Multimedia Tools and Applications (2021) 80:32111–32130
https://doi.org/10.1007/s11042-021-11193-4

Human action recognition based on multi-scale feature maps from depth video sequences

Chang Li[1] · **Qian Huang**[1] · **Xing Li**[1] · **Qianhan Wu**[1]

Received: 14 October 2020 / Revised: 22 April 2021 / Accepted: 24 June 2021 /
Published online: 24 July 2021

Abstract

Human action recognition is an active research area in computer vision. Although great progress has been made, previous methods mostly recognize actions from depth video sequences at only one scale, and thus they often neglect multi-scale spatial changes that provide additional information in practical applications. In this paper, we present a novel framework with a multi-scale mechanism to improve scale diversity of motion features. We propose a multi-scale feature map called Laplacian pyramid depth motion images(LP-DMI). First, We employ depth motion images (DMI) as the templates to generate the multi-scale static representation of actions. Then, we caculate LP-DMI to enhance multi-scale dynamic information of motions and reduce redundant static information in human bodies. We further extract the multi-granularity descriptor called LP-DMI-HOG to provide more discriminative features. Finally, we utilize extreme learning machine (ELM) for action classification. The proposed method yeilds the recognition accuracy of 93.41%, 85.12%, 91.94% on the public MSRAction3D, UTD-MHAD and DHA dataset. Through extensive experiments, we prove that our method outperforms the state-of-the-art benchmarks.

Keywords Action recognition · Laplacian pyramid · Multi-scale motion representation · Extreme learning machine

✉ Qian Huang
huangqian@hhu.edu.cn

Chang Li
lichang@hhu.edu.en

Xing Li
lixing@hhu.edu.cn

Qianhan Wu
wuqianhan@hhu.edu.cn

[1] School of Computer and Information, Hohai University, Nanjing, China

1 Introduction

Human action recognition is a hot topic in computer vision, which aims to automatically interpret the semantic information conveyed by human actions and interactions with the external environment. It has many real-world applications, such as security monitoring, intelligent human-computer interaction, smart home, and elderly healthcare etc. [12, 19, 41, 43, 44]. However, this task is still challenging because of problems like illumination, occlusion, varying spatio-temporal scale, clothing, and viewing angles.

Initially, action recognition technology was mainly based on RGB videos acquired by ordinary cameras [10, 49, 52]. However, RGB information is tempted with external factors, such as shooting environment, lighting, and wearing texture, which has limited the development of action recognition. With the introduction of low cost depth sensors, such as Microsoft Kinect, ASUS Xtion and SR-4000, major breakthroughs have been made in human action recognition. Compared to traditional RGB data, depth video sequences provide 3D structure of actions. The pixels of depth maps describe the distance between the surface of objects and sensors [4]. This range information provides convenience for segmenting the foreground person and eliminates the interference caused by complex backgrounds. Therefore, depth maps have better invariance to illumination and texture changes. Actually, human behavior is a tricky task in application scenarios, which contains abundant spatial information in different scales. Over the past few decades, a variety of methods have been investigated to describe depth videos for action recognition [3, 14, 18, 53]. However, the descriptors mentioned in these methods all lack of scale diversity and fail to capture more discriminative features.

Aiming at mining additional multi-scale spatial information from depth video sequences, we motivate to study a novel human action recognition framework with a multi-scale mechanism as illustrated in Fig. 1. We project each frame of depth videos onto three orthogonal Cartesian planes to obtain three-view depth motion images (DMI) which constitutes the 3D action model. After that, we apply the Gaussian pyramid to simulate the scale changes of human eyes and obtain the static multi-scale representation of human motions. Then, we construct Laplacian pyramids to generate the compact feature map LP-DMI which enhances the dynamic multi-scale information for action recognition, thus LP-DMI-HOG capturing multi-granularity motion features can be extracted following the pyramid structure. Finally,

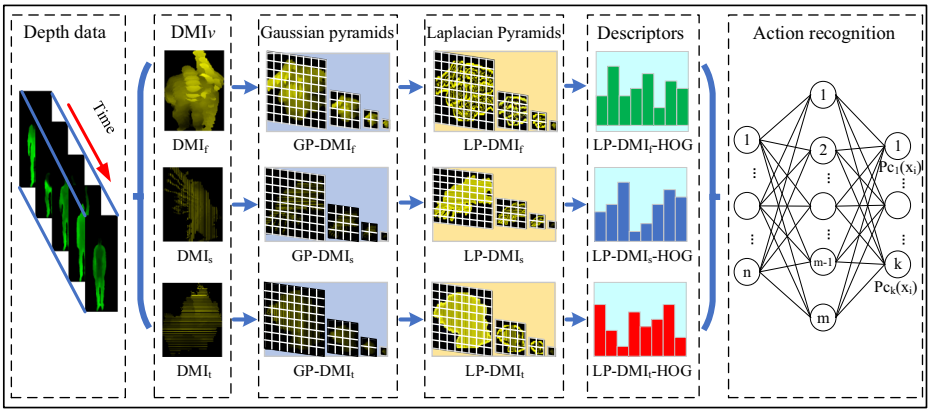

Fig. 1 The framework of our proposed human action recognition method

we employ ELM to classify actions. Specifically, the main contributions of this article are summarized as follows:

1) We study a compact multi-scale feature map based on depth video sequences called LP-DMI. Due to its superiority of enhancing multi-scale dynamic information of actions, the proposed feature map outperforms other existing maps. Moreover, some redundant static information inside the body is excluded simultaneously.
2) We introduce a feature extraction scheme according to the hierarchical structure of Laplacian pyramids. We extract HOG features and cascade them as LP-DMI-HOG. This descriptor captures multi-granularity features therefore it is more discriminative than others.
3) We propose a multi-scale human action recognition framework in which we generate compact multi-scale feature maps through the Laplacian pyramid of three-view DMI and then extract multi-granularity features. In addition, we use extreme learning machine for action classification.
4) We conduct experiments on the public MSRAction3D, UTD-MHAD and DHA dataset, and the experimental results demonstrate that our method surpasses the state-of-the-art benchmarks.

The rest of this article is organized as follows. Section 2 reviews the previous work related to ours. In Section 3, the proposed method is presented in detail, including building Laplacian pyramids of DMI, extracting LP-DMI-HOG feature and action classification. Section 4 discusses the experimental results compared to other human action recognition methods. At last, the conclusions of this paper are drawn in Section 5.

2 Related work

According to the type of input data, human action recognition technologies consist of RGB video based methods [10, 49, 52], depth video based methods [13, 25, 59, 62], 3D skeleton based methods [9, 38, 51], and multi-modal data fusion based methods [7, 16, 57]. Due to the convenience of data acquisition and invariance to illumination and texture changes, many researchers focus on the second methods which generally contain three steps: computing depth feature maps from depth video sequences, generating feature descriptors for motion representaion and recognizing actions by classifiers or neural networks [47, 55]. For higher accuracy, tremendous effort has been made to investigate representation and feature extraction strategy for human action recognition. Bobick and Davis [3] introduced a view-based approach on the basis of a temporal template that contains two component versions: the presence and recency of motion in sequence. They computed motion energy images (MEI) and motion history images (MHI) to model spatial and temporal characteristics of human actions. Mohammad et al. [4] utilized the static history images (SHI) as the complementary components of MHI. Motivated by MHI and MEI, Yang et al. [59] projected each depth frame onto three orthogonal Cartesian planes, then the subtraction operations between successive projections were carried out to obtain depth motion maps (DMM). On the contrary to DMM, Kamel et al. [18] investigated the depth motion images (DMI) in which the pixel value is the minimum value of the position of the same pixels over time to describe the overall action appearance from the front view. Since the DMM fails to recognize two actions with reverse temporal orders, Elmadany et al. [13] divided the depth video sequences into multiple partitions with the equal number of frames. Then they constructed the

hierarchical pyramid depth motion maps (HP-DMM) so as to capture more detailed information of human movements.

Based on the depth feature maps above, many descriptors have been studied for human action recognition. The histogram of oriented gradients (HOG) [26], the local binary pattern (LBP) [8], and other shape and texture features [11] were calculated from DMM for more accurate description. Oreifej and Liu [27] introduced the histogram of oriented 4D normals (HON4D) in order to describe the action in 4D space, including depth, spatial, and time coordinates. Li et al. [23] introduced Local Ternary Pattern (LTP) as an image filter for DMMs and applied CNN to classify corresponding LTP-encoded images. Tian et al. [35] employed Harris detector and local HOG descriptor on MHI for action recognition and detection. Furthermore, Gu et al. [14] selected ResNet-101 as the deep learning model and fed it with MHI. Aly et al. [2] calculated global and local features using Zernike moments with different polynomial orders to represent global and local motion patterns respectively. Kamel et al. [18] presented a feature fusion method for human action recognition from DMI and moving joints descriptor (MJD) data using convolutional neural networks (CNN). Mohammad et al. [4] extracted the gradient local auto-correlations (GLAC) features from the MHI along with SHI to represent the movements. Chen et al. [6] computed GLAC features based on DMM and put them into the extreme learning machine for activity recognition. Space time occupancy patterns (STOP) was proposed by Vieira et al. [40] in which space and temporal axes were divided into several partitions for each sequence. Besides, the bag of angles (BoA) applied to skeleton sequences and the other descriptor called Hierarchical pyramid DMM deep convolutional neural network (HP-DMM-CNN) for depth videos were presented in [13].

In addition, some new methods have emerged in the latest work. Sun et al. [32] presented a global and local histogram representation model using the joint displacement between the current frame and the first frame, and the joint displacement between pairwise fixed-skip frames, respectively. Ahmad et al. [61] fed feature maps into the CNN architecture rather than using any conventional method, and ulteriorly Trelinski et al. [37] computed concatenated handcrafted and action-specific CNN-based descriptors together to obtain action feature vectors. Li et al. [21] generated 3D body mask and then formed the depth spatial-temporal maps (DSTMs) which provided compact global spatial and temporal information of human motions. Wei et al. [51] modeled human actions with a hierarchical graph in which the depth video sequence was represented as sequential atomic actions. Every atomic action was denoted as a composite latent state consisted by a latent semantic attribute and a latent geometric attribute. However, the methods above fail to capture the multi-scale features for action recognition, and thus have poor robustness. Recently, more attention has been paid to multi-scale motion information. Ji et al. [17] embedded the skeleton information into depth feature maps to divide the human body into several parts. The surface normals of local motion part sequence were partitioned into different space-time cells to obtain local spatio-temporal scaled pyramid which was applied to extract local feature representation. Yao et al. [60] studied parallel pair discriminant correlation analysis (PPDCA) to fuse the multi-scale temporal information with a lower dimension. However, the multi-scale temporal information in this method means features related to different numbers of frames. These methods obtain multi-scale information by different number of frames and cells or various sampling rate, which is only the scale change in the temporal level in essence. In this paper, we present a multi-scale method based on the Scale-space theory in [1]. Note that rather than realize multi-temporal scale, we focus on spatial multi-scale of feature maps to tackle the problem of complex model representation and low implementation efficiency.

3 Proposed method for human action recognition

A typical action contains characteristic information in different scales, and it can be represented by the structured multi-scale features. Learning the information in single spatial scale is deficient to provide discriminative feature sufficiently for human action recognition. In order to increase the scale diversity, we propose a novel method to represent actions by multi-scale feature map LP-DMI and extract multi-granularity feature with hierarchical pyramid structure. Then, extreme learning machine is utilized to recognize human actions.

3.1 Calculation of depth motion images

With the advent of depth cameras, a lot of approaches have been introduced based on the depth videos for human action recognition. Each frame of the depth camera records a snapshot of the action at a certain point in time. In general, DMI is considered as an effective representation of depth video sequences. It captures not only the overall appearance of actions but also the dense range changes in the moving parts. In this paper, we project the frames obtained by the depth camera onto three orthogonal Cartesian coordinate planes, thus each 3D depth frame generates three 2D maps. We record them as $\text{map}_v(v \in \{f, s, t\})$ corresponding to the front, side, and top view respectively. The pixel value of DMI is the minimum value of the same spatial position of the depth maps. The three-view DMI of a depth video sequence with N frames can be calculated by the following equation.

$$DMI_v(i, j) = 255 - \min\left(\text{map}_v(i, j, t)\right), \quad \forall t \in [k, \ldots, (k + N - 1)] \tag{1}$$

where $\text{map}_v(i, j, t)$ is the pixel value of (i, j) position of 2D map at time t from the perspective of v. k represents the index of the frame. The maps are processed by dividing each pixel value by the maximum value of all the pixels contained in the image for normalization. We crop the region of interest (ROI) in DMI to exclude excess black pixels. This normalization contributes to eliminating intra-class differences and reducing the nuisances caused by body shape and motion amplitude. The generative process of DMI is depicted in Fig. 2.

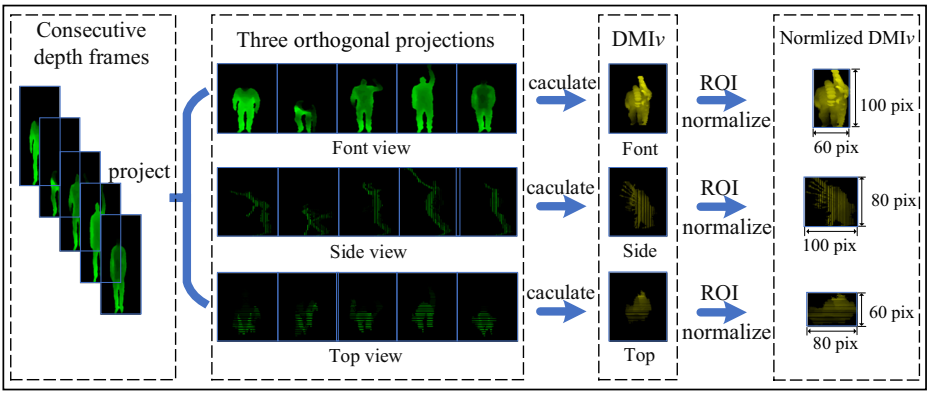

Fig. 2 The process of calculating DMI_v from depth video sequences

3.2 Multi-scale represtation of depth video sequences

However, DMI simply reflects spatial information of actions in single scale. In order to capture multi-scale changes of human motions, we adopt the Gaussian pyramid transform which has been demonstrated the practicability in increasing scale diversity [20, 31]. As shown in Fig. 3, we acquire a cluster of multi-scale feature maps shaped like several pyramids. We stipulate that the number of layers goes up in a bottom-up manner. G_l is used to represent the image of l_{th} layer of a Gaussian pyramid, that is to say, the size of G_{l+1} is smaller than that of the G_l. We need to perform Gaussian kernel convolution and downsampling on the G_l to produce G_{l+1}. Mathematically, the gray value corresponding to the (i, j) position of G_l can be formulated as:

$$G_l(i,j) = \sum_{m=-c}^{c} \sum_{n=-c}^{c} \varpi(m,n) \otimes G_{l-1}(2i+m, 2j+n), \\ (1 \le l \le L, 0 \le i \le R_l, 0 \le j \le C_l) \tag{2}$$

where $\otimes$ is a convolution operator and L is the total number of layers in every Gaussian pyramid. (m, n) is the position of the convolution kernel. R_l and C_l are the number of rows and columns relative to the l_{th} layer image of the Gaussian pyramid. c determines the size of ϖ and ϖ is a Gaussian window of size $(2c+1) \times (2c+1)$ satisfying the following formula:

$$\varpi(m,n) = \frac{1}{2\pi\sigma^2} e^{-(m^2+n^2)/2\sigma^2} \tag{3}$$

where σ is the standard deviation of the normal distribution. It refers to the variance related to the Gaussian filter which reflects the degree to which the image is blurred. We regard DMI as the lowest layer of the Gaussian pyramid denoted as G_1. Then, a set of images $\{G_1, G_2, \ldots, G_L\}$ in which G_{l+1} is $1/c^2$ size of G_l can be generated by (2), and constitutes

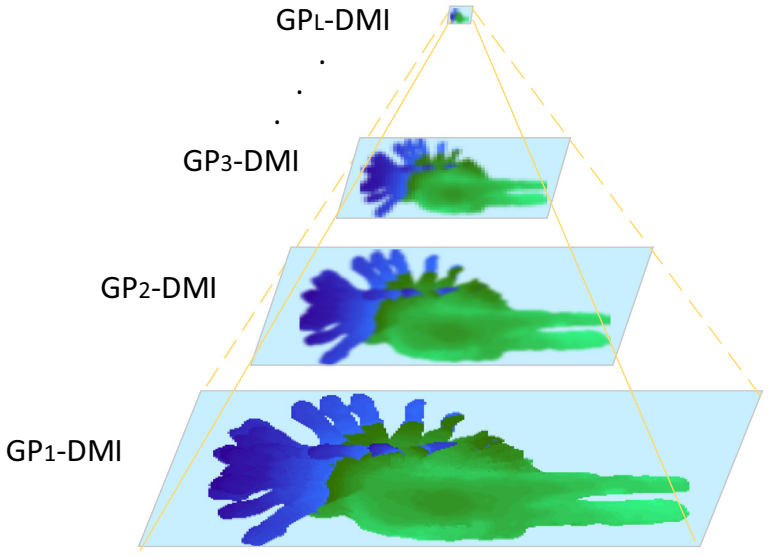

Fig. 3 The hierarchical structure of GP-DMI

an L-layer Gaussian pyramid. Thus, a series of Gaussian pyramids represented as GP_L-DMI are simply calculated by this iterative scheme. In this paper, we set c to 2 and utilize a 5×5 Gaussian kernel as (4). The pyramid algorithm reduces the filter band limit between layers by an octave, and chops the sampling interval by the same factor. The frequency of downsampling operations is related to the size of the original image. For the Gaussian pyramid based on an $M \times N$ image, the maximum number of layers is $\lfloor \log_2 \min\{M, N\} \rfloor$.

$$\varpi = \begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix} \tag{4}$$

Influenced by the complexity and concurrency of human behaviors, a simple action may involve the movement of multiple body parts. We view the inherent characteristic inside the body as static information, while the contour information that can better describe the changing of movements as dynamic information. For the majority of actions, the static information inside the human body is highly similar. Take waving arms in different directions for instance, the information of the abdomen and legs are constant to some extent, and cannot provide the discriminative feature for recognition very well. On the contrary, the dynamic information of different body parts can better reflect the spatial changes of actions in the interval, thus reflecting the specific feature of certain action. Inspired by this, we motivate to obtain the multi-scale dynamic information for human action recognition. We interpolate the l_{th} layer of the Gaussian pyramid, that is, insert 0 in even rows and columns. Then, we utilize Gaussian filter to get G_l^* which has the equal size as the image one layer below it. We caculate the difference between G_l and G_l^* to get the multi-scale dynamic infomation. At the same time, this operation removes a lot of reduntant static information, making LP-DMI more compact than GP-DMI. As in the Gaussian pyramid, we set c to 2. Mathematically:

$$G_l^*(i, j) = 4 \sum_{m=-2}^{2} \sum_{n=-2}^{2} \varpi(m, n) \otimes G_l\left(\frac{i+m}{2}, \frac{j+n}{2}\right), \quad (1 \le l \le L, 0 \le i \le R_l, 0 \le j \le C_l) \tag{5}$$

and

$$G_l\left(\frac{i+m}{2}, \frac{j+n}{2}\right) = \begin{cases} G_l\left(\frac{i+m}{2}, \frac{j+n}{2}\right), & \text{if } \frac{i+m}{2}, \frac{j+n}{2} \in \mathbb{N}^+ \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Therefore, the Laplacian pyramid can be calculated as follows.

$$\begin{cases} LP_l = G_l - G_{l+1}^*, & 1 \le l < L \\ LP_L = G_L, & l = L \end{cases} \tag{7}$$

where LP_l is the l_{th} layer of the Laplacian pyramid. Considering the integrity of motion infomation, we directly take the top layer of Gaussian pyramids as that of the Laplacian pyramid. Consequently, they have equal number of layers. Specifically, each depth frame produces three depth feature maps according to three views, thereby, it has three generated Laplacian pyramids. As shown in Fig. 4, the Laplacian pyramids cut down a large amount of static information inside the body meanwhile strengthen the dynamic information of body boundaries, which is more conducive to extracting discriminative features. In Sec. 4, we will further evaluate the proposed multi-scale feature map LP-DMI.

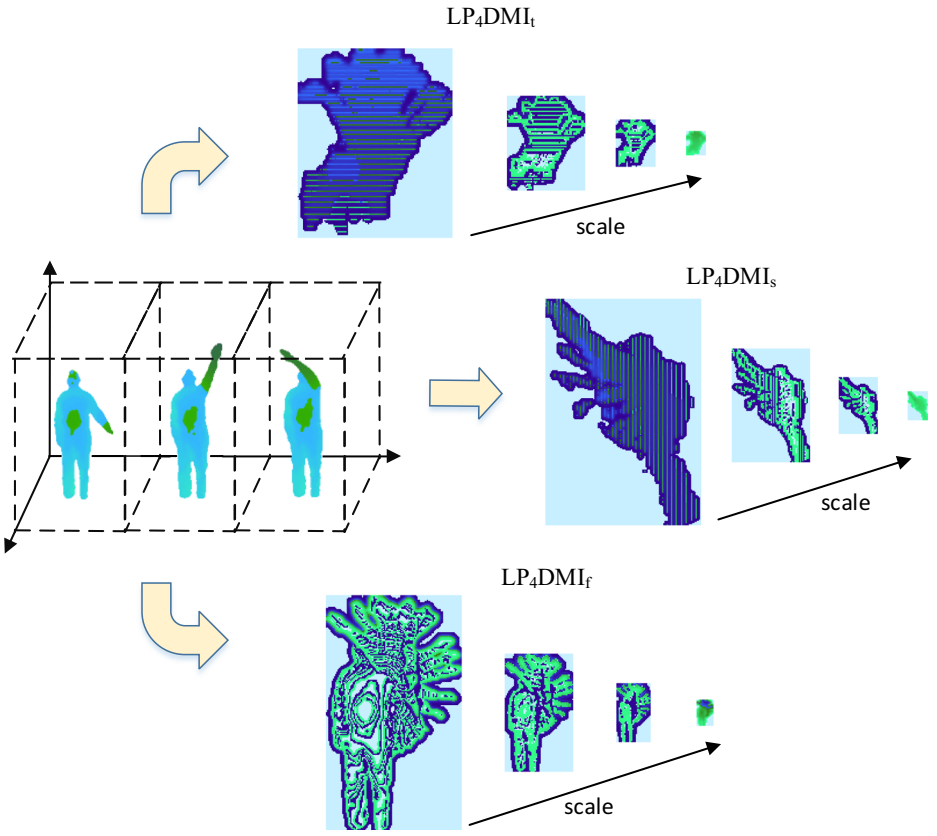

Fig. 4 An example of a four-layer LP_4-DMI with three angles

3.3 Feature extraction with hierarchical pyramid structure

There are several reasonable options for determining which feature to extract [42, 45, 46]. In this paper, we utilize HOG descriptors to extract the local features of LP-DMI denoted as LP-DMI-HOG. HOG feature is sensitive to the distribution of gradient and edge information, thus it characterizes gradient changes especially the shape of objects pretty well. The basic idea is to compute gradient orientation histograms on a dense grid of uniformly spaced cells and perform local contrast normalization [59]. Before extracting features, we copy adjacent pixels to normalizing the feature maps from the same view to the same size. The interpolated pixel values are the same as the neighboring pixels, so they will not interfere with the multi-scale information and we can compute multi-granularity motion features effectively. Moreover, this step is beneficial to solve the problem of too small pictures caused by incremental layers. We cascade the HOG feature extracted from LP-DMI in the same layer to obtain the three-view features at the same scale. Then we derive LP-DMI-HOG from coarse-grained to fine-grained as the layer increases. We normalize the resulting feature vectors using min-max scaling, and the principal component analysis (PCA) is applied to reduce the dimension for the sake of computational efficiency.

We normalize the depth feature maps projected onto the same planes to a uniform size, and the specific parameter settings are shown in Fig. 2. We set the size of each cell to 10×10 pixels and the number of gradient orientation bins is 9. The size of block is 2×2. Furthermore, the step is 10 pixels. The remained principal components of MSRAction3D, UTD-MHAD, and DHA is 550, 860, and 450. So that, each action sample is a total of 15444 and 20592 dimensions when the number of layers is 3 and 4 respectively. Note that, we consider this as the default setting of feature extraction. Then, the resulting feature will be fed into ELM for action classification.

3.4 Action recognition by extreme learning machine

In this work, we employ extreme learning machine (ELM) for action classification which was proposed by Huang et al. for training single-hidden layer feed-forward neural networks (SLFNs) [63]. The weight between the input layer and hidden layer can be initialized randomly as well as the bias of the hidden nodes. Therefore, the ELM just calculates the weight matrix between the hidden layer and output layer without the need to tune parameters. The matrix can be figured out by finding the generalized inverse matrix, thus the extreme learning machine has distinct advantages in parameter selection and computational efficiency. That is why we use extreme learning machine for action recognition. Given a training set with n samples and m classes $D = \{(x_i, y_i) | x_i \in R^n, y_i \in R^m, i = 1, 2, \ldots, n\}$, the SLFNs with N hidden nodes can be expressed as:

$$f(x_i) = \sum_{j=1}^{N} \beta_j g\left(w_j \cdot x_i + b_j\right) = o_i, i = 1, 2, \ldots, N \tag{8}$$

where $w_j = (w_{j1}, w_{j2}, ..., w_{jd})^T$ is the weight vector connecting the j_{th} hidden node with the input nodes. $\beta_j = (\beta_{j1}, \beta_{j2}, ..., \beta_{jm})^T$ is the weight vector connecting the j_{th} hidden node with the output nodes. b_j represents the threshold of the j_{th} hidden neuron, and $g(x)$ denotes the activation function. Note that w_j and b_j are assigned randomly. The goal of ELM is to minimize the training error as far as possible, which can be depicted as $\sum_{i=1}^{N} \|o_i - y_i\| = 0$. Therefore, parameters $\beta_j = (\beta_{j1}, \beta_{j2}, ..., \beta_{jm})^T$ can be estimated by least-square tting with the given training data D. In other words, the problem can be written as the following equation.

$$Y = H\beta \tag{9}$$

with

$$H = \begin{pmatrix} g(w_1 \cdot x_1 + b_1) & \cdots & g(w_m \cdot x_1 + b_m) \\ \vdots & \ddots & \vdots \\ g(w_1 \cdot x_n + b_1) & \cdots & g(w_m \cdot x_n + b_m) \end{pmatrix} \tag{10}$$

$$\beta = \left(\beta_1^T, \beta_2^T, \ldots, \beta_m^T\right)^T,$$

$$Y = \left(y_1^T, y_2^T, \ldots, y_n^T\right)^T$$

H is the hidden layer output matrix of the network, in which the j_{th} column is the j_{th} hidden nodes output vector concerning inputs $(x_1, x_2, \ldots, x_m)$. The i_{th} row of H is the output vector of the hidden layer about input x_i. Once the input weight w_j and the hidden layer bias b_j are determined, the output matrix H of the hidden layer is unique. The number of

hidden nodes is usually much smaller than that of training samples. In this case, the smallest norm least-squares solution of (9) is equivalent to solving the following equation.

$$\hat{\beta} = H^{\dagger} Y \tag{11}$$

where $H^{\dagger}$ is the Moore-Penrose generalized inverse of matrix H [33].

4 Experiment results and analysis

In order to evaluate the effectiveness of the proposed framework, we conduct experiments on the public MSRAction3D [18], UTD-MHAD [7], and DHA dataset [4]. In Fig. 5, the depth video sequence of pickup and throw is shown as an example of action samples. We investigate how many layers are sufficient to capture multi-scale features for action recognition and compare several ways of extracting local features. In this section, we present the results of the ablation experiment, optimizing and confirming the effectiveness of the multi-scale mechanism in the proposed framework. Meanwhile, we show the advantages of our proposal over other state-of-the-art methods.

4.1 Datasets and experimental settings

4.1.1 Datasets description

The MSRAction3D is a dataset for action recognition which contains 557 depth video sequences and 557 skeleton sequences for 20 actions captured by Kinect sensor. The actions including high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, and pickup and throw are taken by 10 subjects. Every action is repeated by all the subjects two or three times.

The UTD-MHAD includes 861 samples of 8 subjects. There are 27 actions in total, and every subject performed each action 4 times. The actions are: right arm swipe to the left, right arm swipe to the right, right hand wave, two hand front clap, right arm throw, cross arms in the chest, basketball shoot, right hand draw x, right hand draw circle (clockwise), right hand draw circle (counter clockwise), draw triangle, bowling, front boxing, baseball swing from right, tennis right hand forehand swing, arm curl, tennis serve, two hand push, right hand knock on door, right hand catch an object, right hand pick up and throw, jogging in place, walking in place, sit to stand, stand to sit, forward lunge, and squat.

Fig. 5 The depth video sequence of pickup and throw in MSRAction3D dataset

The DHA database is orgnized with 483 depth video sequences for 23 actions. Each sample video was performed by 2 or 3 times by 21 subjects (12 males and 9 females). The list of action classes are: bend, jack, jump, pjump, run, side, skip, walk, one-hand-wave, two-hand-wave, front-clap, side-clap, arm-swing, arm-curl, leg-kick, leg-curl, rod-swing, golf-swing, front-box, side-box, tai-chi, pitch, and kick.

4.1.2 Experimental setups

We conduct experiments with the following experimental settings.

Setup 1: Cross-subject. In order to have fair experimental results, we perform the cross-subject tests on the three benchmark datasets according to the experimental settings of [18, 28]. More precisely, we use odd subjects for training, whereas even subjects are applied for testing.

Setup 2: Subset partition. We divide the MSRAction3D dataset into three subsets as shown in Table 1, and three different tests are conducted on these subsets following the settings as [4]. In test 1, 1/3 action samples in each subset are employed as the training set, and the remaining samples are used for validation. On the contrary, test 2 uses 2/3 samples for training, and the rest samples are taken in the testing set. Test 3 has a cross-subject test on each subset abide by setup 1, that is to say, the action samples corresponding to the odd subjects in each subset are used for training and the rest for testing.

Setup 3: K-Fold cross-validation. In order to further prove the scientific nature of multi-scale feature maps, we carried out k-fold cross-validation (KFCV) experiments. In this setting, every dataset is divided into ten portions in which the nine pieces are combined as the training set, and the remaining parts are used as the testing set. The above process is repeated for ten times testing all the parts one by one, and then the average score is taken as the final recognition accuracy. Furthermore, in each fraction we keep the categories ratios same as the original data.

4.2 Ablation study

4.2.1 Influence of layer parameter

To exploit the optimal muti-scale feature map of different datasets, we construct LP-DMI with different layers in a step-wise manner and perform experiments according to setup 1

Table 1 Three subsets of the MSRAction3D dataset

Label	AS1	Label	AS2	Label	AS3
2	Horizontal arm wave	1	High arm wave	6	High throw
3	Hammer	4	Hand catch	4	Forward kick
5	Forward punch	7	Draw x	15	Side kick
6	High throw	8	Draw tick	16	Jogging
10	Hand clap	9	Draw circle	17	Tennis swing
13	Bend	11	Two hand wave	18	Tennis serve
18	Tennis serve	14	Forward kick	19	Golf swing
20	Pickup and throw	12	Side boxing	20	Pickup and throw

on three datasets. The experimental results with respect to the GP-DMI and LP-DMI from 2 to 6 layers are presented in Fig. 6. The first thing we noticed is that the motion feature will be too coarse-grained ro recognize similar actions if the number of layers is inadequate. Otherwise, if the number of layers is superfluous, the static information may be more redundant, which leads to low efficiency and accuracy. In addition, the experimental results illustrate that LP-DMI yeilds better recognition accuracy on the whole, which achieves the highest recognition rate of 93.41% when the number of layers is 4 on MSRAction3D dataset. The LP_3-DMI on UTD-MHAD and DHA dataset are the optimum, and the recognition rates are respectively 85.12% and 91.94%. We will abide by the optimal layer setting obtained here in subsequent experiments.

4.2.2 Evaluation of different feature extraction strategies

After that, we compare several strateges of feature extraction and normalization following setup 1. The default feature extraction setting in Sec. 3 was not adopted in this experiment but a combination of two dynamic constraints in order to prevent the feature map and the cell of HOG from being too small. In details, $D^l{}_v(w, h, d)$ is the normalization parameter denoting that the size of LP_l-DMI_f, LP_l-DMI_s, LP_l-DMI_t is $w \times h$, $h \times d$, $w \times d$. Constraint N_1: $D^l{}_v(w/2^{l-1}, h/2^{l-1}, d/2^{l-1})$. Constraint N_2: $D^l{}_v(w, h, d) = D^l{}_v(160, 320, 240)$. Constraint C_1: the size of cell is 20×20. Constraint C_2:the size of cell is $20/2^{l-1} \times 20/2^{l-1}$. These parameters determine various scale of the feature map and the granularity of descriptor, and we report experimental results in Tables 2 and 3. We observe that N_2 combined with

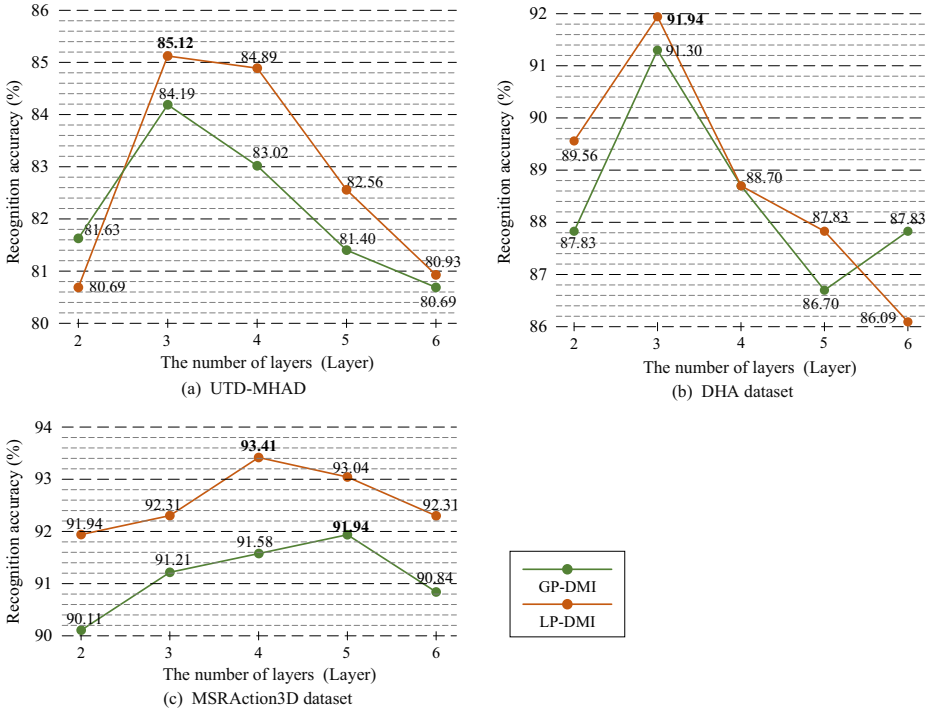

Fig. 6 The recognition accuracy of LP-DMI with different layers

Table 2 The results of various normalization strategies on MSRAction3D dataset

Constraint	N_1			N_2		
	Accuracy(%)	Dimension	Time(s)	Accuracy(%)	Dimension	Time(s)
C_1	89.01	15444	91.25	**91.58**	49968	254.88
C_2	88.64	49968	205.34	87.55	1234512	3151.40

C_1 outperforms other strateges. In other words, the applied normalization method achieve the effect of improving the classification accuracy. We show the recognition accuracy and average computation time of LP-DMI-HOG descriptor and VGG-16 [46] in Table 4, confirming that LP-DMI-HOG is more efficient and more discriminative. Considering the tradeoff between precision and efficiency, we chose HOG descriptor to extract motion features.

4.2.3 Effectiveness of multi-scale feature map LP-DMI

We evaluate the effectiveness of LP-DMI from two aspects. On the one hand, we have proved that LP-DMI is a more discriminative multi-scale feature map compared with GP-DMI. On the other hand, we will certify that the LP-DMI-HOG extracted from LP-DMI excels HOG features based on other feature maps. For the fairness of the results, we follow default feature extraction strategy on these depth maps to obtain HOG descripton, and employ ELM for action recognition. In terms of MSRAction3D dataset, we conduct the experiments following setup 2. The alone and average results with regard to AS1, AS2, and AS3 are presented in Table 5, and the highest rate of each subset has been shown in bold. As can be seen, LP-DMI achieves the highest average recognition rate in three different tests and outperforms than other feature maps. Specifically, in test one, LP-DMI achieved 90.42% accuracy on the three subsets. In addition to the performance on AS2 which is slightly lower than DMM, LP-DMI has an absolute advantage on other two subsets. In the second test, our proposal exceeds others significantly and gets the best recognition rate of 98.63% on AS2. Furthermore, the ELM trained by LP-DMI-HOG even can completely label all the testing samples on AS3. Therefore, in spite of the recognition rate of DMM and HP-DMM on AS1 equals to our method, the average recognition rate we have achieved is still 5% higher than them. In test three, LP-DMI obtains an average recognition rate of 94.59%. The result of LP-DMI on AS1 is 0.95% mildly lower than that of the DMM, but the recognition rates on other subsets are optimal. Overall, LP-DMI surpasses MEI, MHI and GP_5-DMI in all tests. Although DMM, HP-DMM, and DMI on individual subsets are superior to LP-DMI, the average recognition rate of our method is the highest. It should be noted that we almost improved accuracy by 4% in three tests by constructing Laplacian Pyramid pyramid for

Table 3 The results of various feature extraction strategies on UTD-MHAD

Constraint	N_1			N_2		
	Accuracy(%)	Dimension	Time(s)	Accuracy(%)	Dimension	Time(s)
C_1	79.53	15444	120.36	**87.55**	37476	255.38
C_2	77.91	37476	205.34	84.91	296676	272.25

Table 4 The comparsion of different descriptors

Datasets	VGG		HOG	
	Accuracy(%)	Time(s)	Accuracy(%)	Time(s)
MSR	91.94	26.76	**93.41**	0.23
UTD-MHAD	81.86	12.62	85.12	0.17
DHA	84.35	19.44	91.94	0.12

DMI, and this transformation process is very efficient and does not cause too much time consumption.

On UTD-MHAD and DHA dataset, we testify the proposed LP-DMI complying with setup 1, and describe the result in Table 6. LP_3-DMI yeilds the best recognition accuracy of 85.12% on UTD-MHAD. Once more, the experiments of DHA dataset validate our methods in which LP_3-DMI produce the result of 91.94%. For elaborating the performance of our method clearly, the confusion matrix computed from three datasets is depicted in Fig. 7. It can be seen that our method can correctly recognize the majority of actions. After analyzing the accuracy of specific classes, we find that the errors mainly occur in the classification of similar actions. For example, skip and jump, front-box and arm-curl, draw x and draw tick. In a word, this experimrnt further confirms that LP-DMI is a compact multi-scale feature map, and the proposed LP-DMI-HOG descriptor is promising.

In order to further prove the scientific nature of multi-scale action representation, we conduct a k-fold cross-validation experiment additionally complying with setup 3. Figure 8a shows the recognition accuracy of different feature maps corresponding to three datasets, and Fig. 8b depicts the part of LP-DMI that is higher than others. For MSRAction3D dataset, LP-DMI achieves the highest recognition rate of 98.48% with a little difference of 0.43% to GP-DMI. It should be noted that both of them are higher than their template feature map DMI by more than 3%. Compared with the single scale feature map, LP-DMI can improve the recognition accuracy by up to 8.27%. Besides, the experimental results of UTD-MHAD prove the advantages of LP-DMI as well, which are 4.57% and 3.8% higher than MEI and DMM, respectively. The scores of HP-DMM, DMI and GP-DMI are close, which are 0.61%

Table 5 The comparation of other feature maps on MSRAction3D dataset(%)

		MEI	MHI	DMM	HP-DMM	DMI	GP-DMI	LP-DMI
Test One	AS1	75.34	67.81	92.47	89.73	89.04	93.84	**95.21**
	AS2	71.05	69.74	84.21	80.92	**86.84**	86.18	86.18
	AS3	71.62	70.95	86.49	84.46	85.81	86.49	**89.86**
	Average	72.67	69.50	87.72	85.04	87.23	88.84	**90.42**
Test Two	AS1	89.04	84.93	**97.26**	**97.26**	93.42	93.42	**97.26**
	AS2	82.89	82.89	86.84	88.16	93.42	93.42	**98.63**
	AS3	90.54	93.24	97.30	94.59	97.30	97.30	**100.00**
	Average	87.49	87.02	93.80	93.34	94.71	94.71	**98.63**
Test Three	AS1	71.43	72.38	**99.05**	91.43	94.29	97.14	98.10
	AS2	69.64	66.07	85.71	84.82	82.14	89.29	**90.18**
	AS3	74.78	70.27	94.59	92.79	93.69	92.79	**95.50**
	Average	71.95	69.57	93.12	89.68	90.04	93.07	**94.59**

Table 6 The recognition rate of depth feature maps on UTD-MHAD and DHA dataset(%)

	MEI	MHI	DMM	HP-DMM	DMI	GP-DMI	LP-DMI
UTD-MHAD	51.63	56.52	74.19	77.67	76.28	84.19	**85.12**
DHA	62.60	79.10	90.4	91.30	90.80	91.30	**91.94**

lower than LP-DMI on average. LP-DMI also achieves optimistic results 92.39% on the DHA dataset, which is better than GP-DMI by 1.81% and even 9.49% higher than MEI. Compared with common DMI, the accuracy of LP-DMI is markedly improved by 2.18%. In general, the multi-scale feature maps LP-DMI and GP-DMI are significantly superior to other single-scale feature maps. The results further prove that increasing scale diversity can enhance the discriminativeness of motion features, thus achieving higher recognition accuracy.

4.3 Comparisons to other state-of-the-art approaches

In this experiment, we follow setup 1 which is same as the baseline methods for persuasion. The cross-subject test is challenging due to variations in the same actions performed by different subjects, but our method can still achieve high accuracy. In Table 7, our method obtains the promising accuracy of 93.4% compared with other solutions utilizing single

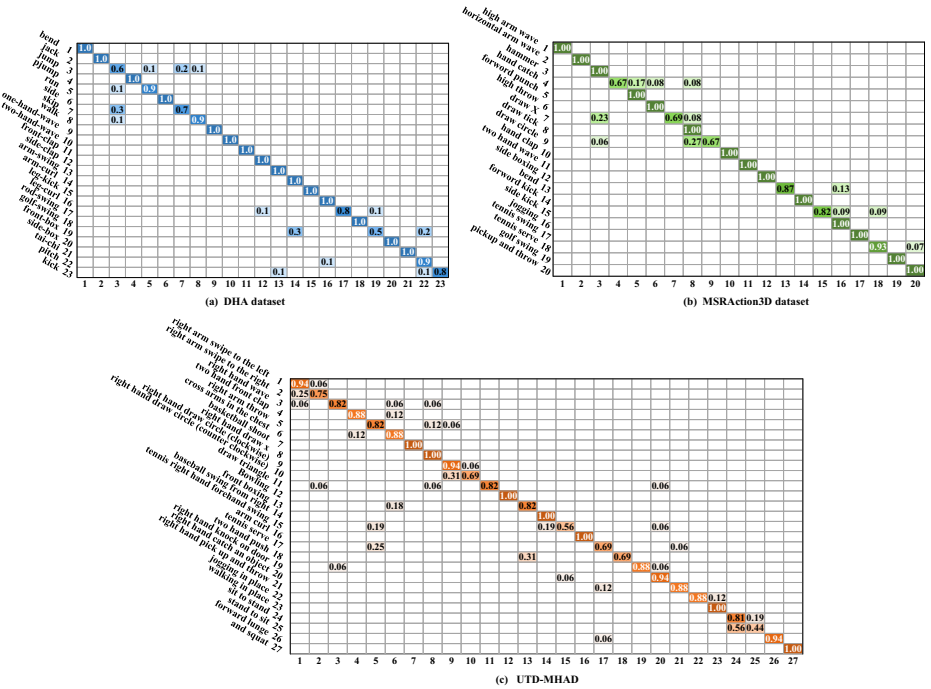

Fig. 7 The confusion matrix of three datasets

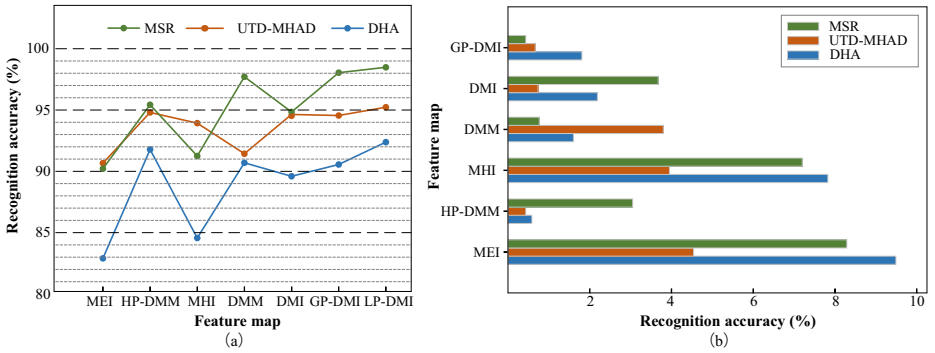

Fig. 8 K-Fold cross-validation results of three datasets

depth modality data on MSRAction3D dataset, and it is 4.5% higher than DMM-GLAC which extracts local feature descriptor from depth feature maps as well. The HP-DMM-CNN, 3D-CNN as well as the method in [61] using convolutional neural networks are 1.1%, 7.3% , 6.3% lower than our method. It should be noted that the method proposed by Ji et al

Table 7 Comparison of our method with baseline methods on MSRAction3D dataset

Methods	Modality	Accuracy(%)
Pose Set [50]	D	90.0
DMM-HOG [59]	D	88.7
DMM-GLAC [6]	D	88.9
HON4D [27]	D	88.9
Skeletons Lie group [38]	S	89.5
HOG3D+LLC [30]	D	90.9
3D-CNN [58]	D	86.1
DSTIP [53]	D	89.3
STOP [40]	D	87.5
HP-DMM-CNN [13]	D	92.3
PointLSTM-late [24]	D+S+RGB	95.4
MMHCCCA [13]	D+S	93.5
Ji et al. [17]	D+S	90.8
PointLSTM-late [24]	D+S	95.4
Trelinski et al. [37]	D	90.6
Ahmad et al. [61]	D	87.1
Wei et al. [51]	S	87.2
Xin et al et al. [15]	D+S	91.6
LP-DMI	D	93.4
LP-DMI+HP-DMM	D	94.9
LP-DMI+MJD	D+S	95.6
LP-DMI+GCN	D+S	94.5

Table 8 Comparison of our method with baseline methods on UTD-MHAD

Methods	Modalities	Accuracy(%)
DMM-HOG [59]	D	81.5
3DHOT-MBC [62]	D	84.4
Hierarchical Gaussian [25]	D	84.1
HP-DMM-CNN [13]	D	82.8
HP-DMM-HOG [56]	D	73.7
MLSL [57]	D+S	88.4
Kamel et al. [18]	D+S	88.1
Chen et al. [7]	D+S+RGB	79.1
STSDDI. [16]	D+S+RGB	91.2
LP-DMI	D	85.1
LP-DMI+HP-DMM	D	89.1
LP-DMI+MJD	D+S	90.5
LP-DMI+GCN	D+S	94.2

[17] is 2.6% lower than ours although they obtain local spatio-temporal scaled pyramid and embed skeleton information. Furthermore, we fuse LP-DMI-HOG descriptor with HP-DMM-HOG, MJD-HOG, GCN [29] by canonical correlation analysis(CCA) [13], which yeilds the recognition rate of 94.9%, 95.6%, 94.5%.

We also demonstrate the generality of our framework on UTD-MHAD and report the results in Table 8. Our methods obtains the recognition accuracy of 85.1% which is 2.3% and 0.7% higher than HP-DMM-CNN and 3DHOT-MBC. The method proposed by Nguyen et al. which employs hierarchical gaussian descriptor is 1% lower than us. In addition, our approach surpasses other methods employing HOG descriptor. For example, LP-DMI is 3.6% higher than DMM-HOG and 11.4% higher than HP-DMM-HOG. With same evaluation strategy, we compare our system with depth-based and multi-modal feature fusion methods as well. The above experiments prove that our method is superior to other deep video based approaches and is able to achieve better performance through fusion techniques.

5 Conclusion

In this paper, we proposed a novel method based on the Laplacian pyramid considering multi-scale information for human action recognition. We calculated LP-DMI to increase the scale diversity of depth motion images in order to capture the multi-scale motion features and strengthen more favorable dynamic information. The experimental results have demonstrated that LP-DMI is more compact and discriminative than existing feature maps. Furthermore, the extracted LP-DMI-HOG which contains multi-granularity features has effectively improved the accuracy of action recognition. The experiments results conducted on MSRAction3D, UTD-MHAD and DHA dataset outperform the baseline methods. However, our method is still flawed in identifying actions with similar motion trajectories.The future work will focus on fusing multimodal features and considering multi-scale temporal information to facilitate the recognition accuracy.

Acknowledgements This work is partly supported by the National Key Research and Development Program of China under grant no. 2018YFC0407905, and the fundamental research funds of China for central universities under grant no. B200202188.

References

1. Alpatov AV, Rybina N, Trynov DY, Vikhrov SP (2018) Scale-space theory application to investigate surface correlation properties. Mediterranean Conference on Embedded Computing (MECO), pp 1–3
2. Aly S, Sayed A (2019) Human action recognition using bag of global and local Zernike moment features. Multimed Tools Appl 78:24923–24953. https://doi.org/10.1007/s11042-019-7674-5
3. Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 23(3):257–267. https://doi.org/10.1109/34.910878
4. Bulbul MF, Islam S, Ali H (2019) 3D human action analysis and recognition through GLAC descriptor on 2D motion and static posture images. Multimed Tools Appl 78(15):21085–21111. https://doi.org/10.1007/s11042-019-7365-2
5. Burt P, Adelson E (1987) The laplacian pyramid as a compact image code. IEEE Trans Commun 31(4):532–540. https://doi.org/10.1109/TCOM.1983.1095851
6. Chen C, Hou Z, Zhang B, Jiang J, Yang Y (2015) Gradient local Auto-Correlations and extreme learning machine for Depth-Based activity recognition. Adv Vis Comput 9474:613–623. 978-3-319-27856-8
7. Chen C, Jafari R, Kehtarnavaz N (2015) UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. IEEE International Conference on Image Processing (ICIP). https://doi.org/10.1109/icip.2015.7350781
8. Chen C, Jafari R, Kehtarnavaz N (2015) Action recognition from depth sequences using depth motion Maps-Based local binary patterns. IEEE Winter Conf Appl Comput Vis:1092–1099
9. Cheng K, Zhang Y, He X, Chen W, Cheng J, Lu H (2020) Skeleton-Based Action recognition with shift graph convolutional network. IEEE/CVF conference on computer vision and pattern recognition (CVPR), Seattle, 180–189. https://doi.org/10.1109/CVPR42600.2020.00026
10. Crasto N, Weinzaepfel P, Alahari K, Schmid C (2019) MARS: Motion-Augmented RGB stream for action recognition. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 7874–7883
11. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pp 886–893
12. Dhiman C, Vishwakarma DK (2018) A review of state-of-the-art techniques for abnormal human activity recognition. Eng Appl Artif Intell 77:21–45
13. Elmadany NED, He Y, Guan L (2018) Information Fusion for Human Action Recognition via Biset/Multiset Globality Locality Preserving Canonical Correlation Analysis, in IEEE Transactions on Image Processing, 27(11):5275–5287. https://doi.org/10.1109/TIP.2018.2855438
14. Gu Y, Ye X, Sheng W (2018) Depth MHI Based Deep Learning Model for Human Action Recognition. 13th World Congress on Intelligent Control and Automation (WCICA), pp 395–400
15. Hou CX, Liang Z, Jiuzhen Yang T (2020) Integrally Cooperative Spatio-Temporal Feature Representation of Motion Joints for Action Recognition. Sensors (Basel, Switzerland). vol 20. https://doi.org/10.3390/s20185180
16. Hou Y, Wang S, Wang P, Gao Z, Li W (2018) Spatially and Temporally Structured Global to Local Aggregation of Dynamic Depth Information for Action Recognition. IEEE Access 6:2206–2219. https://doi.org/10.1109/ACCESS.2017.2782258
17. Ji X, Cheng J, Feng W, Tao D (2017) Skeleton embedded motion body partition for human action recognition using depth sequences. Signal Process 143:56–68. https://doi.org/10.1016/j.sigpro.2017.08.016
18. Kamel A, Sheng B, Yang P, Li P, Shen R, Feng DD (2019) Deep convolutional neural networks for human action recognition using depth maps and postures. IEEE Trans Syst Man Cybern Syst 49(9):1806–1819. https://doi.org/10.1109/TSMC.2018.2850149
19. Kim H, Kim GY, Kim JY (2019) Music recommendation system using human activity recognition from accelerometer data. IEEE Trans Consum Electron 65(3):349–358. https://doi.org/10.1109/TCE.2019.2924177
20. Li S, Hao Q, Kang X, Benediktsson JA (2018) Gaussian pyramid based multiscale feature fusion for hyperspectral image classification. Sel Top Appl Earth Observ Remote Sens 11(9):3312–3324. https://doi.org/10.1109/JSTARS.2018.2856741
21. Li X, Hou Z, Liang J et al (2020) Human action recognition based on 3D body mask and depth spatial-temporal maps. Multimedia Tools and Applications

22. Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3D points. IEEE Comput Soc Conf Comput Vis Pattern Recogn:9–14
23. Li Z, Zheng Z, Lin F et al (2019) Action recognition from depth sequence using depth motion maps-based local ternary patterns and CNN. Multimedia Tools Appl 78:9587–19601. https://doi.org/10.1109/WACV.2015.150
24. Min Y, Zhang Y, Xiujuan C, Xilin C (2020) An Efficient pointLSTM for Point Clouds Based Gesture Recognition. IEEE/CVF Conf Comput Vis Pattern Recogn:5761–5770
25. Nguyen X, Son M, Thanh A-I et al (2018) Action recognition in depth videos using hierarchical gaussian descriptor. Multimed Tools Appl 77(16):21617–21652
26. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987. https://doi.org/10.1109/tpami.2002.1017623
27. Oreifej O, Liu Z (2013) HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. IEEE Conf Comput Vis Pattern Recogn:716–723
28. Padilla-López JR, Chaaraoui AA, Flórez-Revuelta F (2014) A discussion on the validation tests employed to compare human action recognition methods using the msr action3d dataset. Computer Science
29. Peng W., Shi J, Zhao G. (2021) Spatial Temporal Graph Deconvolutional Network for Skeleton-based Human Action Recognition. IEEE Signal Processing Letters. https://doi.org/10.1109/LSP.2021.3049691
30. Rahmani H, Huynh DQ, Mahmood A, Ajmal M (2016) Discriminative human action classification using locality-constrained linear coding. Pattern Recogn Lett 72:62–71
31. Sujee R, Padmavathi S (2018) Pyramid-based Image Interpolation. International Conference on Computer Communication and Informatics (ICCCI), pp 1–5
32. Sun B, Kong D, Wang S, Wang L, Wang Y, Yin B (2019) Effective human action recognition using global and local offsets of skeleton joints. Multimed Tools Appl 78:6329–6353. https://doi.org/10.1007/s11042-018-6370-1
33. Tan Z, Xiao L, Chen S, Lv X (2020) Noise-Tolerant And Finite-Time convergent ZNN models for dynamic matrix Moore–Penrose inversion. IEEE Trans Indust Inf 16(3):1591–1601. https://doi.org/10.1109/TII.2019.2929055
34. Teng Y, Liu F, Wu R (2013) The research of image detail enhancement algorithm with laplacian pyramid. IEEE international conference on green computing and communications and IEEE internet of things and IEEE cyber Physical and Social Computing, pp 2205–2209
35. Tian Y, Cao L, Liu Z, Zhang Z (2012) Hierarchical filtered motion for action recognition in crowded videos. IEEE Trans Syst Man Cybern 42(3):313–323. https://doi.org/10.1109/TSMCC.2011.2149519
36. Tran DT, Yamazoe H, Lee JH (2020) Multi-scale affined-HOF and dimension selection for view-unconstrained action recognition. Appl Intell 50(4):1468–1486. https://doi.org/10.1007/s10489-019-01572-8
37. Trelinski J, Kwolek B (2019) Ensemble of classifiers using CNN and Hand-Crafted features for Depth-Based action recognition. Int Conf Artif Intell Soft Comput:91–103
38. Vemulapalli R, Arrate F, Chellappa R (2014) Human action recognition by representing 3D skeletons as points in a lie group. IEEE Conf Comput Vis Pattern Recogn:588–595
39. Vieira AW, Nascimento ER, Oliveira GL, Liu Z, Campos MF (2012) Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences. Iberoamerican Congress Pattern Recogn:252–259
40. Vieira AW, Nascimento ER, Oliveira GL, Liu Z, Campos MF (2014) On the improvement of human action recognition from depth map sequences using space-time occupancy patterns. Pattern Recogn Lett 36: 221–227
41. Vishwakarma DK, Kapoor R (2012) Simple and intelligent system to recognize the expression of speech-disabled person. 2012 4th International Conference on Intelligent Human Computer Interaction (IHCI), Kharagpur, pp 1–6
42. Vishwakarma DK, Kapoor R (2015) Integrated approach for human action recognition using edge spatial distribution, direction pixel and -transform. Adv Robot 29(23):1553–1562. https://doi.org/10.1080/01691864.2015.1061701
43. Vishwakarma DK, Kapoor R, Maheshwari R, Kapoor V, Raman S (2015) Recognition of abnormal human activity using the changes in orientation of silhouette in key frames. In: 2015 2nd International Conference on Computing for Sustainable Global Development. IEEE, pp 336–341
44. Vishwakarma DK, Kapoor R (2017) An efficient interpretation of hand gestures to control smart interactive television. Int J Comput Vis Robot 7(4):454–471
45. Wan GY, Gai S, Yang Z (2017) Two-dimensional discriminant locality preserving projections (2ddlpp) and its application to feature extraction via fuzzy set. Multimedia Tools and Applications

46. Wan M, Yang G, Sun C, Liu M (2019) Sparse two-dimensional discriminant locality-preserving projection (S2DDLPP) for feature extraction
47. Wang P, Li W, Gao Z, Tang C, Ogunbona PO (2018) Depth pooling based Large-Scale 3-D action recognition with convolutional neural networks. IEEE Trans Multimedia 20(5):1051–1061. https://doi.org/10.1109/TMM.2018.2818329
48. Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. IEEE Conf Comput Vis Pattern Recogn:1290–1297
49. Wang H, Schmid C (2013) Action recognition with improved trajectories. IEEE Int Conf Comput Vis:3551–3558
50. Wang C, Wang Y, Yuille AL (2013) An Approach to Pose-Based Action Recognition. IEEE Conference on Computer Vision and Pattern Recognition, Portland, pp 915–922
51. Wei P, Sun H, Zheng N (2018) Learning composite latent structures for 3D human action representation and recognition. IEEE Trans Multimed 21:2195–2208. https://doi.org/10.1109/TMM.2019.2897902
52. Wiliem A, Madasu V, Boles W, Yarlagadda P (2010) An Update-Describe approach for human action recognition in surveillance video. Int Conf Digit Image Comput Techn Appl:270–275
53. Xia L, Aggarwal JK (2013) Spatio-temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera. IEEE Conf Comput Vis Pattern Recogn:2834–2841
54. Xia L, Chen C, Aggarwal JK (2012) View invariant human action recognition using histograms of 3D joints. IEEE Comput Soc Conf Comput Vis Pattern Recogn Worksh:20–27
55. Xiao Y, Chen J, Wang YC, Cao ZG, Zhou JT, Bai X (2019) Action recognition for depth video using multi-view dynamic images. Inf Sci 480:287–304. https://doi.org/10.1016/j.ins.2018.12.050
56. Yang X. (2017) Super normal vector for human activity recognition with depth cameras. IEEE Trans Pattern Anal Mach Intell 39(5):1028–1039
57. Yang T, Hou Z, Liang J, Gu Y, Chao X (2020) Depth Sequential Information Entropy Maps and Multi-Label Subspace Learning for Human Action Recognition. In: IEEE Access, vol 8, pp 135118–135130. https://doi.org/10.1109/ACCESS.2020.3006067
58. Yang R, Yang R (2014) DMM-Pyramid based deep architectures for action recognition with depth cameras. Asian Conf Comput Vis:37–49
59. Yang X, Zhang C, Tian Y (2012) Recognizing actions using depth motion maps-based histograms of oriented gradients. ACM Multimed:1057–1060. https://doi.org/10.1145/2393347.2396382
60. Yao GL, Lei T, Zhong JD, Jiang P (2019) Learning multi-temporal-scale deep information for action recognition. Appl Intell 49:2017–2029. https://doi.org/10.1007/s10489-018-1347-3
61. Zeeshan A, Kandasamy I, Naimul K, Dimitri A (2019) Human action recognition using convolutional neural network and depth sensor data. Int Conf Inf Technol Comput Commun:1–5
62. Zhang B, Yang Y, Chen C, Yang L, Han J, Shao L (2017) Action recognition using 3D histograms of texture and a Multi-Class boosting classifier. IEEE Trans Image Process 26(10):4648–4660. https://doi.org/10.1109/tip.2017.2718189
63. Zhu QY, Siew CK (2004) Extreme learning machine: a new learning scheme of feedforward neural networks. IEEE Int Joint Conf Neural Netw 2:985–990. https://doi.org/10.1109/IJCNN.2004.1380068