Check for updates

International Journal of Advanced Robotic Systems
January-February 2019: 1–16
© The Author(s) 2019
DOI: 10.1177/1729881418825093
journals.sagepub.com/home/arx
SAGE

Hierarchical dynamic depth projected difference images–based action recognition in videos with convolutional neural networks

Hanbo Wu, Xin Ma and Yibin Li

Abstract

Temporal information plays a significant role in video-based human action recognition. How to effectively extract the spatial–temporal characteristics of actions in videos has always been a challenging problem. Most existing methods acquire spatial and temporal cues in videos individually. In this article, we propose a new effective representation for depth video sequences, called hierarchical dynamic depth projected difference images that can aggregate the action spatial and temporal information simultaneously at different temporal scales. We firstly project depth video sequences onto three orthogonal Cartesian views to capture the 3D shape and motion information of human actions. Hierarchical dynamic depth projected difference images are constructed with the rank pooling in each projected view to hierarchically encode the spatial–temporal motion dynamics in depth videos. Convolutional neural networks can automatically learn discriminative features from images and have been extended to video classification because of their superior performance. To verify the effectiveness of hierarchical dynamic depth projected difference images representation, we construct a hierarchical dynamic depth projected difference images–based action recognition framework where hierarchical dynamic depth projected difference images in three views are fed into three identical pretrained convolutional neural networks independently for finely retuning. We design three classification schemes in the framework and different schemes utilize different convolutional neural network layers to compare their effects on action recognition. Three views are combined to describe the actions more comprehensively in each classification scheme. The proposed framework is evaluated on three challenging public human action data sets. Experiments indicate that our method has better performance and can provide discriminative spatial–temporal information for human action recognition in depth videos.

Keywords

Human action recognition, depth videos, rank pooling, dynamic images, CNN

Date received: 1 September 2018; accepted: 25 December 2018

Topic: AI in Robotics; Human Robot/Machine Interaction
Topic Editor: Henry Leung
Associate Editor: Huaping Liu

Introduction

Human action recognition has attracted increasing attention throughout the computer vision community over the past years. The traditional methods based on the red, green and blue (RGB) data for action recognition usually focus on body shape feature,[1] key poses[2] and son on. Although they may have achieved high performance in some specific

School of Control Science and Engineering, Shandong University, Jinan, China

Corresponding authors:
Xin Ma and Yibin Li, School of Control Science and Engineering, 17923 Jingshi Road, Shandong University, Jinan 250061, Shandong, China.
Emails: maxin@sdu.edu.cn; liyb@sdu.edu.cn

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (http://www.creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

contexts, however, RGB action recognition methods are sensitive to changes of lighting conditions and fail to recognize actions in more challenging scenarios when there exist occlusions and clutter backgrounds.

The introduction of low-cost integrated depth sensors (such as Microsoft Kinect™, Redmond, Washington) that can capture both RGB video and depth information has significantly advanced the research of human action recognition. Compared with conventional RGB cameras, Kinect depth sensors provide us the 3D structural information of the scene that is useful in facilitating the recognition task by simplifying intra-class motion variations and removing cluttered background noise. Furthermore, the depth information can eliminate the effects of illumination and colour variations. Therefore, researchers have put lot of attentions to the depth-data-based human action recognition and proposed effective features such as depth motion maps (DMM),[3] local occupancy pattern (LOP),[4] histogram of oriented 4D normal,[5] super normal vector (SNV),[6] depth cuboid similarity feature (DCSF)[7] and {Xia, 2013 #8}range-sample depth feature.[8]

In the last decade, thanks to the significant advancements in computational capabilities and the availability of large amount of annotated data sets,[9,10] deep learning has gained a lot of focus and been widely used to address various computer vision challenges. The most popular deep neural network model is convolutional neural networks (CNNs) introduced by LeCun et al.[11] CNN can automatically learn powerful and discriminative image features and has been demonstrated as an effective model for understanding image content. An increasing number of researchers start to apply CNN in video-based action recognition tasks.[12–21] However, most of the existing action recognition works rely on RGB data or skeleton data, moreover, currently existing public available action recognition data sets that almost all deep neural network models are evaluated on are composed of RGB videos alone, such as UCF-101,[22] HMDB51[23] and Kinetics.[24] There are only few researches on depth-based human action recognition using CNN,[25–28] because the depth training data are relatively small-scale. Recently, a new large-scale benchmark data set named NTU RGB + D data set[29] is proposed to overcome the limitations of depth data-based human action recognition with CNN.

In this article, we construct a CNN-based action recognition framework with the proposed hierarchical dynamic depth projected difference images (HDDPDI). HDDPDI are presented as a simple and efficient descriptor to extract the spatial–temporal motion information in depth videos. For a depth video sequence, each depth frame is projected onto three orthogonal Cartesian views. Then depth maps in each projected view are sampled at several different temporal scales. Depth projected difference image (DPDI) is defined as the absolute difference image between two consecutive depth maps. We compute DPDI sequences at different temporal scales for each projected view, which can reflect the spatial motion and variation of an action more comprehensively. Dynamic image is introduced as a simple and powerful representation for a video. Bilen et al.[30] applied rank pooling[31] that is an effective temporal pooling method on the raw image pixels of a RGB video sequence to produce the RGB dynamic image. Inspired by this idea, we extend dynamic image to depth data and propose that utilizing rank pooling encodes a DPDI sequence to generate the dynamic depth projected difference image (DDPDI). DDPDI integrates the whole changing process of an action into a single dynamic image in time order and captures the spatial–temporal variations of a depth video effectively. DDPDIs at different temporal scales in each view form HDDPDI. Finally, the HDDPDI in three views for depth videos are fed into three identical CNNs independently. Three CNNs pretrained on ImageNet are finely retuned using HDDPDI in corresponding view, respectively. To fully verify the validity of the proposed HDDPDI video representation as well as to compare the influences of different CNN layers on action recognition, we design three classification schemes in the recognition framework where different CNN layers are used. Since three projected views can offer complementary characteristics for human actions, multimodal information fusion[32–34] is applied in each classification scheme and results of three views are combined for action recognition. The proposed action recognition framework is described in Figure 1.

A major source of inspiration comes from DMM.[3] Each frame in a depth video sequence is projected onto three orthogonal Cartesian planes to form three projected image sequences. Under each projection view, the thresholded absolute difference between two consecutive projected maps is accumulated across an entire depth video sequence forming DMM. DMM contain the motion change information in a depth video; however, the accumulation operator ignores the time sequence. Temporal order is an important factor in videos and contributes significantly to the final action recognition. Hence, to capture the temporal information effectively, we apply rank pooling[31] on DPDI sequences in each projected view to get DDPDIs that include the spatial–temporal variances of the whole video. It's worth mentioning that pseudocolour coding[25] can remap the spatiotemporal information of human actions. Compared with this work, rank pooling models the evolution of appearance and dynamics over time in a video. It not only captures the temporal dynamics in videos robustly but also is easily implemented and fast computed. Therefore, rank pooling method is utilized in this article to get the dynamic representation for a depth video.

The major contributions of this article can be summarized as follows: (1) HDDPDI are proposed as a representation of a depth video for extracting the spatial–temporal dynamics. With the help of rank pooling and dynamic image, this method overcomes the drawback of ignoring video temporal information in original DMM[3] and improves the discrimination of human action recognition.

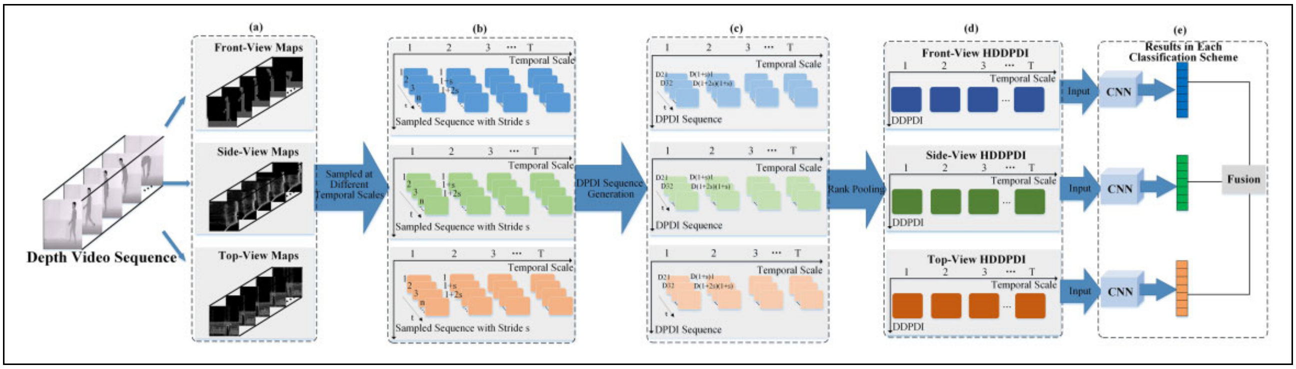


Figure 1. Overview of the proposed HDDPDI-based action recognition framework. (a) Depth videos are projected onto three orthogonal views. (b) Depth maps in each view are sampled at different temporal scales with stride *s*. (c) DPDI sequence is generated by computing absolute difference image for two consecutive images across a sampled depth map sequence. (d) Rank pooling is applied on DPDI sequences at different temporal scales to produce HDDPDI in each view. (e) HDDPDIs of three views are used to train three CNNs independently and results of three views are fused for action recognition in each classification scheme. HDDPDI: hierarchical dynamic depth projected difference images; DPDI: depth projected difference images; CNN: convolutional neural network.

(2) We extend the dynamic image to depth data by applying rank pooling on DPDI sequences. DDPDIs of different temporal scales can hierarchically describe the spatial–temporal dynamics of an action. (3) We construct a HDDPDI-based action recognition framework to demonstrate the effectiveness of HDDPDI representation, where HDDPDIs in three views are inputted into three CNNs independently and three classification schemes are designed using different CNN layers. The results of three views are fused for action recognition. (4) State-of-the-art recognition performance is achieved on three challenging action data sets. The results are analysed in detail for more findings.

The rest of this article is organized as follows: The second section briefly presents the related works. In the third section, we elaborate the construction procedure of the proposed HDDPDI. In the fourth section, we describe the details of the action recognition framework. Experimental results and analysis are reported in the fifth section. The sixth section concludes the article.

Related works

Depth-based action recognition

Emergence of low-cost depth sensors makes depth data available, which extends the researches for human action recognition from RGB to depth. Various algorithms are developed for depth video-based action recognition. We review them from two aspects: handcrafted and deep learning approaches. More comprehensive surveys[35–37] summarize these works in detail.

Handcrafted algorithms. Many approaches for human action recognition in videos are based on depth data. Yang et al.[3] accumulated depth maps projected onto three orthogonal planes to generate DMM. The histograms of the oriented gradients (HOG) were used to extract features from DMM. Chen et al.[38] used local binary patterns (LBPs) to get feature representations based on DMMs as well. Wang et al.[4] proposed a 3D LOP feature for capturing the local depth appearance information based on the joint locations. Oreifej et al.[5] extended surface normals to 4D space and constructed histogram of oriented 4D normals (HON4D) as the feature descriptor. An action recognition scheme was proposed to aggregate the low-level polynormals produced by clustering hypersurface normals in depth sequences into the SNV.[6] The spatial–temporal DCSF[7] was presented to describe the local 3D depth cuboids around the spatial–temporal interest points (STIPs) extracted from depth videos. Lu et al.[8] proposed a binary range-sample feature that can exclude clutter background and complex occlusion to capture shape and motion of the human body in depth sequences.

Deep learning algorithms. A large amount of work[12,21] based on CNN has been done for human action recognition in videos inspired by its remarkable performance. There are some state-of-art achievements that perform well for action recognition in RGB videos, for example, 3D convolutional networks (C3D),[12] two-stream convolutional networks,[13] trajectory-pooled based deep-convolutional descriptors,[15] temporal segment networks (TSN)[16] and so on. However, depth-based action recognition methods with CNN are rare. Wang et al.[25] used weighted hierarchical depth motion maps (WHDMM) as the inputs of CNN and produced a three-channel architecture to acquire the final classification results. Dynamic depth images (DDI), dynamic depth normal images (DDNI) and dynamic depth motion normal images (DDMNI) were proposed as three representations for depth sequences and were fed into CNN for action classification.[26] Features learned from RGB videos are utilized for depth videos directly by domain adaptation to do action recognition.[27] Motion history images (MHI) generated from RGB videos are added into DMM to construct a four-channel deep CNN.[28] In this article, we focus on human action recognition in depth videos with the purpose of taking full advantage of the depth data.

Temporal order modelling

Different from the image classification tasks, action videos are 3D and contain rich spatial–temporal information. Since videos can be represented as image sequences, most existing feature extraction algorithms[3,8] are all frame-level, so how to model the temporal structure within a video is a considerable problem. Some early works used conditional random fields (CRFs)[39] and hidden Markov models (HMMs)[40] to model temporal information. These two methods need a large amount of training samples to learn model parameters. Wang et al.[41] applied Fourier temporal pyramid (FTP) to encode the temporal changes, which is robust to the noisy data. Optical flow[13] was also commonly used to capture the time variations. However, computation of optical flow is heavy and time consuming. It is not suitable for depth videos that lack the texture information. Pseudocolour coding[25] is another commonly used approach to reflect the temporal order of a video. Pseudocolour coding maps indicate the motion temporal order by colour intensity. Recurrent neural networks (RNN)[42] is proposed to handle the sequential data. It can display the action time dynamics by internal states and is generally combined with CNN for action recognition.[43] As a new efficient temporal pooling method, rank pooling[31] captures temporal dynamics of the whole video by modelling the evolution of appearance and dynamics over time. It is easily implemented and fast computed serving as a robust video representation.

Hierarchical dynamic depth projected difference images

To describe the human 3D spatial motion information and temporal dynamics for actions in depth videos, we propose HDDPDI as an effective video representation method. In this section, we firstly introduce the depth video projection and then we describe the sampling procedure on the projected depth maps under different temporal scales. Finally, the construction of HDDPDI using rank pooling is explained elaborately.

Depth video projection

Depth videos contain rich 3D structure and shape information and can help to improve the human action recognition performance significantly. Yang et al.[3] proposed DMM to capture the 3D motion information of human actions in depth videos. Considering this advantage, we use the same approach for depth video projection. Specifically, each frame in a depth video sequence is projected onto three 2D orthogonal Cartesian planes where X-Y plane represents front view, Y-Z plane represents side view and X-Z plane represents top view. For a point (x, y, z) in a depth image, its pixel value in projected front\side\top view is $z\backslash x\backslash y$. Therefore, we get depth maps in three projected views, respectively, for each depth video sequence. Depth maps in three views for some actions in NTU RGB + D data set are illustrated in Figure 2.

Scaled sampling of depth maps

Each frame of a depth video is projected onto three views using the same method as DMM[3] for capturing the 3D human shape and motion information. Besides, we make two extensions based on the projected depth maps. The first one is that to describe the motion changes of human actions from coarse to fine and suppress noise as well, the depth map sequence in each projected view is scaled-sampled progressively along the time dimension, which produces a set of sampled sequences with different lengths named as different temporal scales. This process can also be regarded as a data augmentation method for increasing the number of action samples.

For a depth video projected sequence $\mathrm{V} = \{\mathrm{map}_v^1, \ \ldots, \ \mathrm{map}_v^N\}$, where N is the number of total frames and v represents the front, side or top view. We sample the depth map sequence with a stride s from the start frame f. The sampling stride is s and besides that, we set the stride of start frame to s_f. The original depth map sequence is named as the first temporal scale. Then we initialize the start frame $f = 1$ and get a sampled sequence with [N/s] frames. [N/s] is the nearest integer larger than N/s. We update f with its stride s_f and sample the original depth map sequence with the stride s iteratively. Considering that if a sampled sequence is too short, it won't contain the key information of an action. Therefore, we set a ratio r and define that the start frame index should not exceed $r \times N$ to guarantee the length of a sampled sequence. Ratio r controls the lower limit of the start frame offset and further ends up the sampling process. Finally, a group of sampled sequences with different lengths are generated progressively along the time dimension, which can capture the action information in a coarse-to-fine way. And each sequence represents a temporal scale. The number of temporal scales of a depth video in three projected views is the same and related to the video duration (N); therefore, different depth videos may have different numbers of temporal scales. Suppose that one depth video has T temporal scales, the scaled-sampling process is illustrated in Figure 3. In this way, we get a set of coarse-to-fine depth map sequences at multiple temporal scales in each projected view for a depth video.

HDDPDI construction with rank pooling

For a depth video, we have obtained a group of scaled depth map sequences sampled at several time scales. The thresholded absolute difference between two consecutive maps

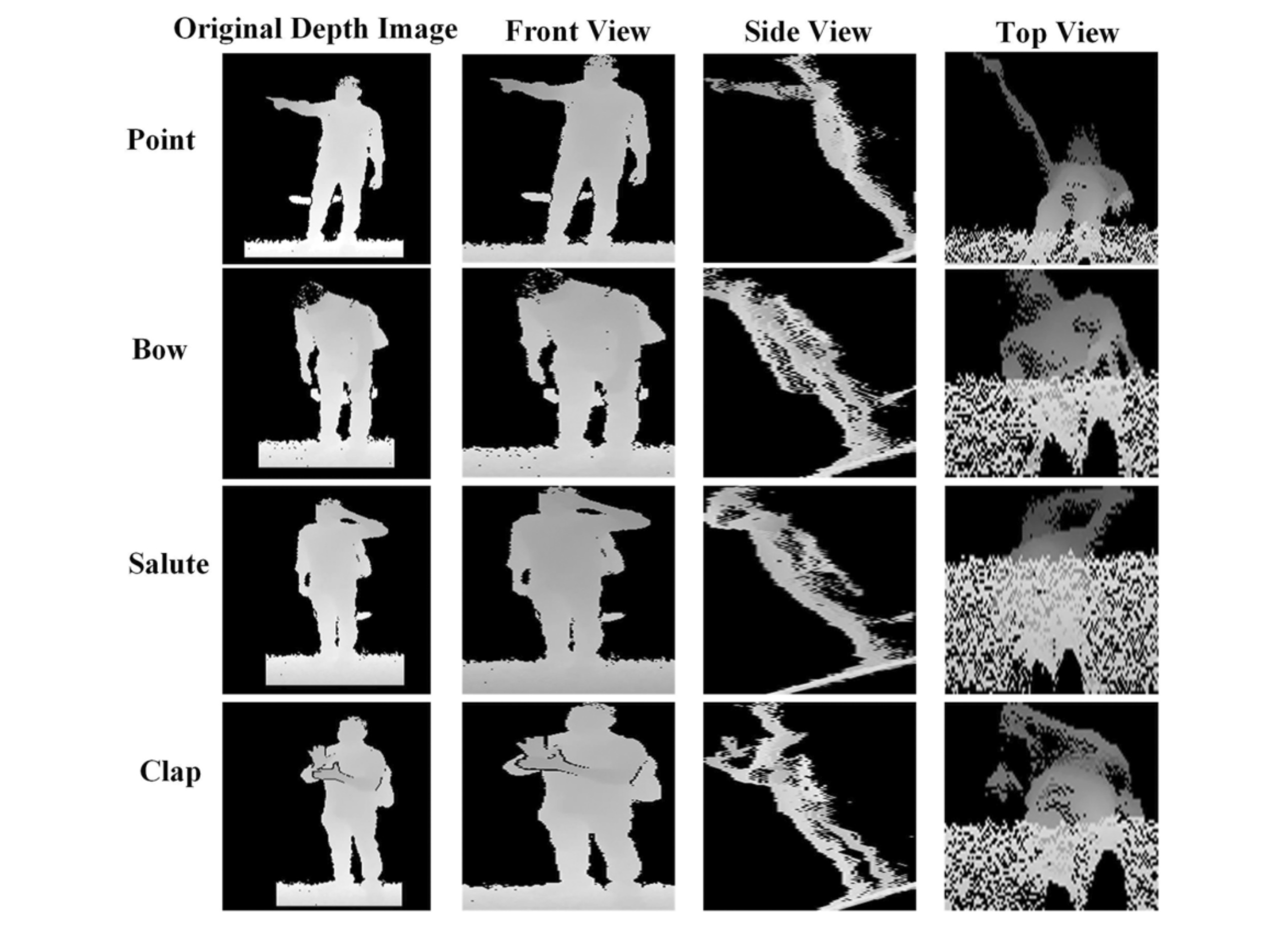


Figure 2. Depth maps in three views for some actions in NTU RGB + D data set.

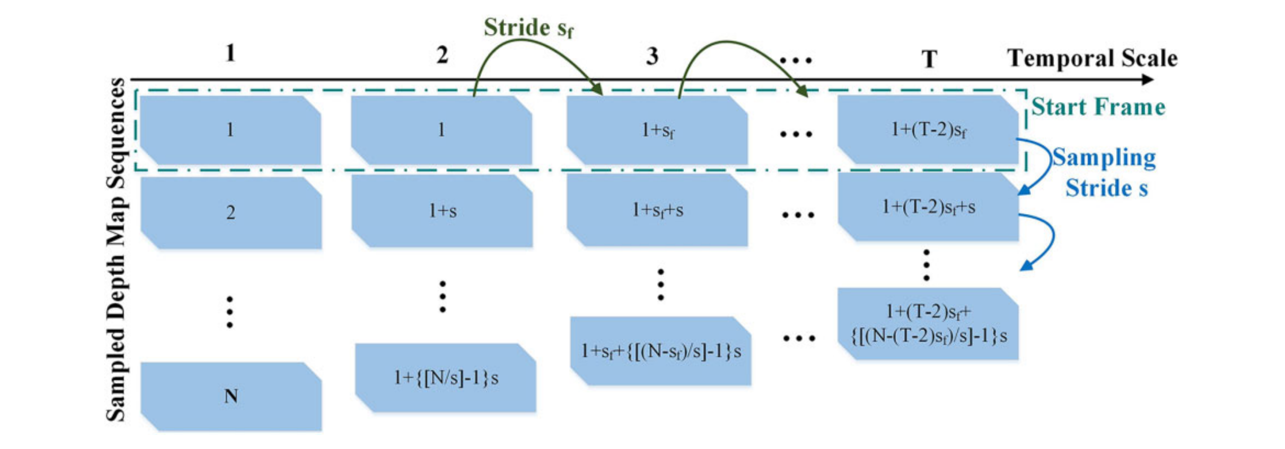


Figure 3. Flow chart of the scaled sampling. The original depth map sequence has *N* frames. *T* is the number of temporal scales. Sampling stride is *s* and start-frame stride is s_f. [*X*] represents the nearest integer larger than *X*.

is accumulated across an entire depth map sequence in the original DMM.[3] Similarly, we compute the absolute difference without thresholding for two consecutive maps to extract the motion changes, referred to as DPDI. It can be expressed as equation (1). Let DPDI_{vt}^{j} be the j^{th} DPDI at temporal scale t in projected view v, $v \in (\text{front}, \text{side}, \text{top})$

$$\mathrm{DPDI}_{vt}^{j} = |\mathrm{map}_{vt}^{j+1} - \mathrm{map}_{vt}^{j}| \tag{1}$$

where map_{vt}^{j} is the jth depth map of temporal scale t in view v. DPDIs across the sequence at each temporal scale form a DPDI sequence, as shown in Figure 1(c).

Different from the original DMM,[3] rather than accumulation, we then apply rank pooling on a DPDI sequence temporally to get a single dynamic image for encoding the spatial–temporal information of human motion changes effectively. Rank pooling[31] is a new temporal pooling method which not only captures the temporal changes of videos robustly but also is easily implemented. It utilizes pairwise linear ranking machines to learn a linear function whose parameters can encode the frame order within a video and be used as a new video representation. We consider the chronological order of one DPDI sequence and aggregate the motion changing information into a dynamic image using rank pooling.

Rank pooling is applied directly on the pixels of DPDIs in this article. Let a DPDI sequence with k frames at temporal scale t in projected view v be represented as $\mathrm{DPDI}_{vt} = \{\overline{\mathrm{DPDI}_{vt}^{1}}, \ldots, \overline{\mathrm{DPDI}_{vt}^{j}}, \ldots, \overline{\mathrm{DPDI}_{vt}^{k}}\}$. $\overline{\mathrm{DPDI}_{vt}^{j}}$ is the vectorized DPDI_{vt}^{j}. Time varying mean vector operation[31] is applied as equation (2) to capture the temporal information from the independent DPDI frames

$$\begin{cases} m_i = \dfrac{1}{i}\displaystyle\sum_{j=1}^{i} \overline{\mathrm{DPDI}_{vt}^{j}} \\ d_i = \dfrac{m_i}{\| m_i \|} \end{cases} \tag{2}$$

The smoothed vector sequence $d = \{d_1, \ldots, d_i, \ldots, d_k\}$ still remains the time order of k frames in original DPDI sequence. A linear rank function is defined as $\varphi(d; \alpha) = \alpha^{\mathrm{T}} \cdot d$, where $\alpha \in R^{D}$. α is a parametric vector of the rank function that can preserve the relative orderings of frames. That is, if $\forall\ t_i > t_j$, the rank value satisfies $\varphi(d_{t_i}; \alpha) > \varphi(d_{t_j}; \alpha)$. The objective function of rank pooling is defined with structural risk minimization as equation (3)

$$\begin{cases} \min\limits_{\alpha} \quad \dfrac{1}{2}||\alpha||^2 + C \displaystyle\sum_{\forall t_i > t_j} \varepsilon_{ij} \\ s.t.\ \alpha^{\mathrm{T}} \cdot (d_{t_i} - d_{t_j}) \geq 1 - \varepsilon_{ij} \\ \qquad \varepsilon_{ij} \geq 0 \end{cases} \tag{3}$$

where ε_{ij} is a slack variable. $\alpha*$ is the found optimal parametric vector and can be used as a descriptor of the DPDI sequence. We transform vector $\alpha*$ into an image called dynamic DPDI (DDPDI) that aggregates the spatial–temporal motion information of the whole video.

In each projected view for a depth video, DPDI sequences of all temporal scales are processed using rank pooling to form DDPDIs, as shown in Figure 1(d). Part of DDPDIs along the different temporal scales in the front view for some actions in NTU RGB + D data set are shown in Figure 4. Since the original depth map sequences are sampled progressively along the time, so DDPDIs are also dynamically progressive along the temporal scale for human actions. These DDPDIs at different temporal scales in each projected view for a depth video are named as HDDPDI, which can be used as an effective representation for the video.

HDDPDI-based action recognition with CNN

CNN training

After the construction of HDDPDIs in three views, VGG16[44] is adopted as the basic network structure of our action recognition framework in this article. VGG16 contains five convolutional layers, three fully connected (FC) layers and a softmax classifier layer. We train a VGG16 network independently for each view, seen in Figure 1. Three VGG16 networks pretrained on the ImageNet data set are fine-tuned to avoid training a lot of parameters from scratch. The implementation is completed using Pytorch.[45]

During the training process of each view, HDDPDIs are human-centric cropped to 224 × 224 as the inputs of CNN. Output of the last FC layer is adjusted to C, which is the number of action categories. Based on cross-entropy loss function, the stochastic gradient descent algorithm is used to learn network weights with a mini-batch size of 32 samples, momentum of 0.9 and weight decay of 10^{-3}. The initial learning rate is set to 10^{-4} and will be decreasing as the training goes on. Iteration number of the training is 100 epochs. Random horizontal flip and rotation are applied for data augmentation.

Classification schemes

To verify the effectiveness of the proposed HDDPDI representation, we construct a CNN-based action recognition framework. CNN features at different layers encode different levels of information. So with the purpose of comparing the influences of different CNN layers on action recognition, three classification schemes are designed in the framework, with the last convolutional (LC) layer, the FC layer and the softmax layer of CNN, respectively. In this article, LC layer is defined as the fifth convolutional layer in VGG16. FC layer is the second FC layer in VGG16. Softmax layer is the final classification function of VGG16. In the classification schemes with LC layer and FC layer, HDDPDIs of a depth video in three views are fed into three corresponding CNNs, respectively, for feature representation. Features of three views are fused to capture the complementary characteristics of human actions. For the scheme with softmax layer, we use CNN in an end-to-end mode and fuse the prediction scores of three views for recognition. We describe the classification schemes in detail as follows.

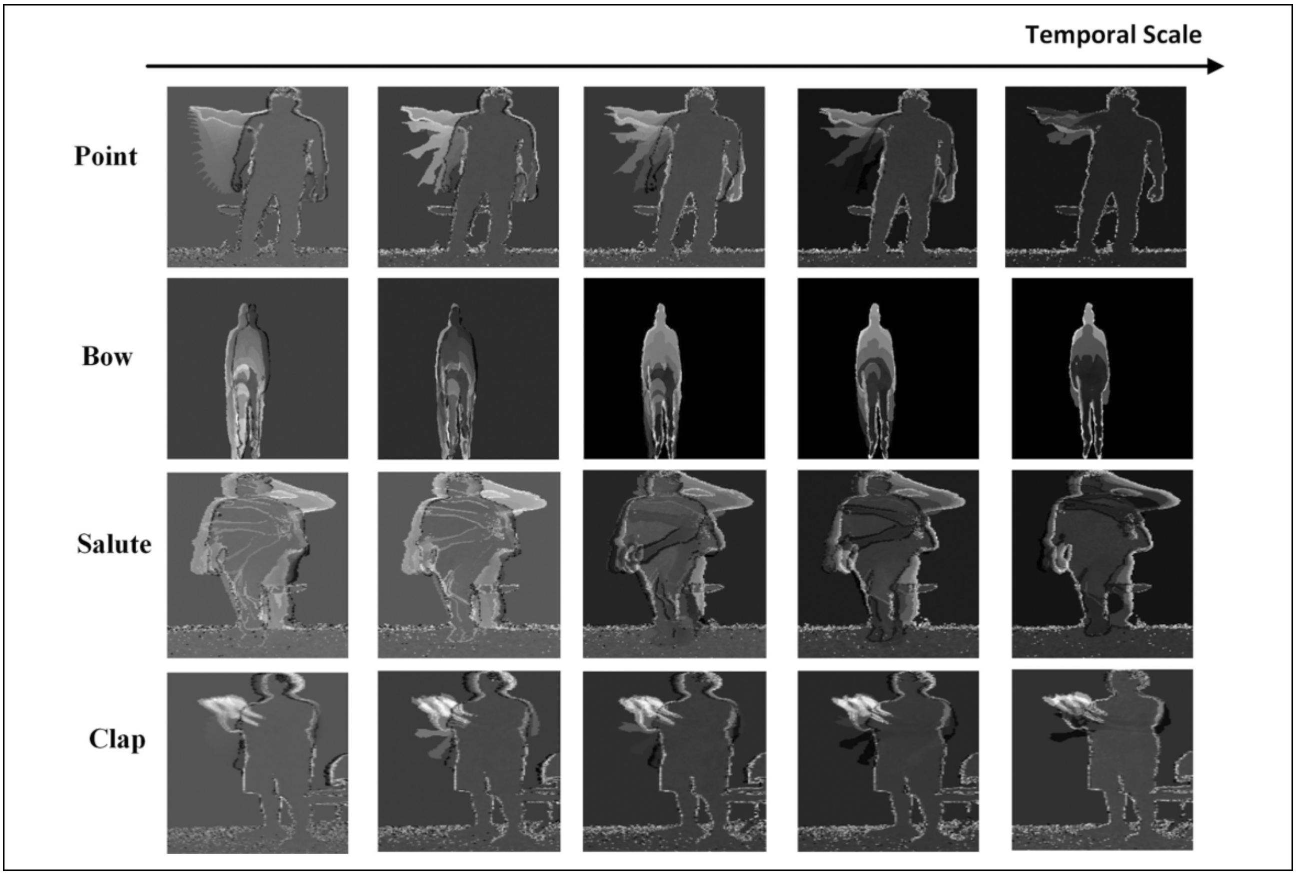


Figure 4. DDPDIs along the temporal scale in the front view for some actions in NTU RGB + D data set. DDPDI: dynamic depth projected difference images.

Recognition with LC layer. Convolutional features focus on spatial structure information of actions in HDDPDIs, such as colour, edge and texture. The LC layer in VGG16 has a larger receptive field and encodes more complex spatial features. Therefore, for each view we, extract the LC layer feature maps for HDDPDI of a depth video to get the corresponding feature representation. Take the front view as an example, we assume that the LC layer feature maps of HDDPDI of a depth video in this view are represented as a 4D tensor $F_{\mathrm{LC}} \in R^{H \times W \times C \times T}$, where H and W are both 7 denoting height and width of the LC layer feature map, C is 512 denoting the number of channels (filters), T is the number of temporal scales of HDDPDI. Considering that HDDPDI are dynamic representations of a depth video under different time scales, we accumulate the feature maps of all temporal scales in each channel for feature enhancement. The accumulation operator is shown in equation (4)

$$\hat{F}^{i}_{\mathrm{LC}} = \sum_{t=1}^{T} F_{\mathrm{LC}}(H, W, i, t) \qquad (4)$$

where $\hat{F}^{i}_{LC}$ is the accumulated feature map of the channel i. $\hat{F}_{\mathrm{LC}} = \{\hat{F}^{i}_{\mathrm{LC}}\}_{i=1\ldots C}, \in R^{H\times W\times C}$ is normalized with channel normalization.[15] And we apply max pooling on the normalized convolutional feature maps to get the max feature response in each channel. The LC layer feature descriptor of a depth video in front view is represented as $V^{\mathrm{Front}}_{\mathrm{LC}} = \{v^{i}_{\mathrm{LC}}\}_{i=1\ldots C}, \in R^{C}$, where v^{i}_{LC} is the max response value on the normalized accumulated feature map in channel i. The generation flow chart of the LC layer feature descriptor is described in Figure 5. In the same way, we can get the feature descriptors in side view and top view. LC layer features of three views of a depth video are concatenated as the final feature representation. A multi-class linear SVM is used for action recognition in this classification scheme.

Recognition with FC layer. Compared with convolutional features, FC layer features pay more attention to the abstract semantic information. HDDPDIs of three views of a depth video serve as inputs to the three trained CNNs and then we obtain the FC layer feature descriptor for each view. For a depth video, FC layer output of HDDPDI in front view is represented as a 2D tensor $F_{\mathrm{FC}} \in R^{T\times D}$, where T is the number of temporal scales in HDDPDI and D is the dimension of FC layer output in VGG16. Since HDDPDI represent dynamic images of a depth video at different

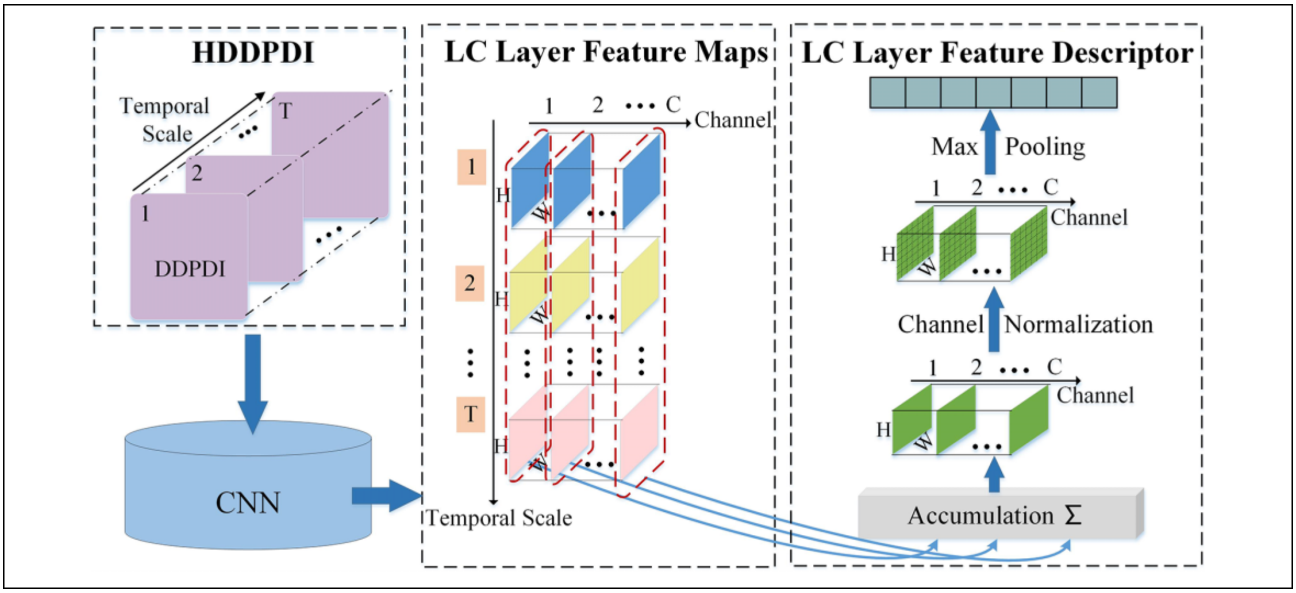


Figure 5. The generation flow chart of the LC layer feature descriptor in each view for a depth video. LC: last convolutional.

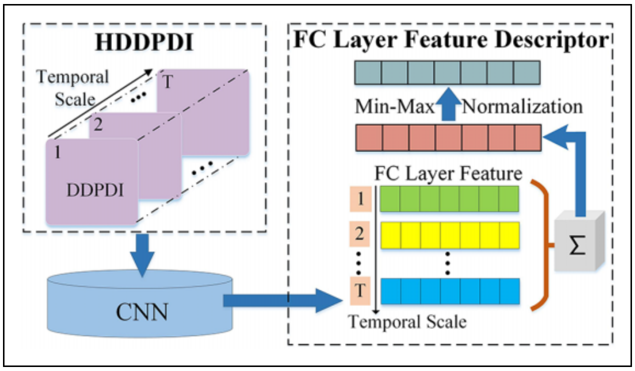


Figure 6. The generation flow chart of the FC layer feature descriptor in each view for a depth video. FC: fully connected.

temporal scales, we accumulate the FC layer features of all-time scales for feature reinforcement, as shown in equation (5)

$$\hat{F}_{\mathrm{FC}} = \sum_{t=1}^{T} F_{\mathrm{FC}}(t, D) \tag{5}$$

where $\hat{F}_{\mathrm{FC}} \in R^D$ is the accumulated FC feature and is normalized with min-max normalization. Finally, the normalized $\hat{F}_{\mathrm{FC}}$ is used as the FC layer feature descriptor of a depth video in front view and is reformulated as $V_{\mathrm{FC}}^{\mathrm{Front}} \in R^D$. We illustrate the generation flow chart of the FC layer feature descriptor in Figure 6. And the FC layer feature descriptors in side view and top view are calculated with the same method. We combine the FC layer feature descriptors in three views as the final FC feature representation for a depth video. A multi-class linear SVM is also applied for action recognition.

Recognition with softmax layer. CNN is a deep neural network which can capture features and make classification automatically for images by end-to-end learning. Output of softmax layer in CNN represents the class probability. For each view, we take HDDPDI of a depth video as the input and get the softmax output denoted as a 2D tensor $\mathrm{P} \in R^{T \times A}$, where T is the number of temporal scales in HDDPDI, A is the number of action categories. $\mathrm{P} = \{p_t^i\}_{t=1\ldots T, i=1\ldots A}$, p_t^i is the probability of ith action class at temporal scale t. We then apply max operator, average operator and multiply operator, respectively, for probabilities of each action class in all scales, as shown in equations (6) to (8).

$$p^i_{\mathrm{max}} = \max_{t=1\ldots T} p_t^i \tag{6}$$

$$p^i_{\mathrm{ave}} = \frac{1}{T} \sum_{t=1}^{T} p_t^i \tag{7}$$

$$p^i_{\mathrm{mul}} = \prod_{t=1}^{T} p_t^i \tag{8}$$

p^i_{max}, p^i_{ave} and p^i_{mul} are the prediction probabilities of ith action class under three operators named as softmax–max, softmax–average and softmax–multiply. We then average the prediction class probabilities of three views of a depth video for each operator as the final prediction class probability under the corresponding operator. Index of the max probability of all action classes corresponds to the recognized class label. The generation flow chart of the class probabilities of a depth video in each view under three operators is described in Figure 7.

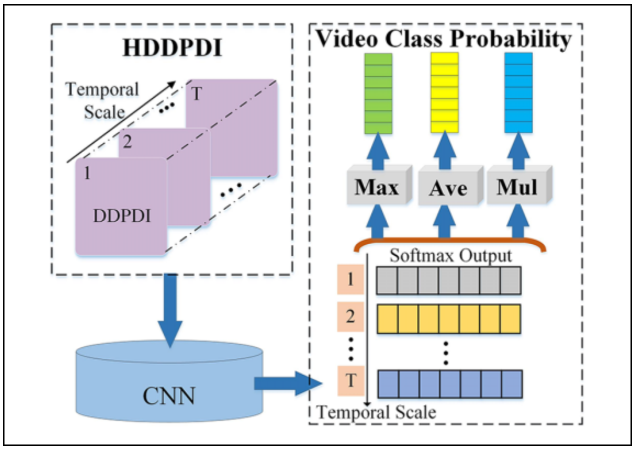


Figure 7. The generation flow chart of the class probabilities of a depth video in each view under three operators.

Experiments and discussions

We evaluate the proposed HDDPDI-based action recognition framework on three challenging data sets. In this section, we first introduce the three human action data sets and the basic experimental settings. Next, we present and analyse the recognition results on the three data sets. Furthermore, we also conduct the experiments separately on each single- view to demonstrate the fusion advantages that three projected views can provide complementary information for action recognition. Finally, we show the comparisons with the state-of-the-art methods.

Data sets

In our experiments, we evaluate the proposed HDDPDI video representation with three CNN-based classification schemes on the following publicly available human action data sets: SDUFall,[46] MSRAction3D[47] and NTU RGB + D.[29]

The SDUFall data set[46] was built by our Robot Research Center in Control Science and Engineering College, Shandong University. The data set is collected by a Kinect camera installed 1.5 m high in a laboratory environment. It contains six action classes: bending, falling, lying, sitting, squatting and walking. Each action is performed 10 times by 20 subjects, and there are total 1200 samples. Furthermore, SDUFall data set contains rich variations such as illumination, direction and position changes.

The MSRAction3D data set[47] was built by the Advanced Multimedia Research Lab in University of Wollongong. It contains 20 action types performed by 10 subjects. Each subject performs each action 2 or 3 times. There are 567 depth sequences in total. The 20 actions are 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 pick up and throw.

The NTU RGB + D dataset[29] was built by the ROSE Lab in the Nanyang Technological University. It is the largest RGB-D action recognition data set till now. This data set is captured by three Microsoft Kinect v.2 cameras concurrently. NTU RGB + D action recognition data set consists of 56,880 action samples, containing 60 different action classes performed by 40 volunteers. The 60 actions are drinking, eating, brushing teeth, brushing hair, dropping, picking up, throwing, sitting down, standing up (from sitting position), clapping, reading, writing, tearing up paper, wearing jacket, taking off jacket, wearing a shoe, taking off a shoe, wearing on glasses, taking off glasses, putting on a hat/cap, taking off a hat/cap, cheering up, hand waving, kicking something, reaching into self-pocket, hopping, jumping up, making/answering a phone call, playing with phone, typing, pointing to something, taking selfie, checking time (on watch), rubbing two hands together, bowing, shaking head, wiping face, saluting, putting palms together, crossing hands in front, sneezing/coughing, staggering, falling down, touching head (headache), touching chest (stomach ache/heart pain), touching back (back-pain), touching neck (neck-ache), vomiting, fanning self, punching/slapping other person, kicking other person, pushing other person, patting other's back, pointing to the other person, hugging, giving something to other person, touching other person's pocket, handshaking, walking towards each other and walking apart from each other. This data set is challenging due to a large number of action samples and classes as well as rich intraclass variations.

Experimental settings

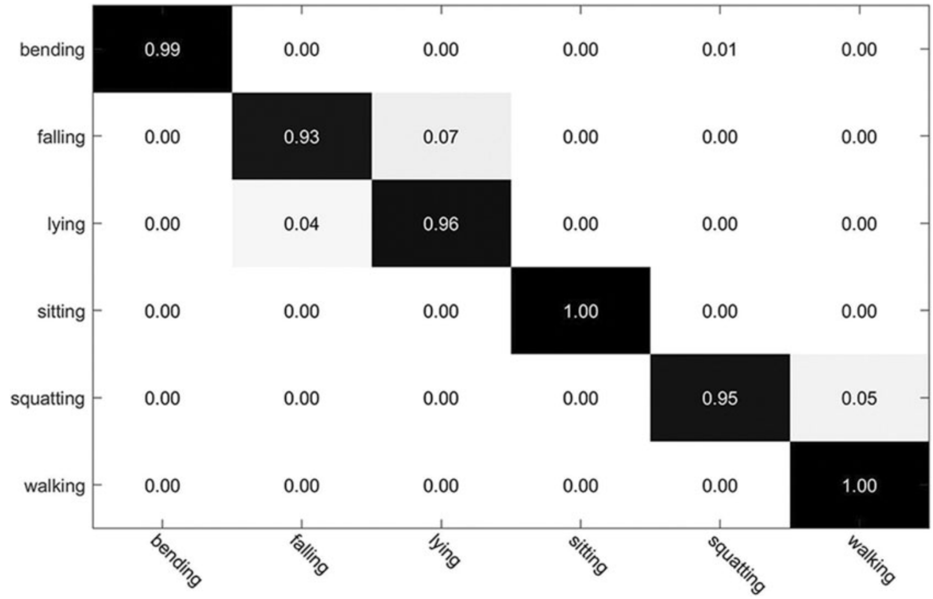
Since the proposed HDDPDI representation is based on depth videos, we project all depth video sequences in a data set onto three 2D orthogonal Cartesian planes to get depth map sequences with the same method as the original DMM.[3] The depth map sequence of a video in each projected view is scaled-sampled at different temporal scales. The start frame stride s_f, sample stride s and ratio r are different for different data sets in the sampling process. We empirically select these parameters depending on characteristics of different data sets. For SDUFall data set, $s_f = 5$, $s = 3$ and $r = 0.7$. For MSRAction3D data set, $s_f = 2$, $s = 3$ and $r = 0.3$. For NTU RGB + D data set, $s_f = 5$, $s = 3$ and $r = 0.3$. In our experiments, the number of temporal scales in HDDPDI representations of depth videos in different data sets ranges from 3 to 20, which is far less than the number of total frames. Therefore, the HDDPDI representation can not only capture the informative spatial–temporal dynamics for human actions but also help to reduce the computation complexity.

In each projected view, we take HDDPDIs of depth videos as inputs of VGG16 for finely retuning. To avoid overfitting, drop out ratio after FC layer is adjusted to 0.5. Moreover, data augmentation methods such as horizontal flip and rotation are also applied. For action recognition, we design three classification schemes based on CNN, with LC layer, FC layer and softmax layer, respectively. The spatial size of the feature map in LC layer is 7 × 7. The dimension of the feature in FC layer is 4096. Softmax output size is

Table 1. Recognition accuracies on the SDUFall data set with different classification schemes in the proposed HDDPDI-based action recognition framework.

	Classification scheme				
Data set	LC layer	**FC layer**	Softmax-max	Softmax-average	Softmax-multiply
SDUFall	93.64%	**97.08%**	95.83%	96.25%	96.04%

Bold values denote the highest recognition accuracy and the corresponding classification scheme in SDUFall. HDDPDI: hierarchical dynamic depth projected difference images; LC: last convolutional; FC: fully connected.



	bending	falling	lying	sitting	squatting	walking
bending	0.99	0.00	0.00	0.00	0.01	0.00
falling	0.00	0.93	0.07	0.00	0.00	0.00
lying	0.00	0.04	0.96	0.00	0.00	0.00
sitting	0.00	0.00	0.00	1.00	0.00	0.00
squatting	0.00	0.00	0.00	0.00	0.95	0.05
walking	0.00	0.00	0.00	0.00	0.00	1.00

Figure 8. Confusion matrix on the SDUFall data set for the FC layer classification scheme. FC: fully connected.

related to the number of action classes, which is 6 for SDUFall, 20 for MSRAction3D and 60 for NTU RGB + D.

HDDPDI performance evaluation with three classification schemes

SDUFall data set. We evaluate the proposed HDDPDI-based action recognition framework with CNN on the SDUFall data set. In our framework, HDDPDI are used as a new representation of a depth video and then fed into CNN for action recognition with three classification schemes. Three views are fused for recognition in each scheme. Three-fifth of subjects in the SDUFall data set are selected randomly for training and the remaining for testing. Three CNNs pretrained on ImageNet are retuned independently using the corresponding HDDPDIs of depth videos in the training set.

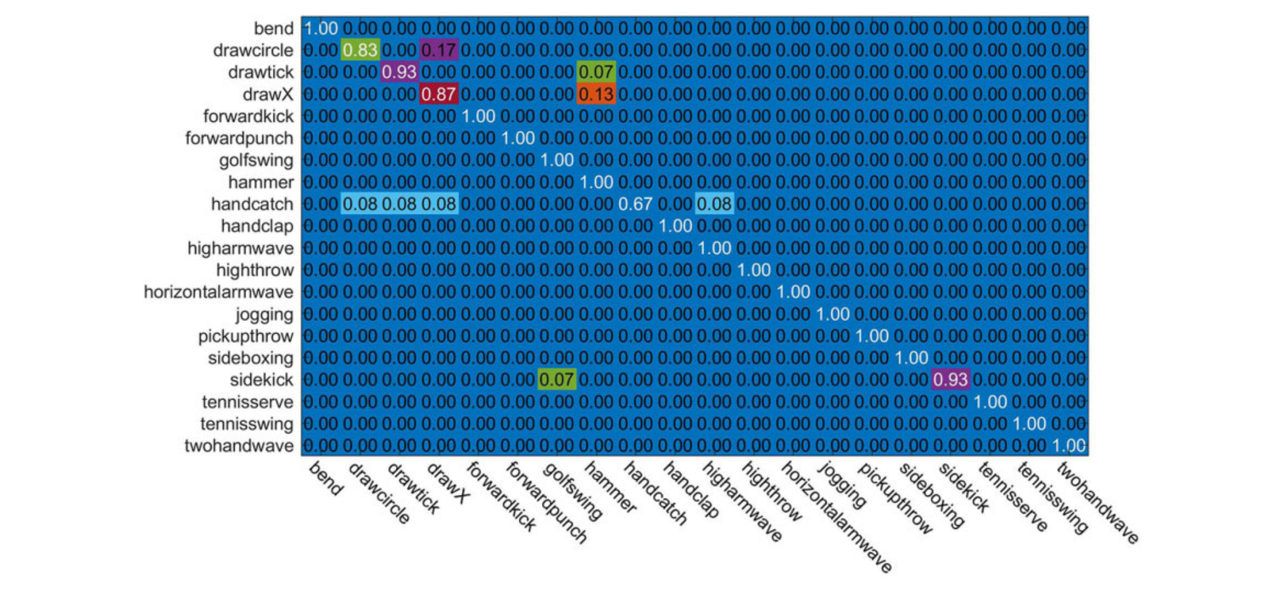
Table 1 shows the recognition accuracies of the proposed method with different classification schemes. The FC layer gets the best recognition result with the highest accuracy of 97.08%. Compared with the LC layer, it achieves an improvement of 3.44%, demonstrating that high-layer features in CNN are more effective for action recognition. Figure 8 is the confusion matrix of six actions in SDUFall data set with the FC layer classification scheme. From Figure 8, we observe that all actions are classified extremely correctly. The action lying and the action falling are also well classified although these two actions have great similarity. The superior experimental results prove that the proposed HDDPDI representation is effective and discriminative for human actions in the SDUFall data set. Moreover, action characteristics in three views are fused for recognition, which is helpful to improve the recognition performance by capturing the 3D motion information for actions.

MSRAction3D data set. We test the HDDPDI representation with the three classification schemes on MSRAction3D data set. For this data set, the cross-subject setting is used to get the training set and the testing set: samples of subjects 1, 3, 5, 7, 9 for training and samples of the remaining subjects for testing. The experimental results are shown in Table 2. From the table, it can be seen that the FC layer classification scheme also achieves the best recognition accuracy on MSRAction3D data set, which illustrates that the HDDPDI can

Table 2. Recognition accuracies on the MSRAction3D data set with different classification schemes in the proposed HDDPDI-based action recognition framework.

	Classification scheme				
Data set	LC layer	**FC layer**	Softmax-max	Softmax-average	Softmax-multiply
MSRAction3D	91.56%	**96.15%**	86.12%	86.12%	85.43%

Bold values denote the highest recognition accuracy and the corresponding classification scheme in MSRAction3D. HDDPDI: hierarchical dynamic depth projected difference images; LC: last convolutional; FC: fully connected.



	bend	drawcircle	drawtick	drawX	forwardkick	forwardpunch	golfswing	hammer	handcatch	handclap	higharmwave	highthrow	horizontalarmwave	jogging	pickupthrow	sideboxing	sidekick	tennisserve	tennisswing	twohandwave
bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
drawcircle	0.00	0.83	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
drawtick	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
drawX	0.00	0.00	0.00	0.87	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
forwardkick	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
forwardpunch	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
golfswing	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
hammer	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
handcatch	0.00	0.08	0.08	0.08	0.00	0.00	0.00	0.00	0.67	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
handclap	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
higharmwave	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
highthrow	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
horizontalarmwave	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
jogging	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
pickupthrow	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
sideboxing	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
sidekick	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00
tennisserve	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
tennisswing	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
twohandwave	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Figure 9. Confusion matrix on the MSRAction3D data set for the FC layer classification scheme. FC: fully connected.

Table 3. Recognition accuracies on the NTU RGB + D data set with different classification schemes in the proposed HDDPDI-based action recognition framework.

		Classification scheme				
Data set	Baseline	LC layer (%)	**FC layer** (%)	Softmax-max (%)	Softmax-average (%)	Softmax-multiply (%)
NTU RGB+D	Cross-subject	78.47	**82.43**	78.73	80.00	78.92
	Cross-view	83.11	**87.56**	84.19	86.00	84.76

Bold values denote the highest recognition accuracies and the corresponding classification scheme in NTU RGB+D. HDDPDI: hierarchical dynamic depth projected difference images; LC: last convolutional; FC: fully connected.

effectively capture the spatial–temporal dynamics of actions for improving the recognition performance significantly. The accuracy 96.15% demonstrates that abstract high-level features in FC layer are more discriminative for actions in this data set, and features fusion of three views is more effective compared with the classification results fusion in softmax layer.

The confusion matrix of the best classification scheme on MSRAction3D data set is shown in Figure 9. We can see that the most actions are recognized well except for several confused actions such as ‘draw circle’, ‘draw tick’ and ‘draw X’. Accuracies of these actions are relatively lower due to their similar HDDPDI representations.

NTU RGB + D data set. There are two evaluation criteria on the NTU RGB + D data set: cross-subject and cross-view. We evaluate the proposed HDDPDI video representation with three classification schemes on these two baselines, respectively. The training and testing sets are the same with the original protocol.[29] The results are shown in Table 3. From the table, we can conclude that the HDDPDI-based FC classification scheme still achieves the highest recognition accuracies of 82.43% in the cross-subject evaluation and 87.56% in the cross-view evaluation.

Figure 10 presents the confusion matrix of the FC classification scheme in the cross-subject evaluation. From the confusion matrix, we can see that most actions are

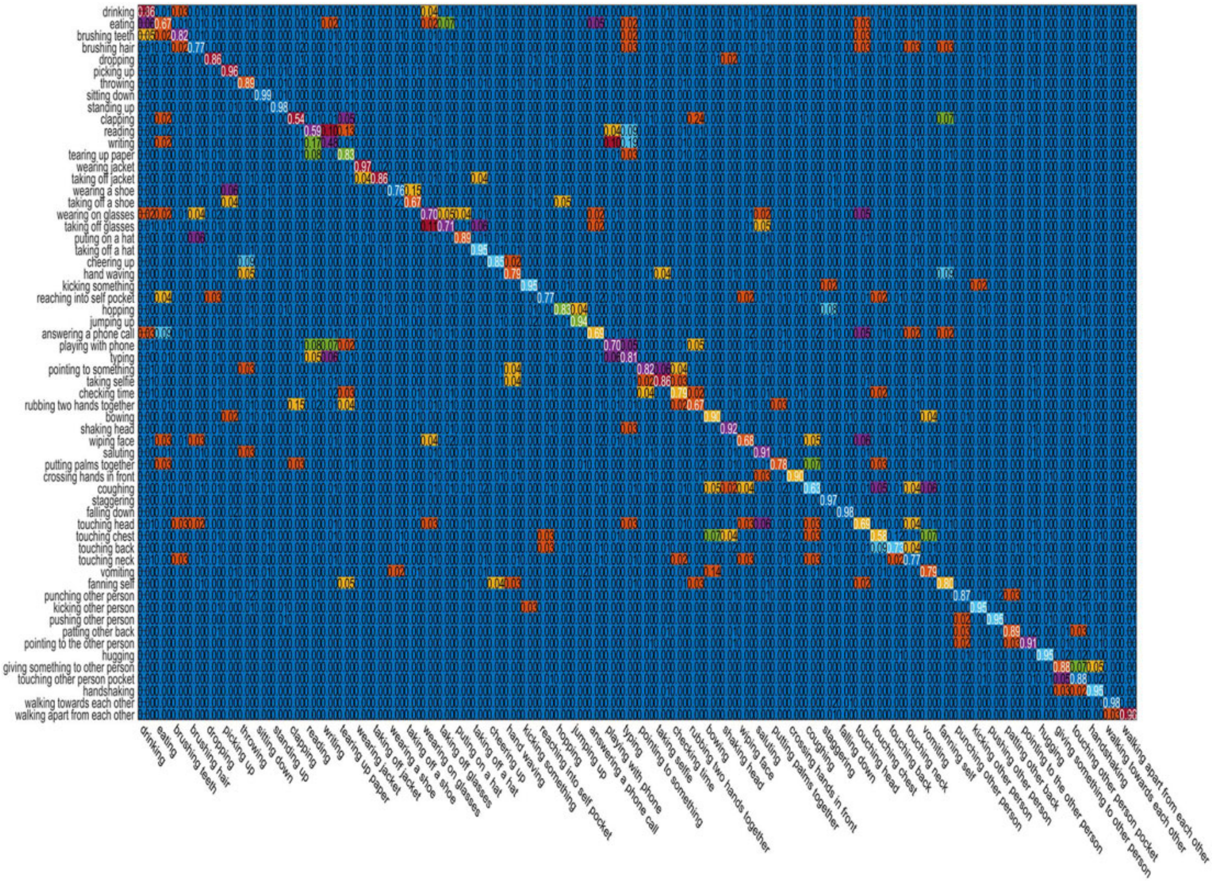


Figure 10. Confusion matrix on the NTU RGB + D data set for the FC layer classification scheme in the cross-subject evaluation. FC: fully connected.

recognized correctly including the mutual actions. And even for some close actions such as ‘wearing jacket’ and ‘taking off jacket’, ‘putting on a hat’ and ‘taking off a hat’, the proposed HDDPDI video representation with the FC classification scheme still achieves good results although only time orders are reversed for these actions. This demonstrates that the HDDPDI representation can well capture the temporal dynamics for actions in videos. However, our method cannot distinguish some actions that have similar motion changes, such as ‘clapping’ and ‘rubbing two hands together’. Moreover, since the objects in the actions are difficult to be recognized for depth videos and the HDDPDI representation is not discriminative enough for actions that contain fine-grained small motion changes, actions such as ‘reading’ and ‘writing’ are easily confused. Such cases may be improved by combining the information extracted from the RGB modality.

The proposed HDDPDI-based action recognition framework achieves the best experimental results with the FC classification scheme on all three data sets, which verifies the effectiveness of the HDDPDI video representation. Compared with the results of the LC classification scheme, it can be seen that the FC layer features of CNN are more discriminative for action recognition. Furthermore, since the HDDPDI representation of a depth video in one projected view contains several dynamic images at different temporal scales, and the softmax classification scheme processes the prediction class probabilities of all these dynamic images, misclassification of one dynamic image will affect the recognition result in this view. Therefore, the softmax classification scheme may cause misclassification more easily than the FC scheme that aggregates features for effective action recognition. Besides, Tables 1 to 3 show that the advantages of FC layer scheme over other mechanisms are bigger on MSRAction3D and NTU than on SDUFall. The main reason is that SDUFall data set is relatively small and contains only six simple human actions that have discriminative features. So, differences among the results of three classification schemes in SDUFall data set are small. However, actions in other two data sets are more complex and, especially in MSRAction3D, some human actions are similar and easily confused, which may increase the misclassification possibility of the softmax scheme that aggregates the recognition results directly.

Contribution evaluation of three projected views

For the construction of the HDDPDI video representation, we firstly project a depth video onto three views to capture the human 3D structure and shape information. Depth

Table 4. Comparisons of recognition accuracies on the three data sets with different classification schemes for the single-view recognition and the fusion-view recognition.

Classification scheme Dataset		LC layer (%)	FC layer (%)	Softmax-max (%)	Softmax-average (%)	Softmax-multiply (%)
SDUFall	Front	89.08	94.77	94.79	95.62	95.20
	Side	82.10	91.12	91.87	91.87	91.87
	Top	88.70	90.41	90.83	90.62	91.66
	Fusion	**93.64**	**97.08**	**95.83**	**96.25**	**96.04**
MSRAction3D	Front	80.17	86.41	80.37	80.94	80.94
	Side	72.15	79.68	73.92	73.04	72.15
	Top	74.68	81.82	79.30	79.73	79.30
	Fusion	**91.56**	**96.15**	**86.12**	**86.12**	**85.43**
NTU RGB + D (cross-subject)	Front	76.59	78.89	78.54	78.52	78.81
	Side	59.06	66.64	66.62	67.02	67.21
	Top	39.03	47.00	43.82	44.52	44.77
	Fusion	**78.47**	**82.43**	**78.73**	**80.00**	**78.92**
NTU RGB + D (cross-view)	Front	78.56	83.56	83.61	83.76	83.91
	Side	57.33	67.31	68.84	69.74	69.76
	Top	42.06	46.92	40.77	41.90	42.29
	Fusion	**83.11**	**87.56**	**84.19**	**86.00**	**84.76**

Bold values denote the highest recognition accuracies achieved by fusion-view method in different datasets. FC: fully connected; LC: last convolutional.

maps in three views describe the human motion from different perspectives. In the proposed HDDPDI-based action recognition framework, feature descriptors or softmax prediction results of three views are fused in different recognition schemes for expressing the spatial–temporal dynamic information of human actions more comprehensively. However, in this section, we evaluate the contribution of each projected view to the action recognition. We use the three classification schemes without fusion and conduct the recognition separately in each view. The training set and the testing set remain unchanged for the three data sets. We present the recognition results on the three data sets with different classification schemes for the single-view recognition and the fusion-view recognition in Table 4.

From Table 4, we can see that for the three data sets, action recognition with the three-view fusion outperforms the single view in each classification scheme. Since HDDPDIs in three projected views can describe the spatial–temporal motion dynamics from different perspectives and offer the complementary characteristics for human actions, multimodal information fusion is necessary and plays an important role for improving the recognition performance. For the single-view experiments, because most of the human actions are facing the camera, recognition results in the front view are the best for the three data sets in each classification scheme. HDDPDI in the front view are more discriminative for the most human actions, while the side view and the top view are more effective for actions such as 'forward punch' and 'forward kick'. In the NTU RGB + D data set, recognition accuracies in the top view are much lower than that in the side view. This is because the ground background of this data set produces much noise in depth maps of the top view.

Table 5. Performance comparison of the proposed HDDPDI-based action recognition with the FC layer classification scheme with the state-of-the-art methods on the SDUFall data set.

Method	Modality	Accuracy (%)
Shape feature encoding[48]	Depth	64.67
Slow feature analysis[49]	RGB	81.33
Silhouette orientation volumes[50]	Depth	89.63
HDDPDI-based recognition with FC layer classification scheme	Depth	**97.08**

Bold value denotes the highest recognition accuracy achieved by our method for SDUFall. HDDPDI: hierarchical dynamic depth projected difference images.

Comparison with the state of the arts

Tables 5, 6 and 7 compare the performance of the proposed method with the previous works, respectively, on the SDUFall data set, the MSRAction3D data set and the NTU RGB + D data set. From the tables, we can conclude that the FC layer classification scheme in the proposed HDDPDI-based action recognition framework outperforms those previous methods for all three data sets. The possible reasons are summarized as follows: (1) The HDDPDI representations of depth videos can describe the spatial–temporal motion dynamics for human actions from different temporal scales and contain rich action changing information that is effective for recognition. (2) FC layer in CNN provides the discriminative abstract features for different actions. (3) Three CNNs are finely retuned based on the pretrained models on ImageNet, which can ensure that the model parameters are well initialized for action classification. (4) HDDPDIs in three projected views offer 3D human motion information and the complementary features for

Table 6. Performance comparison of the proposed HDDPDI-based action recognition with the FC layer classification scheme with the state-of-the-art methods on the MSRAction3D data set.

Method	Modality	Accuracy (%)
Bag of 3D Points[47]	Depth	74.70
Actionlet ensemble[4]	Depth	82.22
DMM[3]	Depth	88.73
HON4D[5]	Depth	88.89
SNV[6]	Depth	93.09
Range sample[8]	Depth	95.62
HDDPDI-based recognition with FC layer classification scheme	Depth	**96.15**

Bold value denotes the highest recognition accuracy achieved by our method for MSRAction3D. HON4D: histogram of oriented 4D normal; DMM: depth motion map; SNV: super normal vector; FC: fully connected; HDDPDI: hierarchical dynamic depth projected difference images.

Table 7. Performance comparison of the proposed HDDPDI-based action recognition with the FC layer classification scheme with the state-of-the-art methods on the NTU RGB + D data set.

Method	Modality	Cross-subject accuracy (%)	Cross-view accuracy (%)
SNV[6]	Depth	31.82	13.61
HON4D[5]	Depth	30.56	7.26
Lie group[51]	Skeleton	50.08	52.76
HBRNN[42]	Skeleton	59.07	63.97
Skeletal quads[52]	Skeleton	38.62	41.4
Dynamic skeletons[53]	Skeleton	60.23	65.2
Two-layer RNN[29]	Skeleton	56.29	64.09
Two-layer LSTM[29]	Skeleton	60.69	67.29
Part-aware LSTM[29]	Skeleton	62.93	70.27
ST-LSTM[54]	Skeleton	69.20	76.10
DSSCA-SSLM[55]	RGB + depth	74.86	—
Different skeleton features encoding[20]	Skeleton	—	82.31
Clips + CNN + MTLN[18]	Skeleton	79.57	84.83
HDDPDI-based recognition with FC layer classification scheme	Depth	**82.43**	**87.56**

Bold values denote the highest recognition accuracies achieved by our method in cross-subject and cross-view respectively for NTU RGB+D. RNN: recurrent neural network; HBRNN: hierarchically bidirectional RNN; LSTM: long-short term memory; ST-LSTM: spatio-temporal LSTM; DSSCA-SSLM: deep shared-specific component analysis-structure sparsity learning machine; MTLN: multi-task learning network; HON4D: histogram of oriented 4D normal; HDDPDI: hierarchical dynamic depth projected difference images; FC: fully connected; CNN: convolutional neural network; SNV: super normal vector.

actions. Therefore, fusion of three views can improve the action recognition performance significantly.

For the SDUFall data set, human actions change significantly. HDDPDI can well capture the large body motion changes, so the recognition results in this data set are extremely superior. From the Table 7, we can see that our method outperforms the previous works significantly in the NTU RGB + D data set, which verifies that the proposed video representation is effective for describing the spatial–temporal information. However, in the MSRAction3D data set, our method is only slightly higher than the range sample.[8] This is because that our method is limited to differentiate some similar actions containing fine-grained small motions, such as 'draw tick' and 'draw X'. Besides, our method is applied on depth videos where the absence of colour and texture may reduce the discriminative power of CNN models which are more suitable for texture-based feature learning and classification.

Conclusion and future work

In this article, we propose the HDDPDI representation for a depth video to describe the spatial–temporal dynamics of human actions from different temporal scales. We project a depth video sequence onto three orthogonal planes to capture the 3D human shape and motion information. HDDPDI are produced in each projected view by applying rank pooling on DPDI sequences at different sampling temporal scales. In addition, we construct a HDDPDI-based action recognition framework that contains three classification schemes to verify the effectiveness of the HDDPDI representation. Information of three views is fused for recognition in each scheme. We test the framework on three publicly available data sets and compare the recognition results of the three classification schemes. Presented experimental results show that the HDDPDI representation is efficient and practicable for human action recognition.

Although the proposed HDDPDI-based action recognition with the FC classification scheme has achieved outstanding results, there are some limitations of the work that need to be solved in the future. Firstly, HDDPDI representation performs better for actions containing significant human motions. For some similar actions that have tiny motion variations such as 'read' and 'write', HDDPDI representation is not discriminative enough. Therefore, to differentiate these actions, colour and texture features extracted from the static RGB images are considered to be combined with the HDDPDI representation. Secondly, feature descriptors in the three views for the LC\FC classification scheme or prediction results for the softmax classification scheme are fused for action recognition. But how to fuse multimodal information effectively is still a challenging problem in our future work.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of following financial support for the research, authorship, and/or publication of this article: This study has been funded with the National High-tech Research and Development (863 Program) of China under Grant No.

2015AA042307 and the National Natural Science Foundation of China under Grant No. U1706228 and No. 61673245.

ORCID iD

Xin Ma https://orcid.org/0000-0003-4402-1957

References

1. Liu J, Ali S, and Shah M. Recognizing human actions using multiple features. In: *IEEE conference on computer vision and pattern recognition*, Anchorage, Alaska, USA, 23–28 June 2008, pp. 1–8. IEEE Computer Society.
2. Raptis M and Sigal L. Poselet key-framing: a model for human activity recognition. In: *Proceedings of the IEEE conference on Computer vision and pattern recognition*, Portland, Oregon, USA, 23–28 June 2013, pp. 2650–2657. IEEE Computer Society.
3. Yang X, Zhang C, and Tian Y. Recognizing actions using depth motion maps-based histograms of oriented gradients. In: *Proceedings of the 20th ACM international conference on Multimedia*, Nara, Japan, 29 October–2 November 2012, pp. 1057–1060. ACM.
4. Wang J, Liu Z, Wu Y, et al. Mining actionlet ensemble for action recognition with depth cameras. In: *2012 IEEE conference on Computer vision and pattern recognition (CVPR)*, Providence, Rhode Island, USA, 16–21 June 2012, pp. 1290–1297. IEEE Computer Society.
5. Oreifej O and Liu Z. HON4D: histogram of oriented 4D normals for activity recognition from depth sequences. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Portland, Oregon, USA, 23–28 June 2013, pp. 716–723. IEEE Computer Society.
6. Yang X and Tian Y. Super normal vector for activity recognition using depth sequences. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Columbus, OH, USA, 23–28 June 2014, pp. 804–811. IEEE Computer Society.
7. Xia L and Aggarwal J. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Portland, OR, USA, 23–28 June 2013, pp. 2834–2841. IEEE Computer Society.
8. Lu C, Jia J, and Tang C-K. Range-sample depth feature for action recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Columbus, OH, USA, 23–28 June 2014, pp. 772–779. IEEE Computer Society.
9. Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Columbus, OH, USA, 23–28 June 2014, pp. 1725–1732. IEEE Computer Society.
10. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis* 2015; 115: 211–252.
11. LeCun Y, Kavukcuoglu K, and Farabet C. Convolutional networks and applications in vision. In: *IEEE international symposium on circuits and systems: nano-bio circuit fabrics and systemsI*, Paris, France, 30 May–2 June 2010, pp. 253–256. IEEE Computer Society.
12. Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks. In: *Proceedings of the IEEE international conference on computer vision*, Santiago, Chile, 7–13 December 2015, pp. 4489–4497. IEEE Computer Society.
13. Simonyan K and Zisserman A. Two-stream convolutional networks for action recognition in videos. In: *Advances in neural information processing systems*, Montréal, Canada, 8–13 December 2014, pp. 568–576. MIT Press.
14. Zhang B, Wang L, Wang Z, et al. Real-time action recognition with enhanced motion vector CNNs. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 27–30 June 2016, pp. 2718–2726. IEEE Computer Society.
15. Wang L, Qiao Y, and Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, MA, USA, 7–12 June 2015, pp. 4305–4314. IEEE Computer Society.
16. Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition. In: *European conference on computer vision*, Amsterdam, Netherlands, 11–14 October 2016, pp. 20–36. Springer.
17. Du Y, Fu Y, and Wang L. Skeleton based action recognition with convolutional neural network. In: *The 3rd IAPR Asian Conference on Pattern Recognition*, Kuala Lumpur, Malaysia, 3–6 November 2015, pp. 579–583. IEEE Computer Society.
18. Ke Q, Bennamoun M, An S, et al. A new representation of skeleton sequences for 3D Action Recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017, pp. 4570–4579. IEEE Computer Society.
19. Ke Q, An S, Bennamoun M, et al. SkeletonNet: mining deep part features for 3-D action recognition. *IEEE Signal Process Lett* 2017; 24: 731–735.
20. Ding Z, Wang P, Ogunbona PO, et al. Investigation of different skeleton features for CNN-based 3D action recognition. In: *2017 IEEE international conference on multimedia & expo workshops (ICMEW)*, Hong Kong, China, 10–14 July 2017, pp. 617–622. IEEE Computer Society.
21. Hou Y, Li Z, Wang P, et al. Skeleton optical spectra-based action recognition using convolutional neural networks. *IEEE Trans Circuits Syst Video Technol* 2018; 28(3): 807–811.
22. Soomro K, Zamir AR, and Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:12120402 2012.
23. Kuehne H, Jhuang H, Stiefelhagen R, et al. Hmdb51: a large video database for human motion recognition. In: *High performance computing in science and engineering '12*. Berlin: Springer, 2013, pp. 571–582. IEEE Computer Society.
24. Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset. arXiv preprint arXiv:170506950 2017.
25. Wang P, Li W, Gao Z, et al. Action recognition from depth maps using deep convolutional neural networks. *IEEE Trans Hum Mach Syst* 2016; 46: 498–509.

26. Wang P, Li W, Gao Z, et al. Depth pooling based large-scale 3-D action recognition with convolutional neural networks. *IEEE Trans Multimedia* 2018; 20: 1051–1061.
27. Chen J, Xiao Y, Cao Z, et al. Action recognition in depth video from RGB perspective: a knowledge transfer manner. In: *Conference on pattern recognition and computer vision*, Guangzhou, China, 23–25 November 2018, pp. 1060916–1060911. Proceedings of SPIE.
28. Imran J and Kumar P. Human action recognition using RGB-D sensor and deep convolutional neural networks. In: *2016 international conference on advances in Computing, Communications and Informatics (ICACCI)*, Jaipur, India, 21–24 September 2016, pp. 144–148. IEEE Computer Society.
29. Shahroudy A, Liu J, Ng TT, et al. NTU RGB + D: a large scale dataset for 3D human activity analysis. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.
30. Bilen H, Fernando B, Gavves E, et al. Dynamic image networks for action recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3034–3042.
31. Fernando B, Gavves E, Oramas J, et al. Rank pooling for action recognition. *IEEE Trans Pattern Anal Mach Intell* 2017; 39: 773–787.
32. Liu H, Yu Y, Sun F, et al. Visual–tactile fusion for object recognition. *IEEE Trans Autom Sci Eng* 2017; 14: 996–1008.
33. Liu H, Wu Y, Sun F, et al. Weakly paired multimodal fusion for object recognition. *IEEE Trans Autom Sci Eng* 2018; 15: 784–795.
34. Liu H, Deng C, Fernández-Caballero A, et al. *Multimodal fusion for robotics*. London: SAGE, 2018.
35. Chen L, Wei H, and Ferryman J. A survey of human motion analysis using depth imagery. *Pattern Recognit Lett* 2013; 34: 1995–2006.
36. Herath S, Harandi M, and Porikli F. Going deeper into action recognition: a survey. *Image Vis Comput* 2017; 60: 4–21.
37. Wang P, Li W, Ogunbona P, et al. RGB-D-based human motion recognition with deep learning: a survey. *Computer Vision & Image Understanding* 2018; 171: 118–139.
38. Chen C, Jafari R, and Kehtarnavaz N. Action recognition from depth sequences using depth motion maps-based local binary patterns. In: *2015 IEEE winter conference on applications of computer vision (WACV)*, Waikoloa, HI, USA, 5–9 January 2015, pp. 1092–1099. IEEE Computer Society.
39. Song Y, Morency L-P, and Davis R. Action recognition by hierarchical sequence summarization. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Portland, OR, USA, 23–28 June 2013, pp. 3562–3569. IEEE Computer Society.
40. Yamato J, Ohya J, and Ishii K. Recognizing human action in time-sequential images using hidden Markov model. In: *1992 proceedings CVPR'92, 1992 IEEE computer society conference on computer vision and pattern recognition*, Champaign, IL, USA, 15–18 June 1992, pp. 379–385. IEEE Computer Society.
41. Wang J, Liu Z, Wu Y, et al. Learning actionlet ensemble for 3D human action recognition. *IEEE Trans Pattern Anal Mach Intell* 2014; 36: 914–927.
42. Du Y, Wang W, and Wang L. Hierarchical recurrent neural network for skeleton based action recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, MA, USA, 7–12 June 2015, pp. 1110–1118. IEEE Computer Society.
43. Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, MA, USA, 7–12 June 2015, pp. 2625–2634. IEEE Computer Society.
44. Simonyan K and Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556 2014.
45. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in pytorch. In: *NIPS autodiff workshop: the future of gradient-based machine learning software and techniques*, Long Beach, CA, USA, 9 December 2017.
46. Ma X, Wang H, Xue B, et al. Depth-based human fall detection via shape features and improved extreme learning machine. *IEEE J Biomed Health Inform* 2014; 18: 1915–1922.
47. Li W, Zhang Z, and Liu Z. Action recognition based on a bag of 3D points. In: *2010 IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW)*, San Francisco, CA, USA, 13–18 June 2010, pp. 9–14. IEEE Computer Society.
48. Aslan M, Sengur A, Xiao Y, et al. Shape feature encoding via fisher vector for efficient fall detection in depth-videos. *Appl Soft Comput* 2015; 37: 1023–1028.
49. Fan K, Wang P, and Zhuang S. Human fall detection using slow feature analysis. *Multimed Tools Appl* 2018; 77: 1–28.
50. Akagündüz E, Aslan M, Şengür A, et al. Silhouette orientation volumes for efficient fall detection in depth videos. *IEEE J Biomed Health Inform* 2017; 21: 756–763.
51. Vemulapalli R, Arrate F, and Chellappa R. Human action recognition by representing 3D skeletons as points in a lie group. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Columbus, OH, USA, 23–28 June 2014, pp. 588–595. IEEE Computer Society.
52. Evangelidis G, Singh G, and Horaud R. Skeletal quads: human action recognition using joint quadruples. In: *The 22nd international conference on pattern recognition*, Stockholm, Sweden, 24–28 August 2014, pp. 4513–4518. IEEE Computer Society.
53. Hu JF, Zheng WS, Lai J, et al. Jointly learning heterogeneous features for RGB-D activity recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, MA, USA, 7–12 June 2015, pp. 5344–5352. IEEE Computer Society.
54. Liu J, Shahroudy A, Xu D, et al. Spatio-temporal LSTM with trust gates for 3D human action recognition. In: *European conference on computer vision*, Amsterdam, Netherlands, 11–14 October 2016, pp. 816–833. Springer.
55. Shahroudy A, Ng TT, Gong Y, et al. Deep multimodal feature analysis for action recognition in RGB + D videos. *IEEE Trans Pattern Anal Mach Intell* 2017; 40(5): 1045–1058.