs).

Let $p_i = (x, y, d)$ denote a joint of k^{th} frame in a skeleton sequence with N frames, and let n denote the number of joints in one skeleton map, $i = 1, \ldots n$. The PDFs are obtained by calculating the difference between the location of joint i and joint j :

$$p_{i,j} = p_i - p_j, \quad i, j = 1, \ldots n, i \neq j \tag{5}$$

The RDFs are obtained by calculating the coordinate difference of all joints with respect to a reference joint. Given the location of a joint $p_i = (x, y, d)$ and a given reference joint $p_c = (x_c, y_c, d_c)$ in the world coordinate system, the reference joint based distance are obtained using the following formula:

$$\Delta p = p_i - p_c, \quad i, c = 1, \ldots n \tag{6}$$

Then we cascade all the relative displacements from n joints to generate our RDF :

$$RDF_k = [\Delta p_1, \Delta p_2, \ldots, \Delta p_n] \tag{7}$$

The joint distances of all frames are expressed as follows:

$$RDF = \{RDF_1, RDF_2, \ldots RDF_N\} \tag{8}$$

The number of columns of RDF is then scaled from N to N' through bilinear interpolation:

$$RDF' = f(RDF) = \left\{RDF'_1, RDF'_2, \ldots RDF'_{N'}\right\} \tag{9}$$

where $f(RDF)$ is the interpolation function. The RDF' is utilized as the final skeleton feature. In this paper, the hip center joint is used as the reference, since the hip center joint typically keeps stable for most actions.

D. COLLABORATIVE REPRESENTATION CLASSIFIER

Great classification performance and computational efficiency of the collaborative representation classifier (CRC) with l_2 -norm regularization have been shown in image classification [30], face recognition [29] and action recognition [20].

Supposing that there are M training samples from C classes of actions, each action sequence generates a feature vector $\mathbf{g}$ with d dimensions. The training set can be denoted as $\mathbf{G} = [\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_C] \in R^{d \times M}$, where $\mathbf{G}_j = [\mathbf{g_1}, \mathbf{g_2}, \ldots, \mathbf{g_{m_j}}]$ denotes m_j training samples from the j-th class, $(j = 1, 2, \ldots, C)$. Let $\mathbf{x} \in R^d$ denote a testing sample, the collaborative representation [40] with l_2 -norm regularization can be mathematically represented as follows:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ \|\mathbf{x} - \mathbf{G}\boldsymbol{\alpha}\|_2^2 + \lambda \|\mathbf{L}\boldsymbol{\alpha}\|_2^2 \right\} \tag{10}$$

where λ is a regularization parameter, $\boldsymbol{\alpha}$ is a coefficient vector corresponding to all training samples, and $\mathbb{L}$ is the Tikhonov

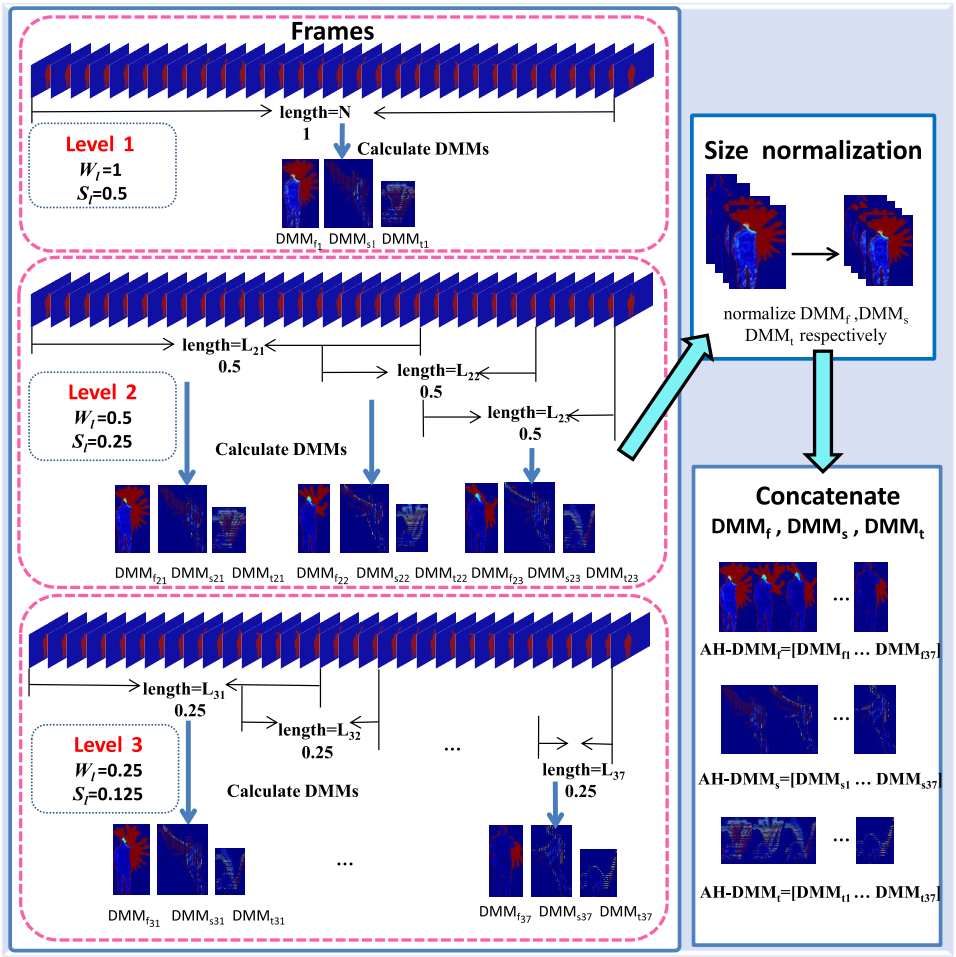

FIGURE 3. The flowchart of generating AH-DMMs with three levels from a depth sequence. L_{lm} denotes the frame length of the m-th window in the l-th level and DMM_{lm} refers to the m-th group DMMs in the l-th level.

regularization matrix, denoted as

$$\mathbf{L} = \begin{bmatrix} \|y - G_1\|_2 & & 0 \\ & \ddots & \\ 0 & & \|y - G_C\|_2 \end{bmatrix} \tag{11}$$

The addition of the regularization term in formula (10) can make the least square solution stable, what's more, it introduces a certain amount of ''sparsity'' to the solution $\hat{\alpha}$, however this sparsity is much weaker than that by l_1 -norm regularization. According to [40], the solution of CRC can be derived as:

$$\hat{\boldsymbol{\alpha}} = \left(\mathbf{G}^T\mathbf{G} + \lambda \cdot \mathbf{L}^T\mathbf{L}\right)^{-1}\mathbf{G}^T\mathbf{x} \tag{12}$$

After obtaining the coefficient vector, the residual errors between the feature vector $\mathbf{x}$ and the approximations can be calculated by:

$$r_j(\mathbf{x}) = \left\|\mathbf{x} - \mathbf{G}_j\hat{\boldsymbol{\alpha}}_j\right\|_2 \tag{13}$$

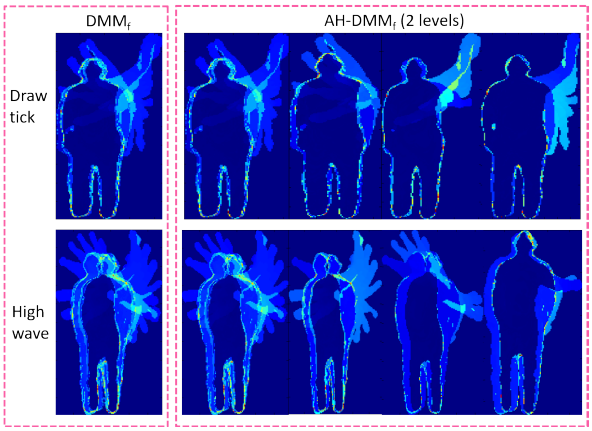

FIGURE 4. Comparison between the proposed AH-DMM_f with traditional DMM_f of actions "*draw tick*" and "*high wave*". It can be seen that AH-DMM_f captures the motion information of both the whole action and sub-movements.

where $\hat{\boldsymbol{\alpha}}_j$ is the coefficient vector associated with class j. Then the class label of $\mathbf{x}$ can be obtained as follows:

$$class(\mathbf{x}) = \arg\min_{j=1,\ldots,C} r_j(\mathbf{x}) \tag{14}$$

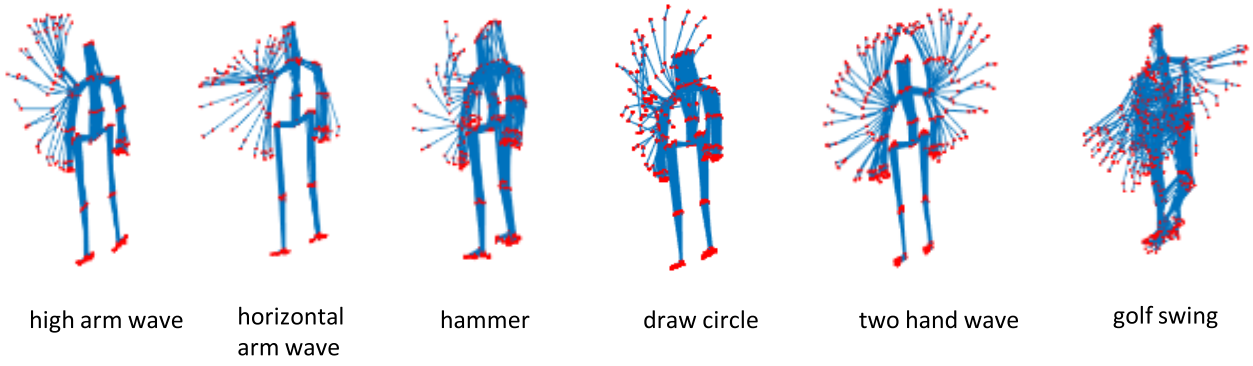

FIGURE 5. Skeleton sequences from MSRAction3D dataset.

E. CLASSIFICATION FUSION

In this paper, two types of fusion consisting of feature-level fusion and decision-level fusion are considered to combine the motion cues and structure information. In the feature-level fusion, features of two modalities are merged before classification. We concatenate AH-DMMs with RDFs as our final representation.

In the decision-level fusion, LOGP is used to combine the classification outcomes. LOGP is a soft decision-fusion rule based on probabilistic decision. Its global membership function is estimated by the posterior probability of all classifiers. The mathematical formula is as follows:

$$P(w_j|x) = \prod_{i=1}^{n} p_i(w_j|x)^{\alpha_i} \tag{15}$$

or

$$log(P(w_j|x)) = \sum_{i=1}^{n} \alpha_i p_i(w_j|x) \tag{16}$$

where $i = 1, 2, \ldots, n$, n is the number of classifiers, C is the number of classes, w is the label of each class and α_i is the weight of the i^{th} classifier. In the CRC, the output function $r_j(x)$ in formula(13) is the residual error between the feature vector and the approximations. According to [31], we can transform a linear function into a probabilistic model.

$$p(w_j|\mathbf{x}) = \frac{e^{Ar_j}}{\sum_{j=1}^{C} e^{(Ar_j+B)}} \tag{17}$$

For simplicity, parameters A and B are set to $A = -1$ and $B = 0$. When $r_j(x)$ is zero, it means that the residual errors is zero, probability value is equal to 1. Conversely the closer the $r_j(x)$ becomes to infinity, the closer the probability value is to zero.

Combining formulae (15) and (17), the final class label of $\mathbf{x}$ is obtained according to

$$class^*(\mathbf{x}) = \arg \max_{j=1,\ldots,C} P(w_j|x)) \tag{18}$$

TABLE 1. The performance of our AH-DMMs using different levels and comparison with the baseline TP-DMMs on MSRAction3D dataset using cross-subject setting.

Levels	TP-DMMs	AH-DMMs	
	Accuracy(%)	Accuracy(%)	Time/Sequence(s)
l=1	90.11	90.11	0.42
l=2	90.91	**94.18**	1.56
l=3	92.36	93.45	3.81

III. EXPERIMENTS AND ANALYSIS

The proposed method is evaluated on two benchmark datasets: MSRAction3D [12] and UTKinectAction [32]. In this section, We firstly introduce the datasets briefly and then describe their evaluation settings respectively, finally we report the experimental results and analysis.

A. MSRAction3D DATASET

MSRAction3D dataset contains 567 videos of 20 actions and each action is performed by 10 subjects for 2 or 3 times. In this dataset, action sequences are collected from a single view and recorded by three channels: RGB, depth and skeletal joint locations. Here, we use the skeleton channel and depth channel. The 3D locations of 20 joints are provided in this dataset. This dataset is challenging since it contains quite similar actions, such as "*draw x*" and "*draw tick*", both of which have similar hands movements, as shown in Figure 6.

1) EXPERIMENTAL SETTINGS

To ensure fair comparisons, we follow the cross-subject setting in [13], this setting uses subject 1, 3, 5, 7, 9 for training and the remaining five subjects for testing. The sizes of DMM_f, DMM_s, and DMM_t are normalized to 102*52, 102*75, and 74*54 respectively, following [19]. The regularization parameter λ of CRC is assigned a value ranging from 0.0001 to 1, the results are shown in Figure 7. It can be seen that $\lambda = 0.001$ leads to the highest recognition accuracy, so we finally choose 0.001 as the default value.

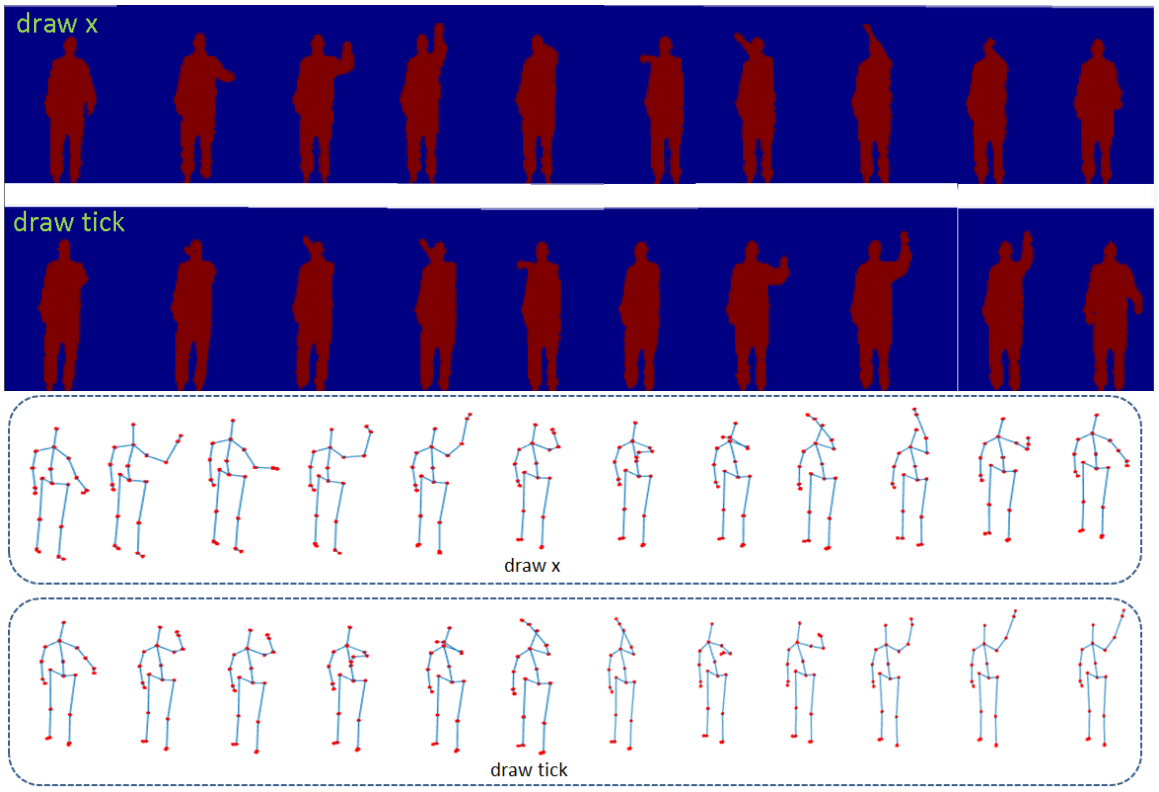

FIGURE 6. Simlar actions "*draw x*" and "*draw tick*" from MSRAction3D dataset.

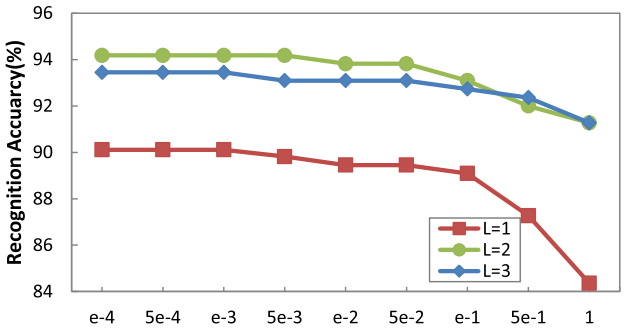

FIGURE 7. Recognition accuracies (%) for AH-DMM with different settings of λ and L.

TABLE 2. Comparison of the performance using different texture descriptors on AH-DMMs (l=2).

Texture Descriptors	HOG	LBP	Gabor	GLCM
Accuracy (%)	93.09	92.00	94.18	93.60

TABLE 3. The recognition accuracies of RDFs and PDFs on MSRAction3D dataset using different classifiers.

Approaches	Accuracy(%)		
	CRC	KLEM	SVM
RDFs	**74.55**	71.64	70.18
PDFs	73.09	69.82	70.18

2) EVALUATION OF THE AH-DMMs

We first evaluate the impact of the level l of our AH-DMMs. Table 1 shows the recognition accuracies while changing the value of l from 1 to 3. It can be observed that when $l = 2$, the proposed AH-DMMs obtains the higest recognition accuracy 94.18% and computation time is 1.56s. The accuracy confusion matrix of twenty actions in MSRAction3D dataset is shown in Figure 8.

In order to verify the effectiveness of AH-DMMs, we compare it with the baseline method: temporal pyramid based DMMs (TP-DMMs),the experimental results are shown in Table 1. It can be seen that AH-DMMs performs better than TP-DMMs when $l = 2$ and 3. It is because that AH-DMMs employed motion energy-based segmentation strategy so it can adapt to action speed variations, whereas TP-DMMs divide video sequence equally without overlapping. What's more, AH-DMMs can capture more discriminative motion cues because they encode the information between two subsequences.

In addition, we have compared the performance of AH-DMMs when using different texture descriptors: Gabor, HOG, LBP and GLCM descriptors. The experimental results are shown in Table 2. It can be seen that Gabor descriptor performs better than other descriptors. Therefore, the Gabor descriptor is chosen to characterize the local appearance and shape on AH_DMMs.

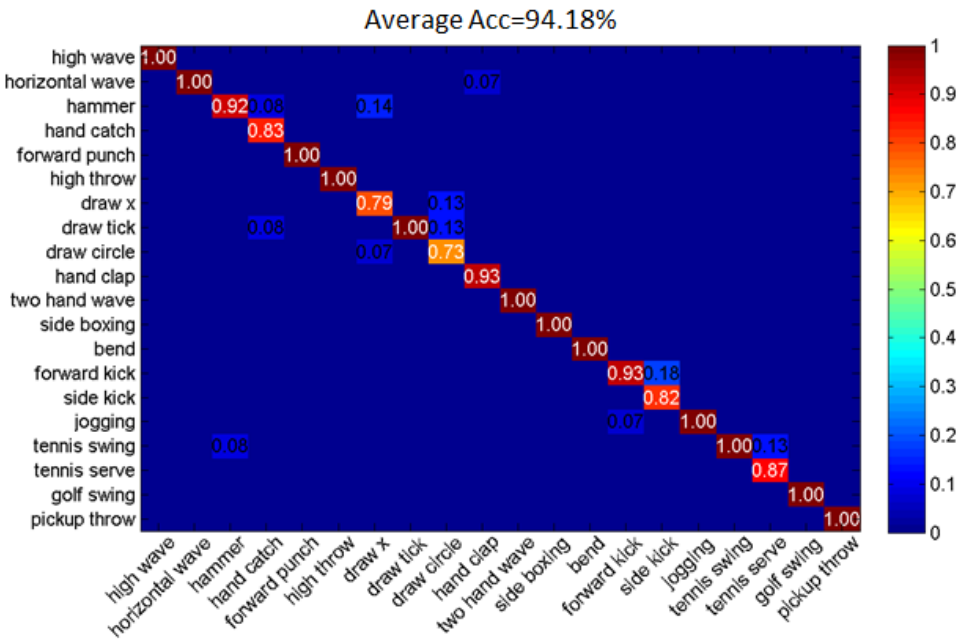

FIGURE 8. The confusion matrix of AH-DMMs for MSRAction3D dataset.

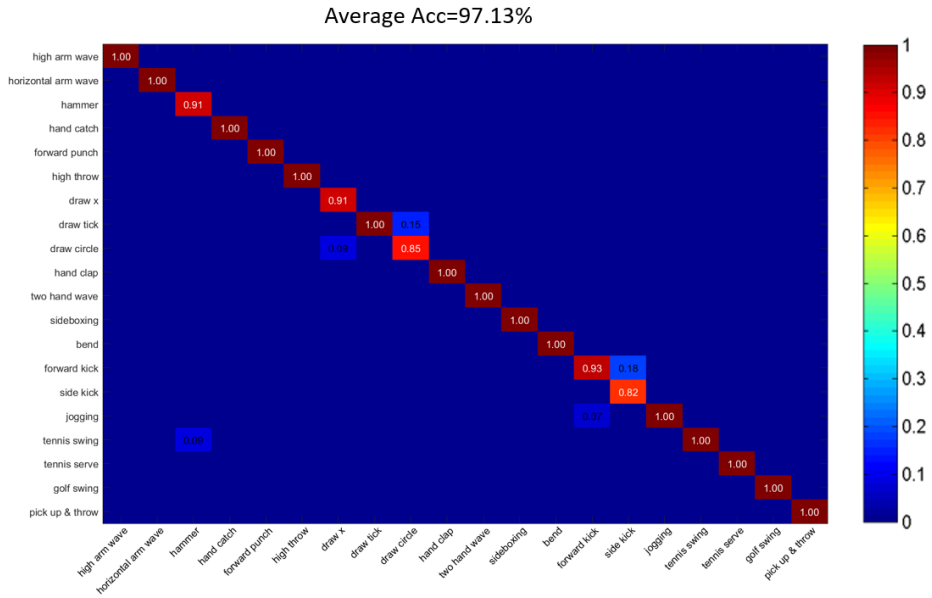

FIGURE 9. The confusion matrix of proposed decision-level fusion method: AH-DMMs + RDFs(DF) for MSRAction3D dataset.

3) EVALUATION OF THE SKELETON FEATURES

We also compare the performance of RDFs with PDFs when using different classifiers : CRC, KELM and SVM. The experimental results are shown in Table 3. It can be observed that RDFs descriptor leads to higher accuracies than PDFs descriptor, and CRC classifier performs better than KELM and SVM.

4) COMPARISON WITH THE STATE-OF-THE-ART

We further compare the performance of our method with several state-of-the-art methods on the MSRAction3D dataset and report the results in Table 4. It can be seen that our feature-level fusion method: AH-DMMs + RDFs(FF) achieves 93.45% recognition accuracy; our decision-level fusion approach AH-DMM + RDFs(DF) achieves the highest recognition accuracy of 97.13%. Decision level fusion approach works better than feature-level fusion approach.

The most comparable methods with our approach are DMM-Quard [44] and Multi-Fused Features [45]. From Table 4, we can see that DMM-Quard [1], [44] approach which employed DMMs achieves 91.30% recognition accuracy, which is much lower than our decision-level fusion

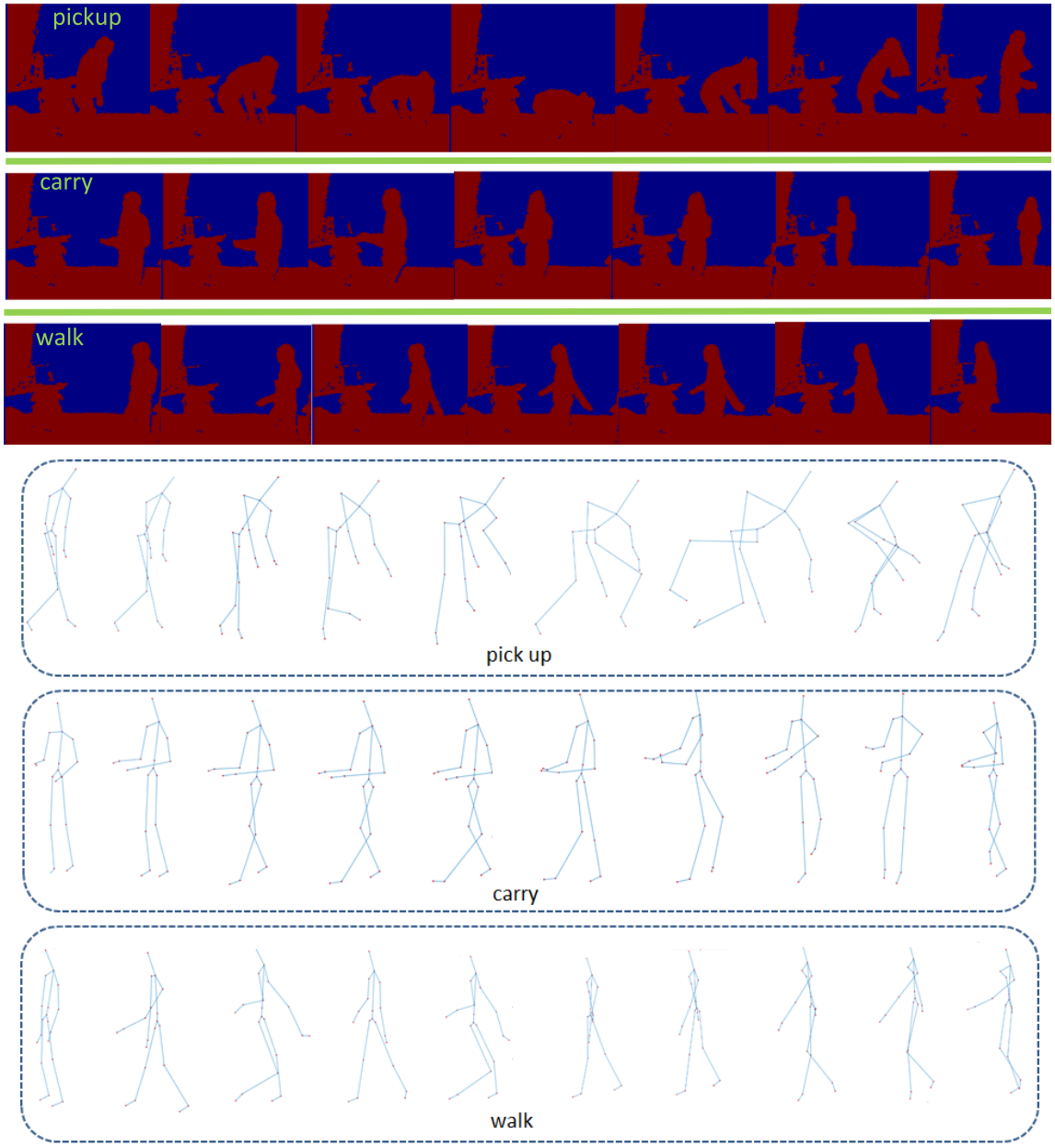

FIGURE 10. Sample images of from UTKinect-Action dataset. Action type from top to bottom: "*pickup*", "*carry*","*walk*". It can be seen that the action"*carry*" and"*walk*" have similar lateral motion patterns, and it has complex background.

approach AH-DMM + RDFs(DF). The Multi-Fused Features [45] approach which fused depth and skeleton informations obtains 93.30% accuracy, which is 3.73% less than the accuracy of our AH-DMMs + RDFs(DF) method. Besides, our AHDMMs + RDFs(DF) descriptor outperforms better than other depth-based methods such as"ROP" [13], "HOG3D + LLC" [33], "Super Normal Vector" [50], "HON4D" [36], "Hierarchical 3D Kernel Descriptors" [34]. Compared with the skeleton + depth based methods, such as "Actionlet ensemble" [41], "Multi-Fused Features" [45], and "Skeleton embedded" [43], the accuracy of our AH-DMMs + RDFs(DF) is higher than them. Our method performs better than these methods mainly attributing to following three aspects. Firstly, our AH-DMMs can sufficiently capture temporal motion information of human actions, and our RDFs can effectively encode the spatial structure information of human body. Secondly, more details and more abundant information of motion can be extracted from human actions; Thirdly, by combining AH-DMMs with RDFs, motion features and structure features can be complementary to each other. The confusion matrix of our decision-level fusion method:AH-DMMs+RDFs(DF) shown in Figure 9. The confusion matrix shows that 15 actions are 100% correctly recognized. The similar actions "*draw x*"

TABLE 4. Comparison with the state-of-the-arts on MSRAction3D dataset under cross-subject setting. ("D" is short or depth feature. "S" is short for skeleton feature.)

Approaches	Year	Accuracy(%)	Type
Bag of 3D Points [12]	2010	74.70	D
DMM-HOG [17]	2012	88.73	D
ROP [13]	2012	86.50	D
Actionlet ensemble [41]	2012	88.20	D+S
HON4D [10]	2013	88.89	D
DSTIP [14]	2013	89.30	D
Lie Group [50]	2014	89.50	S
Skeleton embedded [43]	2018	90.80	D+S
HOG3D+LLC [33]	2016	90.90	D
DMM-Quard [44]	2016	91.30	D+S
Moving Pose [24]	2012	91.70	S
DMM-LBP [20]	2015	93.00	D
Super Normal Vector [11]	2014	93.09	D
Multi-Fused Features [45]	2017	93.30	D+S
MIMTL [48]	2017	93.70	S
Hierarchical 3D Kernel [34]	2016	93.99	D
Trajectorylet detecor [49]	2017	95.90	S
Improved DMMs [1]	2017	96.70	D
AH-DMMs	2017	94.18	D
AH-DMMs+RDFs(FF)	2018	93.45	D+S
AH-DMMs+RDFs(DF)	**2018**	**97.13**	**D+S**

TABLE 5. Comparison with the state-of-the-arts on UTKinect-Action dataset under cross-subject setting. ("D" is short or depth feature. "S" is short for skeleton feature.)

Approaches	Year	Accuracy (%)	Type
DSTIP+DCSF [14]	2013	85.80	D
Synthesized+Pre-trained [22]	2015	90.91	D
histograms of 3D joints [32]	2012	90.92	S
STIP+Joint+RFs [46]	2013	91.90	D+S
eigenjoints [35]	2014	92.38	S
Multiple Features+RFs [51]	2015	92.90	D+S
Lie Group SE [36]	2014	92.97	S
LSTM [37]	2016	97.00	S
Ensemble TS-LSTM [38]	2017	96.97	S
Pairwise joints [39]	2017	97.47	S
AH-DMMs	2017	90.90	D
AH-DMMs+RDFs(FF)	2018	93.94	D+S
AH-DMMs+RDFs (DF)	**2018**	**98.00**	**D+S**

and "*draw tick*" , "*horizontal wave*" and "*high wave*" are successfully distinguished.

B. UTKinect-ACTION DATASET

The UTKinect-Action dataset contains 10 types of human actions in indoor settings and each action is performed twice by 10 subjects. It has totally 199 action sequences. The 3D locations of 20 joints are provided in this dataset. Unlike MSRAction3D dataset, the background of depth maps in UTKinect-Action dataset isn't clear (see Figure 10), which brings difficulty in extracting features based on depth maps. This is a difficult dataset due to its high intra-class variations and complex background. The evaluation setting in this dataset is stili cross-subject setting: subjects (1, 3, 5, 7, 9) are used to train the model and the rest subjects are used for testing. And the parameter settings for UTKinect-Action dataset are the same for MSRAction3D dataset.

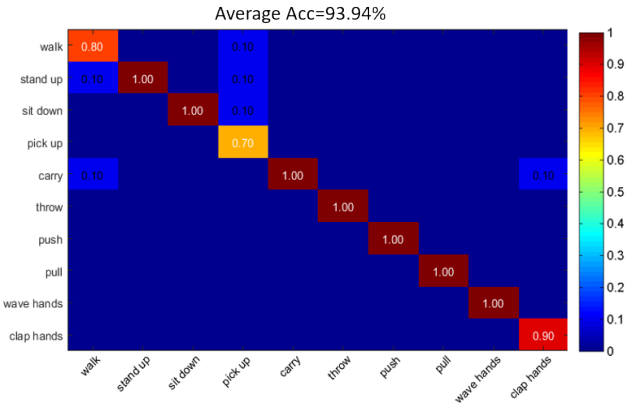

FIGURE 11. The confusion matrix of our feature-level fusion method: AH-DMMs + RDFs(FF) for UTKinect-Action dataset.

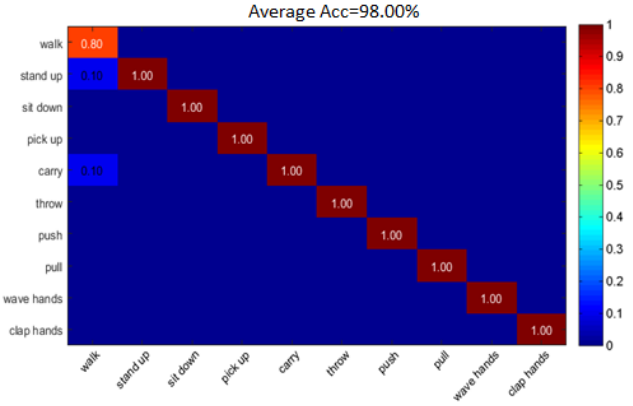

FIGURE 12. The confusion matrix of our decision-level fusion method: AH-DMMs + RDFs(DF) for UTKinect-Action dataset.

Comparison With the State-of-the-Art: We report the experimental results in Table 5. It can be seen that our feature-level fusion method: AH-DMMs + RDFs(FF) achieves 93.94% recognition accuracy in UTKinect-Action dataset. And decision-level fusion method: AH-DMMs + RDFs(DF) achieves the highest recognition accuracy of 98.0% in this dataset. Table 5 also shows the comparison with the state-of-the-art methods, and the proposed AH-DMMs + RDFs(DF) also performs best. In addition, AH-DMMs + RDFs(DF) outperforms some deep learning method such as "LSTM [37] " and "Ensemble TS-LSTM [38]". Compared with the skeleton + depth based methods STIP + Joint + RFs [46] and Multiple Features + RFs [50], the accuracy of our AH-DMMs + RDFs(DF) is 6.1% and 5.1% higher

respectively. These experimental results also proves the effectiveness and robustness of our method.

The confusion matrix of AH-DMMs + RDFs(FF) and AH-DMMs + RDFs(DF) are shown in Figure 11 and Figure 12 respectively. As shown in Figure 12, 9 actions out of 10 actions are 100% correctly recognized. Only one action "*walk*" is confused with the action "*carry*". The similar lateral motion patterns of these two actions (as shown in Figure 10) lead to this classification error.

IV. CONCLUSIONS

This paper shows the complementary property between depth data and skeleton data for human action recognition task. We present the AH-DMMs feature to describe human actions in the depth data. Compared with existing DMMs method, the proposed AH-DMMs feature preserves richer motion and temporal information of human bodies. Meanwhile, the AH-DMMs are adaptive to action speed variations for using energy-based hierarchical structure. To improve the performance of AH-DMMs, we further present the skeletal feature called RDFs to characterize the detail structure information of human body. Both feature-level and decision-level fusion are considered to combine the motion cues and structure information. Extensive experimental results show that our method outperforms these existing approaches on benchmark datasets, which verifies that the combination of depth and skeleton data benefit the human action recognition task.

REFERENCES

[1] C. Chen, M. Liu, H. Liu, B. Zhang, J. Han, and N. Kehtarnavaz, "Multi-temporal depth motion maps-based local binary patterns for 3-D human action recognition," *IEEE Access*, vol. 5, pp. 22590–22604, 2017.

[2] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep Bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2018.

[3] D. Kwon, M. R. Hodkiewicz, J. Fan, T. Shibutani, and M. G. Pecht, "IoT-based prognostics and systems health management for industrial applications," *IEEE Access*, vol. 4, pp. 3659–3670, 2016.

[4] J. K. Aggarwal and L. Xia, "Human activity recognition from 3D data: A review," *Pattern Recognit. Lett.*, vol. 48, no. 1, pp. 70–80, 2014.

[5] I. Rodomagoulakis *et al.*, "Multimodal human action recognition in assistive human-robot interaction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2702–2706.

[6] G. Zhu, "Multimodal gesture recognition using 3-D convolution and convolutional LSTM," *IEEE Access*, vol. 5, pp. 4517–4524, 2017.

[7] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2018, pp. 1159–1168.

[8] W. Lin, M.-T. Sun, R. Poovendran, and Z. Zhang, "Activity recognition using a combination of category components and local models for video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1128–1139, Aug. 2008.

[9] Y. Chen and C. Shen, "Performance analysis of smartphone-sensor behavior for human activity recognition," *IEEE Access*, vol. 5, pp. 3095–3110, 2017.

[10] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 716–723.

[11] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 804–811.

[12] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2010, pp. 9–14.

[13] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3D action recognition with random occupancy patterns," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 872–885.

[14] L. Xia and J. K. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 2834–2841.

[15] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recognit. Lett.*, vol. 34, no. 15, pp. 1995–2006, 2013.

[16] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *Comput. Vis. Image Understand.*, vol. 158, pp. 85–105, May 2017.

[17] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proc. ACM Multimedia*, 2012, pp. 1057–1060.

[18] C. Zhang and Y. Tian, "Edge enhanced depth motion map for dynamic hand gesture recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2013, pp. 500–505.

[19] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps," *J. Real-Time Image Process.*, vol. 21, no. 1, pp. 155–163, 2016.

[20] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in *Proc. WACV*, Jan. 2015, pp. 1092–1099.

[21] M. F. Bulbul, Y. Jiang, and J. Ma, "Human action recognition based on DMMs, HOGs and contourlet transform," in *Proc. Multimedia Big Data*, Apr. 2015, pp. 389–394.

[22] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, and P. Ogunbona, "ConvNets-based action recognition from depth maps through virtual cameras and pseudocoloring," in *Proc. ACM Multimedia*, 2015, pp. 1119–1122.

[23] X. Yang and Y. L. Tian, "Eigenjoints-based action recognition using naive-Bayes-nearest-neighbor," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2012, pp. 14–19.

[24] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 2752–2759.

[25] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from RGBD images," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2012, pp. 842–849.

[26] C. Zhang and Y. Tian, "RGB-D camera-based daily living activity recognition," *J. Comput. Vis. Image Process.*, vol. 2, no. 4, p. 12, 2012.

[27] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2847–2854.

[28] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[29] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 471–478.

[30] M. Yang, L. Zhang, D. Zhang, and S. Wang, "Relaxed collaborative representation for pattern classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2224–2231.

[31] N. M. Nasrabadi, "Pattern recognition and machine learning," *J. Electron. Imag.*, vol. 16, no. 4, p. 049901, 2007.

[32] L. Xia, C. C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2012, pp. 20–27.

[33] H. Rahmani, D. Q. Huynh, A. Mahmood, and A. Mian, "Discriminative human action classification using locality-constrained linear coding," *Pattern Recognit. Lett.*, vol. 72, pp. 62–71, Mar. 2016.

[34] Y. Kong, B. Satarboroujeni, and Y. Fu, "Learning hierarchical 3D kernel descriptors for RGB-D action recognition," *Comput. Vis. Image Understand.*, vol. 144, pp. 14–23, Mar. 2016.

[35] X. Yang and Y. Tian, "Effective 3D action recognition using eigenjoints," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 2–11, 2014.

[36] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 588–595.

[37] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 816–833.

[38] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1012–1020.
[39] M. Liu, C. Chen, and H. Liu, "Learning informative pairwise joints with energy-based temporal pyramid for 3D action recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 901–906.
[40] H. Liu, Q. He, and M. Liu, "Human action recognition using adaptive hierarchical depth motion maps and Gabor filter," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1432–1436.
[41] 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. (CVPR)*, Jun. 2012, pp. 1290–1297.
[42] C. Chen, R. Jafari, and N. Kehtarnavaz, "Fusion of depth, skeleton, and inertial data for human action recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2712–2716.
[43] X. Ji, J. Cheng, W. Fang, and D. Tao, "Skeleton embedded motion body partition for human action recognition using depth sequences," *Signal Process.*, vol. 143, pp. 56–68, Feb. 2018.
[44] Z. Lu, Z. Hou, C. Chen, and J. Liang, "Action recognition based on depth images and skeleton data," *J. Comput. Appl.*, vol. 36, no. 11, pp. 2979–2984, 2016.
[45] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, and D. Kim, "Robust human activity recognition from depth video using spatiotemporal multi-fused features," *Pattern Recognit.*, vol. 61, pp. 295–308, Jan. 2017.
[46] Y. Zhu, W. Chen, and G. Guo, "Fusing spatiotemporal features and joints for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, Jun. 2013, pp. 486–491.
[47] 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. (ICIP)*, Sep. 2015, pp. 168–172.
[48] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, and X. Gao, "Discriminative multi-instance multitask learning for 3D action recognition," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 519–529, Mar. 2017.
[49] R. Qiao, L. Liu, C. Shen, and A. van den Hengel, "Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition," *Pattern Recognit.*, vol. 66, pp. 202–212, Jun. 2017.
[50] Y. Zhu, W. Chen, and G. Guo, "Fusing multiple features for depth-based action recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 2, pp. 1–20, 2015.

RUNWEI DING received the M.S. degree in software engineering from Peking University, China, in 2010, where she is currently pursuing the Ph.D. degree in computer applied technology. She has published articles in the *International Journal of Advanced Robotic Systems*, the IEEE International Conference on Acoustics, Speech, and Signal Processing, and the IEEE International Conference on Image Processing. Her research interests include action recognition and localization.

QINQIN HE received the B.E. degree in electronic information engineering from Peking University, China, in 2015, where she is currently pursuing the master's degree in computer applied technology. She has published articles in the IEEE International Conference on Image Processing and the IEEE International Conference on Acoustics, Speech, and Signal Processing. Her research interests include action recognition and localization.

HONG LIU received the Ph.D. degree in mechanical electronics and automation from the School of EE and CS, Peking University (PKU), China, in 1996, where serves as a Full Professor. He has been selected as the Chinese Innovation Leading Talent supported by National High-level Talents Special Support Plan, since 2013. He is also the Director of the Open Lab on Human Robot Interaction, PKU. He has published more than 150 papers and has received the Chinese National Aero-Space Award, the Wu Wenjun Award on Artificial Intelligence, the Excellence Teaching Award, and was the Candidate of Top Ten Outstanding Professors in PKU. His research fields include computer vision and robotics, image processing, and pattern recognition. He is the Vice President of the Chinese Association for Artificial Intelligent (CAAI) and the Vice Chair of the Intelligent Robotics Society of CAAI. He has served as the keynote speaker, co-chair, session chair, or as a PC member of many important international conferences, such as IEEE/RSJ IROS, IEEE ROBIO, IEEE SMC, and IIHMSP, and recently also serves as a reviewer of many international journals, such as *Pattern Recognition*, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.

MENGYUAN LIU received the Ph.D. degree from the School of Electrical Engineering and Computer Science , Peking University, in 2017. He currently serves as a Research Fellow in EEE, Nanyang Technological University. He has published articles in *Pattern Recognition*, T-CSVT, T-MM, the IEEE ACCESS, *Neurocomputing*, CVPR, IJCAI, ICME, ICASSP, and ICIP. His research interests include human action recognition and action detection.