Neurocomputing 175 (2016) 747–758

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Depth Context: a new descriptor for human activity recognition by using sole depth sequences

Mengyuan Liu, Hong Liu*

Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 6 June 2015
Received in revised form
18 September 2015
Accepted 2 November 2015
Communicated by L. Shao
Available online 12 November 2015

Keywords:
Activity recognition
Bag-of-Visual-Words
Depth data
Shape context

ABSTRACT

Human activity recognition using sole depth information from 3D sensors achieves superior performances to tackle light changes and cluttered backgrounds than using RGB sequences from traditional cameras. However, the noises and occlusions in depth data, which are common problems for 3D sensors, are not well handled. Moreover, many existing methods ignore the strong contextual information from depth data, resulting in limited performances on distinguishing similar activities. To deal with these problems, a local point detector is developed by sampling local points based on both motion and shape clues to represent human activities in depth sequences. Then a novel descriptor named Depth Context is designed for each local point to capture both local and global contextual constrains. Finally, a Bag-of-Visual-Words (BoVW) model is applied to generating human activity representations, which serve as the inputs for a non-linear SVM classifier. State-of-the-art results namely 94.28%, 98.21% and 95.37% are achieved on three public benchmark datasets: MSRAction3D, MSRGesture3D and SKIG, which show the efficiency of proposed method to capture structural depth information. Additional experimental results show that our method is robust to partial occlusions in depth data, and also robust to the changes of pose, illumination and background to some extent.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Human activity recognition based on action sequences has been a core topic in content-based video analysis and intelligent surveillance for decades [1–4], while it is still challenging due to light changes and other common difficulties in video analysis like cluttered backgrounds.

With the advance of imaging technology in capturing depth information in real time, researchers are focusing on utilizing depth data to solve previous problems. Based on Kinect sensor which generates depth sequences, many applications have been developed [5–8]. Compared with conventional RGB data, depth data is more robust to intensive light changes, since the depth value is estimated by infrared radiation and is not related to visible light [9]. Subtracting foregrounds from cluttered backgrounds is also much easier using depth sequences, as the confusing texture and color information from cluttered background are ignored [10].

Common pipeline for human activity recognition includes feature detection, feature encoding, feature pooling and feature classification. Over past decades, various features have been developed, which can be divided into two categories: holistic feature and local feature. Specially for human activity recognition using depth sequences, two holistic features namely *depth motion maps*, *skeleton joints* and two local features namely *surface normals*, *cloud points* have been widely used.

The main idea of developing *depth motion maps* is to find proper projection methods to convert depth sequences into several 2D maps. Yang et al. [11] project depth maps to orthogonal planes and accumulate motions for each plane to obtain the depth motion maps. Then the histograms of oriented gradients (HOG) [12] are computed for these maps as human activity representation. Inspired by motion history images [13], Azary et al. [14] provide motion depth surfaces to track the motion of depth map changes, which serve as inputs for a subspace learning algorithm. These methods [11,14] are effective to encode both body shape and motion information. However, depth motion maps are not robust against partial occlusions, since they belong to the category of holistic feature encoding information from both actors and occlusions.

Human activities can be denoted by the movements of *skeleton joints*, which are distinctive to similar activities. These skeleton joints are recorded by multi-camera motion capture (MoCap) systems [15] or estimated by the OpenNI tracking

* Corresponding author at: Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, Beijing 100871, China.
E-mail addresses: liumengyuan@pku.edu.cn (M. Liu), hongliu@pku.edu.cn (H. Liu).

http://dx.doi.org/10.1016/j.neucom.2015.11.005
0925-2312/© 2015 Elsevier B.V. All rights reserved.

framework [6], where the skeleton joints from MoCap systems are more accurate than that from OpenNI, despite that MoCap systems are marker-based and much more expensive. Taking compromise in accuracy and price, many skeleton joints based features are developed based on the OpenNI framework [16–19]. Yang et al. [16] adopt the differences of joints in temporal and spatial domains to encode the dynamics of joints, and then obtain the EigenJoints by applying Principal Component Analysis (PCA) to joint differences. The EigenJoints contain less redundancy and noises, compared with original joints. Zanfir et al. [17] provide a non-parametric Moving Pose (MP) framework, which considers more features like position, speed and acceleration of joints. To insure precision of the estimated joints, Wang et al. [18] incorporate temporal constraints and additional segmentation cues of consecutive skeleton joints to select the K-best joint estimations. Another way to improve performances of skeleton joints is to associate local features with joints. This idea is named Actionlet Ensemble Model by Wang et al. [20], which combines local occupancy pattern [21] with 3D joints. Pairwise relative positions of skeleton joints are also utilized in [20], which are more discriminative and intuitive than previous skeleton joints based features. Additionally, Luo et al. [19] reduce the irrelevant information of Pairwise skeleton joints feature in [20], and propose a 3D joint feature which selects one joint as reference and uses its differences to the rest joints as features. Beyond [20] and [19], Rahmani et al. [22] encode spatio-temporal depth and depth gradient histograms of local volume around each joint, and convert the 3D joint feature [20] into joint displacement vectors. Both local features and joint displacement vectors are concatenated for classification, which outperform either of them. Despite the efficiency of representing human activities by skeleton joints, these estimated joints may not be accurate in the presence of occlusions or when people are not directly facing camera in upright poses [23]. What is worse, these methods do not work [22] in applications like hand gesture recognition, where joint positions cannot be obtained.

Intuitively, *surface normals* reflect the shape of 3D objects. When human activities are treated as space–time pattern templates [24], the task of human activity classification is converted to 3D object recognition, and surface normals can be utilized for representing human activities [25–27]. Tang et al. [25] form a Histogram of Oriented Normal Vectors (HONV) as a concatenation of histograms of zenith and azimuthal angles to capture local distribution of the orientation of an object surface. Oreifej et al. [26] extend HONV to 4D space of time, depth and spatial coordinates, and provide a Histogram of Oriented 4D Normals (HON4D) to encode the surface normal orientation of human activities. HON4D jointly captures the distribution of motion cues and the dynamic shapes, therefore it is more discriminative than previous approaches which separately encode the motion or shape information. To increase the robustness of HON4D against noise, Yang et al. [27] group local hypersurface normals into polynormal, and aggregate low-level polynormals into the Super Normal Vector (SNV). Surface normals are utilized as local features of activities, which show robustness to occlusions.

Unlike above three features, *cloud points*, which denote human activities as a cloud of local points, are suitable to tackle both partial occlusions and the noise of original depth data. Li et al. [28] extract points from the contours of planar projections of 3D depth map, and employ an action graph to model the distribution of sampled 3D points. Vieira et al. [29] divide 3D points into same size of 4D grids, and apply spatio-temporal occupancy patterns to encode these grids. Wang et al. [21] explore an extremely large sampling space of random occupancy pattern features, and use a sparse coding method to encode these features. Generally speaking, cloud points based methods depend on local features which are robust against partial occlusions. When part of features are destroyed by partial occlusions, the rest of local features are still useful to represent human activities. However, previous works [28,29,21] ignore the global constrains among points and are not distinctive to classify human activities with similar local features.

Nevertheless, previous works focus on exploring holistic or local information separately, ignoring the complementary properties between them. In this work, we develop a Depth Context descriptor, which jointly captures local and global distributions of depth information. This descriptor is inspired by shape context descriptor [30], which is widely used in shape matching and object recognition. Depth Context improves the original shape context, which records the distribution of local points, by encoding the distribution of relative depth values. In previous work [31], original shape context is extended to 3D shape context for human activity recognition. Recently, Zhao et al. [32] present an optimized version of 3D shape context to characterize the distribution of local points. Since these works [31,32] treat human activity as a cloud of 3D local points, different speeds of actors may result in different spatio-temporal distributions of local points. To eliminate the effect of variant speeds, Depth Context ignores the relationships among different frames and encodes the layout of depth information on each frame. Unlike works [30–32] where all detected local points are encoded, we choose a subset of local points with strong motion information for encoding. Since more informative local points are encoded, our final representations show more discriminative power than previous works.

2. Overview of the proposed framework

The pipeline of our human activity recognition framework is shown in Fig. 1, where human activities are treated as sequences of postures changing over time. These postures are described frame by frame, ignoring the temporal relationships of postures among different frames.

In each frame, the posture is described by two local point sets: “reference points” and “target points”. The “reference points” is constructed by local points with salient shape information. From “reference points”, local points with salient motion information are selected to form the “target points”. Intuitively speaking, “reference points” and “target points” respectively reflect the shape and moving regions of the posture. As shown in the second column of Fig. 1, all green points belong to the “reference points”, and a subset of green points with white backgrounds belongs to the “target points”. Note that the white backgrounds stand for the moving regions which are extracted by figuring out the differences between current frame and previous one.

Then, each local point in “target points” is described by a Depth Context descriptor which encodes the spatial relationships between the local point and all local points in “reference points”. In this way, a human activity sequence is represented by a collection of Depth Context descriptors across all frames, which is shown in the third column of Fig. 1.

Afterwards, a classic Bag-of-Visual-Words (BoVW) model is applied to summarize a human activity representation from Depth Context descriptors. As in the training part of BoVW model, we sample a certain number of Depth Context descriptors and cluster

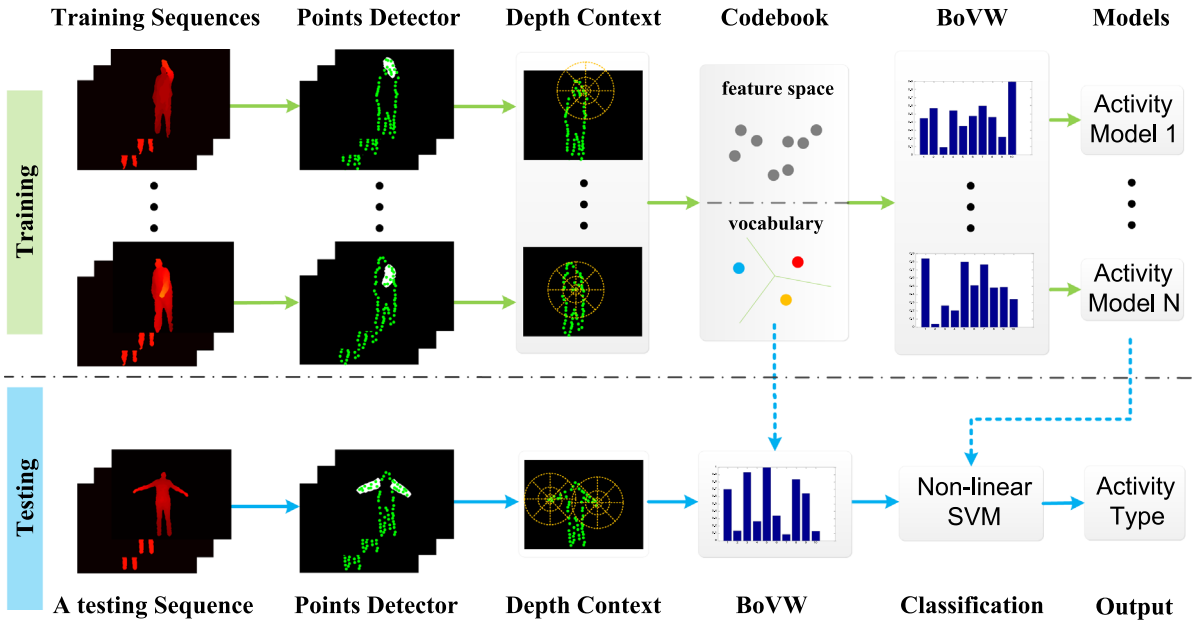

Fig. 1. The pipeline of our human activity recognition framework. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

them by k-means to construct a visual vocabulary. The size of clusters is empirical defined, and each cluster center is called a "word". Then each descriptor is assigned to the closest (we use Euclidean distance) vocabulary "word", and a human activity which contains a collection of descriptors is represented by a histogram of visual "words". As shown in the fourth column of Fig. 1, each gray point in feature space means a Depth Context descriptor, and each colored point means one "word". In the last column of Fig. 1, the action model for each type of human activity is a collection of histograms, which are generated by operating BoVW model on training sequences containing certain type of activity.

Finally, we apply a non-linear SVM classifier to compare the histogram of testing sequence with activity models to decide the activity type. Among above steps, our main work lies in designing a local point detector and a Depth Context descriptor, which will be respectively detailed in Sections 3 and 4. Our human activity recognition framework is constructed in Section 5. Experiments are then conducted in Section 6 to evaluate our framework. In the end, conclusions and future works are drawn in Section 7. Besides, Table 1 demonstrates the symbols used in this paper.

The main contributions of this paper are as follows:

First, we detect local interest points to involve strong shape and motion clues of human activities. Second, a Depth Context descriptor is designed to jointly encode both local and global information. Third, our method is robust to partial occlusions and shows more discriminative power than state-of-the-art methods on three benchmark datasets: MSRAction3D [28], MSRGesture3D [21] and SKIG [33].

Table 1
Definition of symbols.

Symbol	Definition
$\mathcal{I}$	A depth sequence
I_t	The tth frame of $\mathcal{I}$
X, Y	The height and width of I_t
$I_t(x,y)$	The depth value of location (x,y) on I_t
G_t	Shape information of I_t
g_1	Threshold value for G_t
$\mathcal{G}_t$	A local point set generated from G_t
M_t	Motion information from I_t
g_2	Threshold value for M_t
g_3	Minimum area of connected region in binary map
$\mathcal{M}_t$	A local point set generated from M_t
B	A "diamond" structure with parameter 7
$\mathcal{R}_t$	The reference point set of I_t
$\mathcal{P}_t$	The target point set of I_t
$\mathcal{P}_t^j$	The jth local point in $\mathcal{P}_t$
$N_{\mathcal{P}_t}$	Total number of local points in $\mathcal{P}_t$
$\mathcal{R}_t^i$	The ith local point in $\mathcal{R}_t$
$N_{\mathcal{R}_t}$	Total number of local points in $\mathcal{R}_t$
L	Maximum number of local points in each frame
α	Mean distance between all local point pairs in $\mathcal{R}_t$
$I_t(\mathcal{P}_t^j)$	The depth value of point $\mathcal{P}_t^j$ on I_t
$\delta_{\mathcal{P}_t^j,\mathcal{R}_t^i}$	Indicate the type of relative depth feature
K	The number of bins in Depth Context descriptor
k	The kth bin
$\zeta^k_{\mathcal{P}_t^j,\mathcal{R}_t^i}$	A relative depth feature in the kth bin
$h^k_{\mathcal{P}_t^j,+}$	Accumulation of positive relative depth feature in the kth bin for point $\mathcal{P}_t^j$
$h^k_{\mathcal{P}_t^j,-}$	Accumulation of negative relative depth feature in the kth bin for point $\mathcal{P}_t^j$
$\hat{h}_{\mathcal{P}_t^j}$	The Depth Context descriptor of point $\mathcal{P}_t^j$
C	The size of codebook for BoVW model
H	The human activity representation

3. Local point detector

Local features are robust to shelters and can be extracted without pre-processing such as segmentation or tracking, compared with global features. Human action classification methods based on local features and a Bag-of-Visual-Words (BoVW) model have shown promising results in processing RGB sequences [1,34,35]. These methods firstly extract spatio-temporal interest points (STIPs) from training videos and cluster STIPs into "words". Then, BoVW model is adopted to describe the original sequence by a histogram of "words", which is utilized to train classifiers for

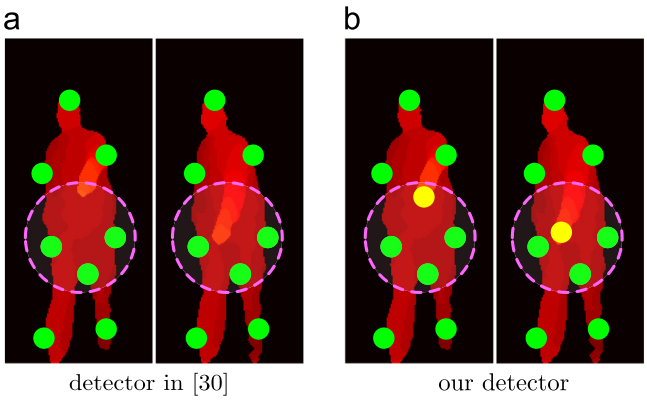

Fig. 2. Comparison of two local interest point detectors. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

classification. The common STIPs detectors for RGB sequences range from Harris3D detector [36], Hessian detector [37] to cuboid detector [1]. Recently, Xia and Aggarwal [38] extend the cuboid detector [1] to form a new DSTIP detector to extract STIPs from depth sequences. Specially for describing depth sequences, Li et al. [28] sample local points from the contours of planar projections of 3D depth map. Different from works in [38,28] which concern either obvious motion or strong shape information, our local point detector considers both clues and thus harvests local points with more details about human activities.

Our local point detector is inspired by Belongie et al. [30], which is originally designed for describing 2-D shapes of objects. In [30], points are sampled randomly from the contours of objects, and all points are treated equally for description. Differently, we sample local points with strong shape information and further select from these detected points to find local points with strong motion information. Only those points with both salient shape and motion information are utilized as inputs for description. To illustrate the differences between the detector in [30] and our local point detector, local points detected from a same human activity sequence by these two methods are shown in Fig. 2, where the human activity is illustrated by two representative frames. Note that the activity means one person puts down his right hand in front of the body. As shown in Fig. 2(a), the contours of two frames keep nearly unchanged, and points selected from the contours by [30] also keep unchanged across two frames. Thus, points from [30] cannot encode the motion. In our method, we select points with rich shape information to encode the activity. These points include green and yellow points in Fig. 2(b). Focusing on the points located in the dashed circle of Fig. 2(b), the location of yellow point changes which denotes the hand movement across two frames. Rather than encoding all points (i.e. the green and yellow point in Fig. 2(b)), only points with salient motion (i.e. the yellow point in Fig. 2(b)) are encoded. Obviously, yellow point shows more distinctive power to represent the "hand moving" than the green points. Without encoding the green points, redundant information is ignored. Therefore, activities with similar redundant information can be distinguished more easily.

Let $\mathcal{I}=\{I_t\}_{t=1}^{T}$ denote a depth sequence which contains T frames and $I_t=\{I_t(x,y)|x\in(1,\ldots,X),y\in(1,\ldots,Y)\}$ denote the tth frame. Note that $I_t(x,y)$ means the depth value in position (x,y) of frame I_t, and X,Y are the height and width of I_t.

For the tth frame, we describe the shape information of each point by G_t in Formula (1), considering the gradient values of I_t in both vertical and horizontal directions:

$$G_t=\sqrt{\left(\frac{\partial I_t}{\partial x}\right)^2+\left(\frac{\partial I_t}{\partial y}\right)^2}, \tag{1}$$

where points with larger values in G_t contain stronger shape information. To eliminate the effect of noises, a threshold g_1 is established to obtain a binary map $\mathcal{G}_t$ in Formula (2), which represents the shape of human posture in the tth frame:

$$\mathcal{G}_t=G_t>\left(\max(G_t)\cdot g_1\right). \tag{2}$$

For the tth frame, we describe the motion information of each point by M_t in Formula (3), which is obtained by figuring out the difference of current frame with previous frame:

$$M_t=|I_t-I_{t-1}|,\quad \text{s.t. } t\in(2,\ldots,T), \tag{3}$$

where points with larger values in M_t contain more salient motion information. To eliminate the effect of light changes and random noises, a threshold g_2 is utilized to obtain a binary map $\mathcal{M}_t$ as follows:

$$\mathcal{M}_t=\left(\mathcal{D}(M_t>g_2)\right)\oplus B, \tag{4}$$

where operator $\mathcal{D}$ removes small connected regions, whose areas are less than a threshold g_3, from a given binary map, $\oplus$ is a math operator for dilation in the field of mathematical morphology [39], B is a flat structuring element for operator $\oplus$ which is chosen as a "diamond" structure with parameter 7.

Taking computation efficiency into consideration, we sample points from $\mathcal{G}_t$ and $\mathcal{M}_t$ to denote the shape and motion of the tth frame. By applying a random sampling method on $\mathcal{G}_t$, we obtain a local point subset $\mathcal{R}_t$ as:

$$\mathcal{R}_t=\mathcal{S}(\mathcal{G}_t), \tag{5}$$

where $\mathcal{S}$ denotes Jitendra's random sampling method[1] [30]. Using the motion information from $\mathcal{M}_t$, we obtain a local point subset $\mathcal{P}_t$ as:

$$\mathcal{P}_t=\mathcal{R}_t\cap\mathcal{M}_t, \tag{6}$$

where operator $\cap$ reserves those local points, which are contained by $\mathcal{R}_t$ and the corresponding values in $\mathcal{M}_t$ are bigger than zero.

The pipeline of local point detector is shown in Fig. 3 where two adjacent depth frames are used as inputs. Refined shape map $\mathcal{G}_t$ and motion map $\mathcal{M}_t$ are illustrated in Fig. 3(b) and (d) respectively. Local points in $\mathcal{R}_t$, which are sampled from refined shape map $\mathcal{G}_t$, are shown in Fig. 3(e). The points in white area of Fig. 3(f) form the local point set $\mathcal{P}_t$, and these points are to be encoded by Depth Context descriptor in the next section. Generally speaking, local points in $\mathcal{R}_t$ provide contextual information for local points in $\mathcal{P}_t$, and all points in $\mathcal{P}_t$ are encoded to represent human posture in the tth frame.

[1] Code in http://www.eecs.berkeley.edu/Research/Projects/CS/vision/code/.

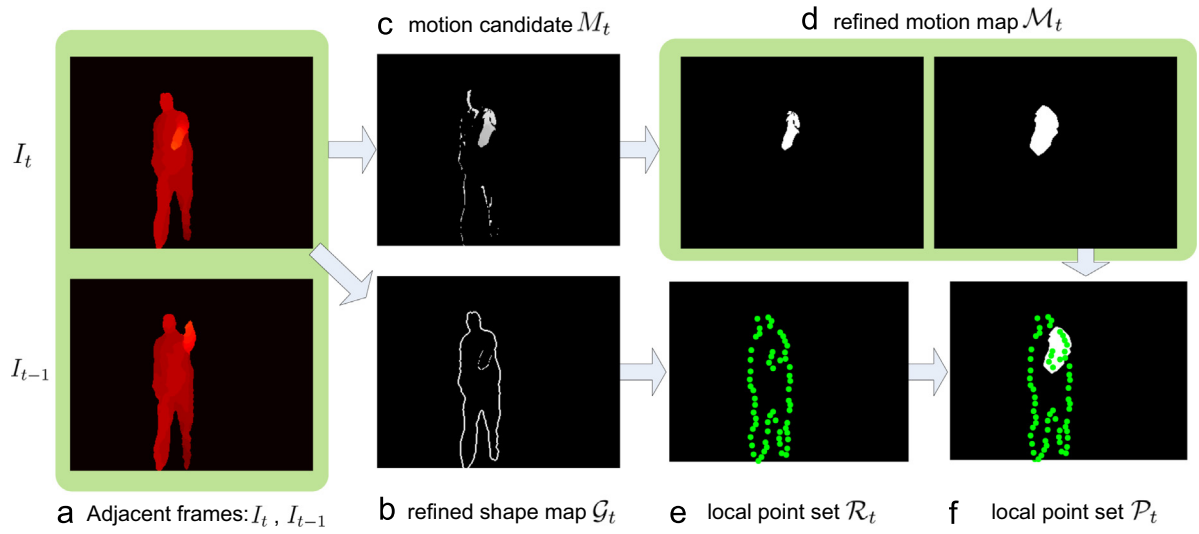

Fig. 3. The pipeline of our local point detector.

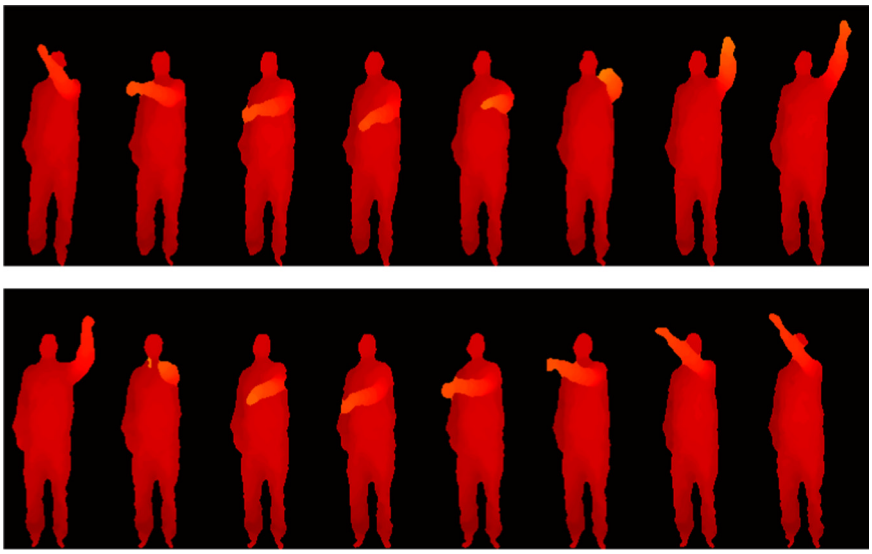

Fig. 4. "drawX" and "drawTick" in MSRAction3D dataset.

4. Depth Context descriptor

To describe the distribution of a cloud of points, Belongie et al. [30] propose a shape context descriptor to encode the relative coordinates of each point with all the remaining points. The shape context is empirically demonstrated to be robust to outliers, noise and slight deformations. In this section, we improve the basic shape context descriptor, and propose a Depth Context descriptor to involve both distribution and depth information of a cloud of points from depth sequences.

As for object recognition, Belongie et al. encode each point in $\mathcal{R}_t$ [30]. While we only encode points in $\mathcal{P}_t$ by exploring their relationships with points in $\mathcal{R}_t$. In this way, the importance of moving regions is emphasized for human activity representation. In Fig. 4, we take similar human activities "drawX" and "drawTick" from MSRAction3D dataset [28] as an example. These two activities own nearly the same human postures through whole sequences except for slight differences in the movements of one hand. In other words, many local points in $\mathcal{R}_t$ for "drawX" and "drawTick" share similar structural information, and only those points in $\mathcal{P}_t$ from the moving parts appear distinctive to represent these activities.

The pipeline of designing Depth Context descriptor for one point in $\mathcal{P}_t$ is illustrated in Fig. 5, where the human activity is denoted by a cloud of green points, i.e. $\mathcal{R}_t$, which is shown in Fig. 5(b). Note that the numbers located in the green points mean depth values. From the green points, we further select local points with salient motions and these points are in yellow color, i.e. $\mathcal{P}_t$, which is shown in Fig. 5(c). Here, yellow and green points are also called "target points" and "reference points". In Fig. 5(d), a target point is encoded by its relationships with reference points, where a Depth Context descriptor with 8∗3 bins is utilized to encode relative depth values between the target point and reference points. In Fig. 5(e), the target point is mathematically denoted as a vector with 8∗3∗2 dimensions, where positive and negative relative depth values are respectively recorded.

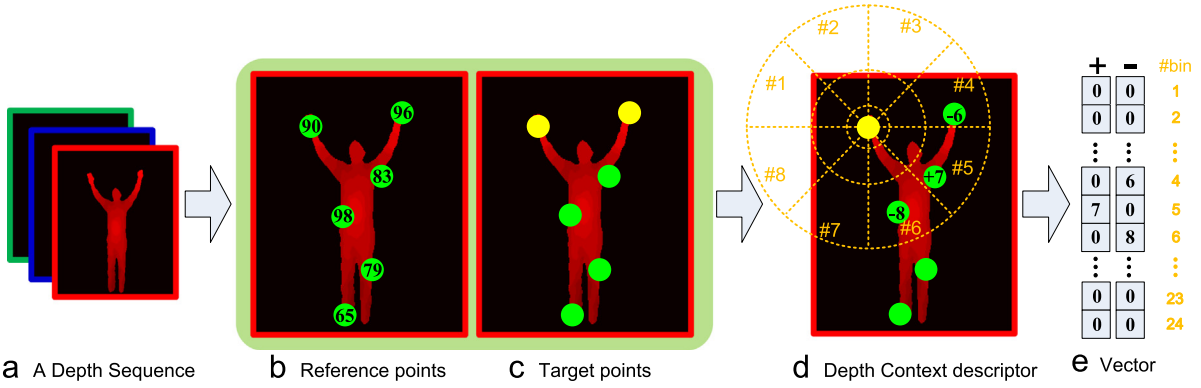

Fig. 5. The pipeline of our Depth Context descriptor. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

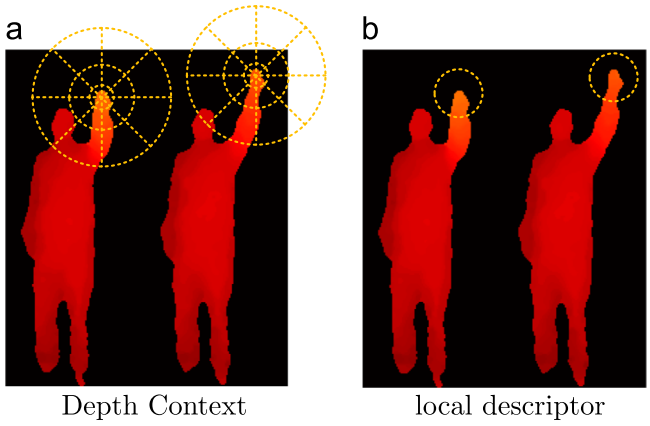

Fig. 6. Comparison of two local interest point descriptors.

Let $\mathcal{P}_t = \{\mathcal{P}_t^j\}_{j=1}^{N_{\mathcal{P}_t}}$ denote target points with a number of $N_{\mathcal{P}_t}$, and let $\mathcal{R}_t = \{\mathcal{R}_t^i\}_{i=1}^{N_{\mathcal{R}_t}}$ denote reference points with a number of $N_{\mathcal{R}_t}$. Note that each point refers to a vector which contains x and y coordinates. To describe a point $\mathcal{P}_t^j$, we firstly divide the space around this point into K bins which can be seen in Fig. 5(d). The bins are normally taken to be uniform in log-polar space. To be invariant to scale, all radial distances are normalized by the mean distance α between all the point pairs:

$$\alpha = \frac{\sum_{\forall i,j \in (1,\ldots,N_{\mathcal{R}_t})} \| \mathcal{R}_t^i - \mathcal{R}_t^j \|_2}{\sum_{\forall i,j \in (1,\ldots,N_{\mathcal{R}_t})} 1}, \tag{7}$$

where $\| \mathcal{R}_t^i - \mathcal{R}_t^j \|_2$ means the Euclidean distance between points $\mathcal{R}_t^i$ and $\mathcal{R}_t^j$. Different from [30] which counts the number of points for each bin, we accumulate the relative depth feature ζ between target point $\mathcal{P}_t^j$ and reference points which are located in corresponding bin. Taking one reference point $\mathcal{R}_t^i$ located in the kth bin as an example, feature ζ is defined as:

$$\zeta^k_{\mathcal{P}_t^j, \mathcal{R}_t^i} = \begin{cases} |I_t(\mathcal{P}_t^j) - I_t(\mathcal{R}_t^i)|, & \text{if } \dfrac{\mathcal{P}_t^j - \mathcal{R}_t^i}{\alpha} \in bin(k) \\ 0, & \text{otherwise} \end{cases}, \tag{8}$$

which records the depth distance between $\mathcal{R}_t^i$ and $\mathcal{P}_t^j$. Note that $I_t(\mathcal{P}_t^j)$ means the depth value in position $\mathcal{P}_t^j$ of frame I_t. We also define an indicator $\delta_{\mathcal{P}_t^j, \mathcal{R}_t^i}$ as:

$$\delta_{\mathcal{P}_t^j, \mathcal{R}_t^i} = \begin{cases} 1, & \text{if } I_t(\mathcal{P}_t^j) > I_t(\mathcal{R}_t^i) \\ 0, & \text{otherwise} \end{cases}, \tag{9}$$

which indicates whether the depth value of point $\mathcal{P}_t^j$ is bigger than $\mathcal{R}_t^i$ or not. For every point located in the kth bin, we calculate the value of δ and feature ζ. Then, positive and negative relative depth features in the kth bin are given as:

$$h^k_{\mathcal{P}_t^j,+} = \sum_{i=1}^{N_{\mathcal{R}_t}} \left(\zeta^k_{\mathcal{P}_t^j, \mathcal{R}_t^i} \cdot \delta_{\mathcal{P}_t^j, \mathcal{R}_t^i} \right), \tag{10}$$

$$h^k_{\mathcal{P}_t^j,-} = \sum_{i=1}^{N_{\mathcal{R}_t}} \left(\zeta^k_{\mathcal{P}_t^j, \mathcal{R}_t^i} \cdot (1 - \delta_{\mathcal{P}_t^j, \mathcal{R}_t^i}) \right). \tag{11}$$

By applying Formula (10), (11) to K bins, the feature vector for point $\mathcal{P}_t^j$ is defined as:

$$h_{\mathcal{P}_t^j} = \left[h^1_{\mathcal{P}_t^j,+}, \ldots, h^K_{\mathcal{P}_t^j,+}, h^1_{\mathcal{P}_t^j,-}, \ldots, h^K_{\mathcal{P}_t^j,-} \right], \tag{12}$$

which records positive and negative relative depth features in all bins. Further, $h_{\mathcal{P}_t^j}$ is normalized as:

$$\hat{h}_{\mathcal{P}_t^j} = \left[\frac{h^1_{\mathcal{P}_t^j,+}}{\| h_{\mathcal{P}_t^j} \|_1}, \ldots, \frac{h^K_{\mathcal{P}_t^j,+}}{\| h_{\mathcal{P}_t^j} \|_1}, \frac{h^1_{\mathcal{P}_t^j,-}}{\| h_{\mathcal{P}_t^j} \|_1}, \ldots, \frac{h^K_{\mathcal{P}_t^j,-}}{\| h_{\mathcal{P}_t^j} \|_1} \right] \tag{13}$$

where $\| \cdot \|_1$ calculates the l_1 norm of given variable.

The Depth Context descriptor $\hat{h}_{\mathcal{P}_t^j}$ can encode both local and global distributions of depth information around local point j, which is superior to local descriptors. As shown in Fig. 6(b), local descriptor only encodes the movement on the hand, which can barely reflect the activity of "raise up one hand". Instead, Depth Context descriptor in Fig. 6(a) not only involves local movements but also records the global relationships between the hand and the body.

5. Human activity recognition framework

We propose a framework for human activity recognition using Depth Context descriptor in this section. The pipeline is presented in Algorithm 1, where sequence $\mathcal{I}$ is processed frame by frame from line 1 to line 17. Local interest points are detected from line 2 to line 5 for the tth frame I_t, and all points from I_t are described from line 6 to line 17 by the Depth Context descriptor. After all frames have been processed, points from frame 2 to frame T are collected to form a point set h, which denotes each point by its Depth Context descriptor. Finally, h is used as the input for the BoVW model named $\mathcal{H}$ to form representation H, where C in line 19 means clusters for clustering method and it determines the length of H. Proper values of C are needed to achieve state-of-the-art results. In experiment section, we test the effect of C and choose default values for different datasets.

Algorithm 1: Extraction of human activity representation

Input: $\mathcal{I} = \{I_t\}_{t=1}^{T}, C$
Output: H

```
for t = 2; t ≤ T do
    G_t ← Formula 1; 𝒢_t ← Formula 2;
    M_t ← Formula 3; ℳ_t ← Formula 4;
    ℛ_t, N_ℛt ← Formula 5;
    𝒫_t, N_𝒫t ← Formula 6;
    for i = 1; i ≤ N_ℛt do
        for j = 1; j ≤ N_ℛt do
            α ← Formula 7;
    for j = 1; j ≤ N_𝒫t do
        for k = 1; k ≤ K do
            for i = 1; i ≤ N_ℛt do
                ζ^k_{𝒫_t^j, ℛ_t^i} ← Formula 8;
                δ_{𝒫_t^j, ℛ_t^i} ← Formula 9;
            h^k_{𝒫_t^j,+} ← Formula 10;
            h^k_{𝒫_t^j,−} ← Formula 11;
        h_{𝒫_t^j} ← Formula 12;
        ĥ_{𝒫_t^j} ← Formula 13;
h = {{ĥ_{𝒫_t^j}}_{j=1}^{N_ℛt}}_{t=2}^{T};
return H = ℋ(h, C);
```

The computational complexity for calculating h is $\sum_{t=2}^{T} O(K \cdot N_{\mathcal{P}_t} \cdot N_{\mathcal{R}_t})$. To reduce the computational complexity, we sample equal number of points denoted by L for each frame using Formula (5). Therefore $N_{\mathcal{P}_t}$ becomes a constant value which equals L. Since $\mathcal{R}_t$ is sampled from $\mathcal{P}_t$, the number of points in $\mathcal{R}_t$ denoted by $N_{\mathcal{R}_t}$ is not bigger than L. With above conditions, the computational complexity is simplified as $K \cdot L^2 \cdot O(T)$, which means the time cost of our framework grows linearly with the number of frames. In other words, our framework is efficient to implement which enables potential usage in real time applications.

6. Experiments

6.1. Datasets and settings

The MSRAction3D dataset stands out as one of the most widely used depth datasets in the literature [40], which is proposed in [28]. It contains 20 actions: "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 & throw". Each action is performed 2 or 3 times by 10 subjects facing the depth camera, and there is totally 567 depth sequences in the dataset.

The MSRGesture3D dataset is proposed in [21], which is a hand gesture dataset of depth sequences. It contains 12 gestures defined by American Sign Language: "bathroom", "blue", "finish", "green", "hungry", "milk", "past", "pig", "store", "where", "j" and "z". Each action is performed 2 or 3 times by each subject, resulting in 336 depth sequences.

The SKIG dataset is proposed in [33], which contains 1080 hand gesture depth sequences. It contains 10 gestures, including "circle", "triangle", "up-down", "right-left", "wave", "Z", "cross", "comehere", "turnaround" and "pat". All gestures are performed with hand postures (i.e., fist, flat and index) by 6 subjects under 2 illumination conditions (i.e., strong and poor light) and 3 backgrounds (i.e., white plain paper, wooden board and paper with characters).

Several action snaps from datasets above are shown in Figs. 7–9, where inter-class-similarities among different types of actions are observed. In MSRAction3D dataset, actions like "drawX" and "drawTick" are similar except for slight differences in the movements of one hand. For MSRGesture3D dataset, actions like "milk" and "hungry" are alike, since both actions need the motion of bending palm. What is more, self-occlusion is also a main issue for the MSRGesture3D dataset. The SKIG dataset is utilized to test the robustness of our method against pose, illumination and background. Note that snaps in Fig. 9 are obtained using foreground extraction [41], where only hand regions are extracted from original depth sequences.

We set the number of sampled points in each frame, which is denoted as L, to 100. As in Fig. 5(d), we respectively set 8 and 3 splits for theta and radius of spatial bins. The inner radius for the log-polar space is set to 0.3 and the outer radius is set to 5. Above settings are the same as shape context [30] except for the splits of theta. To describe the distribution of points more precisely, we change the splits of theta from original value 4 to 8. This change doubles the number of bins denoted by K, which increases from 12 to 24. Default values of C for MSRAction3D dataset, MSRGesture3D dataset and SKIG dataset are 3000, 1000 and 1000. Default values for g_1, g_2, g_3 are 0.1, 10 and 100.

Recognition is conducted using a non-linear SVM with a homogeneous Chi2 kernel [42], and parameter "gamma" which decides the homogeneity degree of the kernel is set to 0.8. We choose the "sdca" solver for SVM and use other default parameters provided in vlfeat library.[2] In order to keep the reported results consistent with

[2] Code in http://www.vlfeat.org/applications/caltech-101-code.html.

Fig. 7. Snaps in MSRGesture3D dataset.

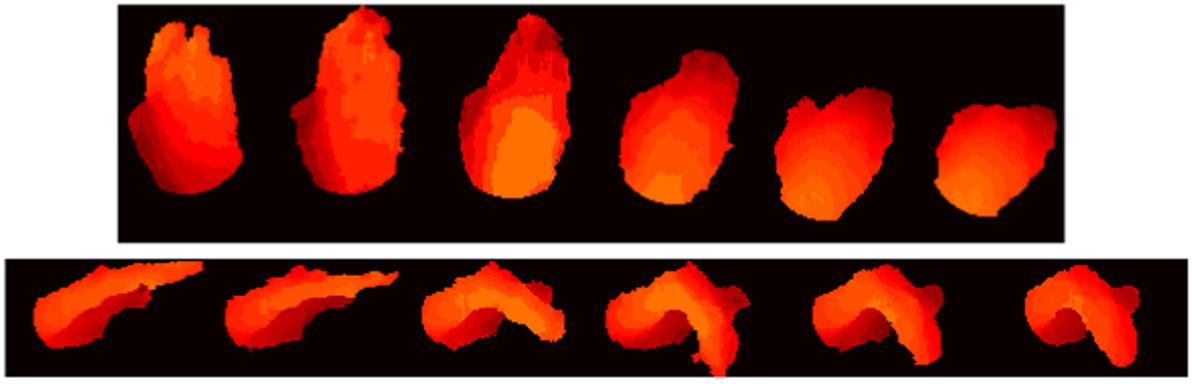

Fig. 8. "milk" and "hungry" in MSRGesture3D dataset.

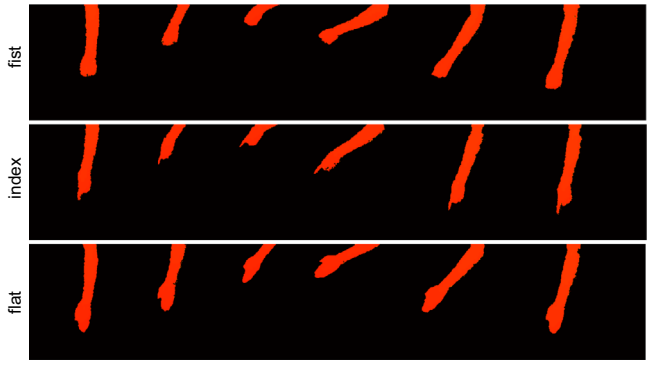

Fig. 9. Three poses for one type of activity named "circle" in SKIG dataset.

other works, we adopt the same cross-validation methods with [28,21,33]. It is noted that cross-subject validation is adopted for MSRAction3D dataset, with subjects # 1,3,5,7,9 for training and subjects # 2,4,6,8,10 for testing [28]. Since random initialization is involved in clustering method of BoVW model, all confusion matrices are average values over 10 times running results.

6.2. Framework evaluation

Previous works observe that the number of cluster, denoted as C, has an effect on the performance of BoVW model [43]. To achieve better performance, parameter C ranging from 1000 to 4000 at 1000 intervals are tested on MSRAction3D, MSRGesture3D and SKIG datasets. As shown in Fig. 10, our framework achieves more than 90% recognition precisions with different C, which shows the robustness of our method. When C respectively equals to 3000, 1000 and 1000, highest precisions are obtained for three datasets.

We evaluate Depth Context and our local point detector in Fig. 11, where "Depth Context" and "shape context" respectively mean applying Depth Context and shape context descriptor to encode all local points in $\mathcal{R}_t$. Since our detector selects local points with salient motions from $\mathcal{R}_t$ to form a new point set $\mathcal{P}_t$, "Depth

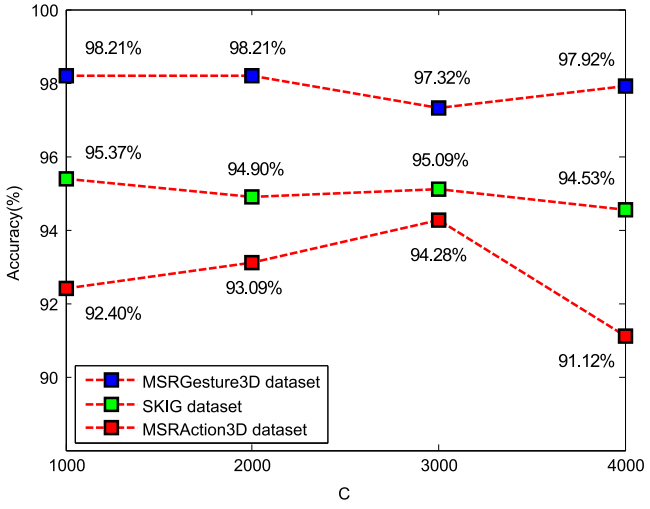

Fig. 10. Evaluation of the effect of parameter C.

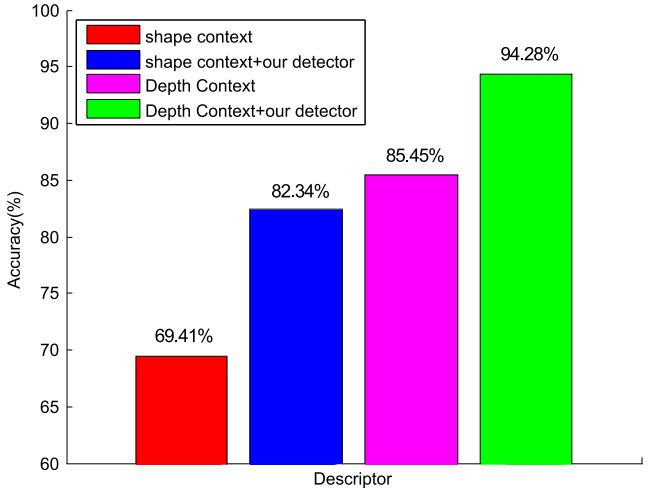

Fig. 11. Evaluation of Depth Context descriptor and our detector on MSRAction3D dataset.

Context+our detector" and "shape context+our detector" respectively mean applying Depth Context and shape context descriptor to encode all local points in $\mathcal{P}_t$.

Our detector improves the performance of "Depth Context" by 8.83% and improves "shape context" by 12.93%, which shows the significance of selecting motion related local points for classification. Since Depth Context not only encodes local point distribution but also involves depth information, "Depth Context" achieves 16.04% higher than "shape context" and "Depth Context+our detector" achieves 11.94% higher than "shape context+our detector". Moreover, "Depth Context+our detector" achieves the highest precision of 94.28%, which verifies the complementary property of motion and depth information.

6.3. Comparison with state-of-the-art methods

We compare our framework with state-of-the-art methods in Tables 2–4, where the original works which collect the datasets are in bold. The best performances of our framework are shown in Figs. 12–14. Compared with HON4D feature [26], 5.39% and 5.76% improvements are respectively achieved for MSRAction3D and MSRGesture3D datasets. Compared with Super Normal Vector feature [27], 1.19% and 3.47% improvements are also achieved. On SKIG dataset, we achieve precision of 95.37%, which is 6.67% higher than original work [33]. Our framework outperforms state-of-the-art works which lack either local or global

Table 2
Comparisons between our method and related works on MSRAction3D dataset.

Methods	Accuracy (%)	Details
Bag of 3D points [28]	74.70	**Li et al. (2010)**
Motion depth surface [14]	78.48	Azary et al. (2013)
STOP [29]	84.80	Vieira et al. (2012)
Random occupancy pattern [21]	86.50	Wang et al. (2012)
Actionlet ensemble [23]	88.20	Wang et al. (2014)
Depth motion maps [11]	88.73	Yang et al. (2012)
HON4D [26]	88.89	Oreifej et al. (2013)
DSTIP [38]	89.30	Xia et al. (2013)
Moving pose [17]	91.70	Zanfir et al. (2013)
Special Euclidean [44]	89.48	Vemulapalli et al. (2014)
Super normal vector [27]	93.09	Yang et al. (2014)
Our method	**94.28**	$C=3000$, protocol of [28]

Table 3
Comparisons between our method and related works on MSRGesture3D dataset.

Methods	Accuracy (%)	Details
Motion depth surface [14]	85.42	Azary et al. (2013)
Random occupancy pattern [21]	88.50	**Wang et al. (2012)**
HON4D [26]	92.45	Oreifej et al. (2013)
Super normal vector [27]	94.74	Yang et al. (2014)
Depth gradients+RDF [22]	95.29	Rahmani et al. (2014)
HOPC [45]	96.23	Hossein et al. (2014)
Our method	**98.21**	$C=1000$, protocol of [21]

Table 4
Comparisons between our method and related works on SKIG dataset.

Methods	Accuracy (%)	Details
DBN + depth [46]	73.80	Hinton et al. (2006)
HOG3D [35]	75.40	Klaser et al. (2008)
RGGP + depth [33]	76.10	**Liu et al. (2013)**
DBN + RGB-D [46]	85.90	Hinton et al. (2006)
RGGP + RGB-D [33]	88.70	**Liu et al. (2013)**
4DCOV [47]	93.80	Cirujeda et al. (2014)
Our method	**95.37**	$C=1000$, protocol of [33]

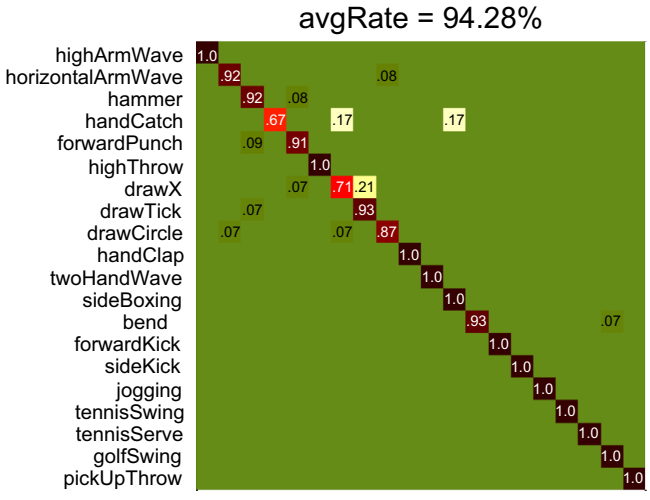

Fig. 12. Our best performance on MSRAction3D dataset.

information, since we not only explore local motion and depth clues for local interest points, but also encode the global constrains among these points.

6.4. Robustness evaluation

6.4.1. Partial occlusion

To evaluate the robustness of proposed framework to occlusion, we adopt the same settings in [21] and divide depth sequences from MSRAction3D dataset into two parts respectively in x, y and t dimensions. The whole sequences are divided into volumes, and eight kinds of occlusions are simulated by ignoring points fall into specified volumes. Fig. 15 illustrates two kinds of occlusions for a same depth sequence, where the shape of actor is dramatically changed and some salient motion is also hidden. The performance of our method is compared with Random Occupancy Pattern (ROP) feature [21] in Table 5. Obviously, our method achieves higher precisions than ROP with all kinds of occlusions. Note that sparse coding method can improve the robustness of given features to occlusions [21]. Without applying sparse coding, our method still outperforms "ROP+sparse coding" under most types of occlusions.

6.4.2. Pose, illumination and background variations

We test the robustness of our method against pose, illumination and background, which are short for "p", "i" and "b". SKIG dataset is utilized, which contains 3 poses (fist, index and flat), 2 illumination conditions (light and dark) and 3 backgrounds (wooden board, white paper and paper with characters). The robustness against pose is shown in Table 6, where "i" and "b" are set to constant values. When "$b=1$, $i=1$", all sequences with the first kind of background and the first kind of illumination are collected. Then these sequences are split into 3 groups for three-fold cross-validation, where each group has sequences of same posture.

In this way, two kinds of postures are utilized for training and a new one is for testing. As in Table 6, we gain higher than 95% precision under different conditions, which shows the robustness against pose changes. Similar experimental settings are applied for

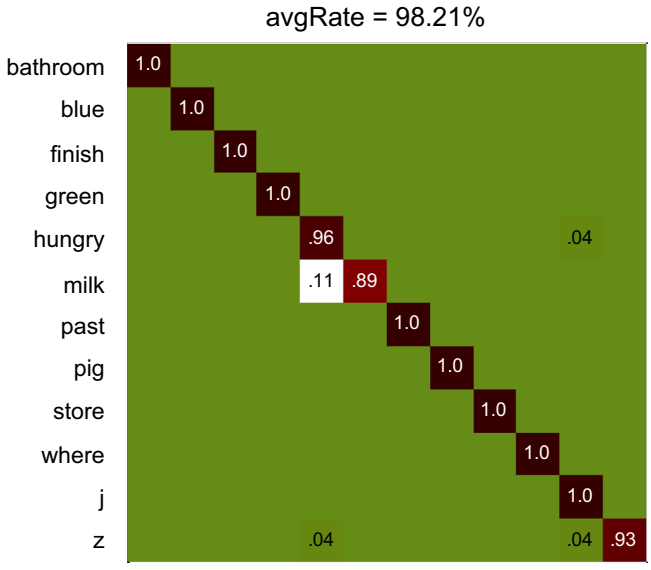

Fig. 13. Our best performance on MSRGesture3D dataset.

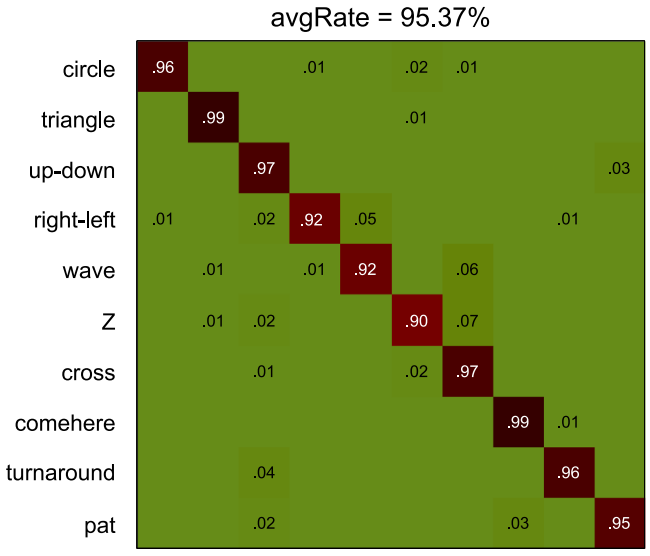

Fig. 14. Our best performance on SKIG dataset.

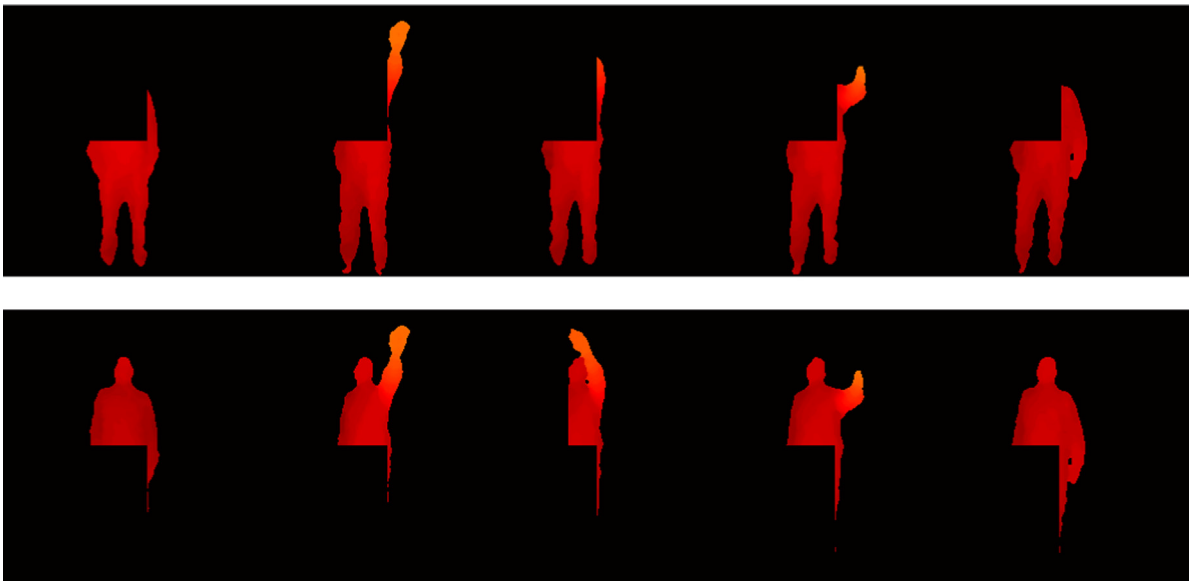

Fig. 15. Partial occlusions in MSRAction3D dataset.

Table 5
Evaluation the robustness of different methods to partial occlusions.

Occlusion	ROP [21] (%)	ROP+sparse coding [21] (%)	Our method (%)
1	83.05	86.17	**91.06**
2	84.18	86.50	**94.39**
3	78.76	80.09	**84.74**
4	82.12	**85.49**	85.13
5	84.48	87.51	**93.23**
6	82.46	87.51	**94.21**
7	80.10	83.80	**93.10**
8	85.83	86.83	**93.36**

Table 6
Evaluating the robustness of our method to pose changes.

	$b = 1$	$b = 2$	$b = 3$
$i = 1$	96.63%	96.67%	96.67%
$i = 2$	96.67%	96.67%	97.22%

Table 7
Evaluating the robustness of our method to illumination changes.

	$b = 1$	$b = 2$	$b = 3$
$p = 1$	96.67%	95.83%	97.50%
$p = 2$	96.67%	99.17%	96.67%
$p = 3$	95.83%	98.33%	95.83%

Table 8
Evaluating the robustness of our method to background changes.

	$p = 1$	$p = 2$	$p = 3$
$i = 1$	97.78%	98.89%	98.89%
$i = 2$	98.33%	99.44%	97.78%

testing the robustness of illumination and background variations. Results in Tables 7 and 8 also illustrate the robustness of our method.

7. Conclusions and future works

We propose a new Depth Context descriptor for human activity recognition using sole depth sequences. Specifically, a local interest points detector is introduced to denote local motion and shape clues. Then the Depth Context descriptor is designed to encode local points with local and global constrains. Finally, BoVW model is utilized to summarize local features and a non-linear SVM classifier is applied for classification. Our Depth Context is able to distinguish similar human activities, and shows robustness against partial occlusions, different poses, various illuminations and backgrounds. What is more, Depth Context can be calculated efficiently, showing potential usage in real time applications. Future works focus on combining extra information from local interest points for more challenging tasks, such as human activity detection using sole depth sequences and tackling with cluttered backgrounds, treating this proposed framework as a fundamental work.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC, No. 61340046), the National High Technology Research and Development Programme of China (863 Programme, No. 2006AA04Z247), the Scientific and Technical Innovation Commission of Shenzhen Municipality (No. JCYJ20120614152234873, JCYJ20130331144716089), and the Specialized Research Fund for the Doctoral Programme of Higher Education (SRFDP, No. 20130001110011).

References

[1] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: Proceedings of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), 2005, pp. 65–72.
[2] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.
[3] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Dense trajectories and motion boundary descriptors for action recognition, Int. J. Comput. Vis. 103 (1) (2013) 60–79.
[4] G. Lu, M. Kudo, Learning action patterns in difference images for efficient action recognition, Neurocomputing 123 (2014) 328–336.
[5] J. Giles, Inside the race to hack the kinect, New Sci. 208 (2789) (2010) 22–23.
[6] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, Real-time human pose recognition in parts from single depth images, ACM Commun. 56 (1) (2013) 116–124.
[7] B. Kwolek, M. Kepski, Improving fall detection by the use of depth sensor and accelerometer, Neurocomputing 168 (2015) 637–645.
[8] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, A. Fitzgibbon, Efficient regression of general-activity human poses from depth images, in: Proceedings of the IEEE Conference on Computer Vision (ICCV), 2011, pp. 415–422.
[9] Z. Zhang, Microsoft kinect sensor and its effect, IEEE MultiMed. 19 (2) (2012) 4–10.
[10] M. Harville, G. Gordon, J. Woodfill, Foreground segmentation using adaptive mixture models in color and depth, in: Proceedings of the IEEE Workshop on Detection and Recognition of Events in Video (EVENT), 2001, pp. 3–11.
[11] X. Yang, C. Zhang, Y. Tian, Recognizing actions using depth motion maps-based histograms of oriented gradients, in: Proceedings of the ACM International Conference on Multimedia (ACM MM), 2012, pp. 1057–1060.
[12] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 886–893.
[13] J.W. Davis, A.E. Bobick, The representation and recognition of human movement using temporal templates, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1997, pp. 928–934.
[14] S. Azary, A. Savakis, Grassmannian sparse representations and motion depth surfaces for 3D action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2013, pp. 492–499.
[15] L.W. Campbell, A.E. Bobick, Recognition of human body motion using phase space constraints, in: Proceedings of the IEEE Conference on Computer Vision (ICCV), 1995, pp. 624–630.
[16] X. Yang, Y. Tian, Eigenjoints-based action recognition using naive-bayes-nearest-neighbor, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012, pp. 14–19.
[17] M. Zanfir, M. Leordeanu, C. Sminchisescu, The moving pose: an efficient 3d kinematics descriptor for low-latency action recognition and detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2752–2759.
[18] C. Wang, Y. Wang, A.L. Yuille, An approach to pose-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 915–922.
[19] J. Luo, W. Wang, H. Qi, Group sparsity and geometry constrained dictionary learning for action recognition from depth maps, in: Proceedings of the IEEE Conference on Computer Vision (ICCV), 2013, pp. 1809–1816.
[20] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1290–1297.
[21] J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, Robust 3d action recognition with random occupancy patterns, in: Proceedings of the European Conference on Computer Vision (ECCV), 2012, pp. 872–885.
[22] H. Rahmani, A. Mahmood, D.Q. Huynh, A. Mian, Real time action recognition using histograms of depth gradients and random decision forests, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2014, pp. 626–633.

[23] J. Wang, Z. Liu, Y. Wu, J. Yuan, Learning actionlet ensemble for 3d human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 36 (5) (2014) 914–927.
[24] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2005, pp. 1395–1402.
[25] S. Tang, X. Wang, X. Lv, T. X. Han, J. Keller, Z. He, M. Skubic, S. Lao, Histogram of oriented normal vectors for object recognition with a depth sensor, in: Proceedings of the Asian Conference of Computer Vision (ACCV), 2013, pp. 525–538.
[26] O. Oreifej, Z. Liu, Hon4d: histogram of oriented 4d normals for activity recognition from depth sequences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 716–723.
[27] X. Yang, Y. Tian, Super normal vector for activity recognition using depth sequences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 804–811.
[28] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, pp. 9–14.
[29] A.W. Vieira, E.R. Nascimento, G.L. Oliveira, Z. Liu, M.F. Campos, Stop: space–time occupancy patterns for 3d action recognition from depth map sequences, in: Proceedings of the Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP), 2012, pp. 252–259.
[30] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (4) (2002) 509–522.
[31] M. Grundmann, F. Meier, I. Essa, 3D shape context and distance transform for action recognition, in: Proceedings of the International Conference on Pattern Recognition (ICPR), 2008, pp. 1–4.
[32] D. Zhao, L. Shao, X. Zhen, Y. Liu, Combining appearance and structural features for human action recognition, Neurocomputing 113 (2013) 88–96.
[33] L. Liu, L. Shao, Learning discriminative representations from RGB-D video data, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2013, pp. 1493–1500.
[34] C. Schüldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proceedings of the International Conference on Pattern Recognition (ICPR), 2004, pp. 32–36.
[35] A. Klaser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: Proceedings of the British Machine Vision Conference (BMVC), 2008, pp. 1–10.
[36] I. Laptev, On space–time interest points, Int. J. Comput. Vis. 64 (2–3) (2005) 107–123.
[37] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: Proceedings of the European Conference on Computer Vision (ECCV), 2008, pp. 650–663.
[38] L. Xia, J. Aggarwal, Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2834–2841.
[39] L. Vincent, Morphological grayscale reconstruction in image analysis: applications and efficient algorithms, IEEE Trans. Image Process. 2 (2) (1993) 176–201.
[40] J.R. Padilla-López, A.A. Chaaraoui, F. Flórez-Revuelta, A discussion on the validation tests employed to compare human action recognition methods using the MSR Action3D dataset, arXiv preprint arXiv:1407.7390, 2014.
[41] C. Zhang, X. Yang, Y. Tian, Histogram of 3D facets: A characteristic descriptor for hand gesture recognition, in: Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013, pp. 1–8.
[42] A. Vedaldi, A. Zisserman, Efficient additive kernels via explicit feature maps, IEEE Trans. Pattern Anal. Mach. Intell. 34 (3) (2012) 480–492.
[43] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: Proceedings of the British Machine Vision Conference (BMVC), 2009, pp. 1–11.
[44] R. Vemulapalli, F. Arrate, R. Chellappa, 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 (CVPR), 2014, pp. 588–595.
[45] H. Rahmani, A. Mahmood, D.Q. Huynh, A. Mian, HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), 2014, pp. 742–757.
[46] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[47] P. Cirujeda, X. Binefa, 4DCov: a nested covariance descriptor of spatio-temporal features for gesture recognition in depth sequences, in: Proceedings of the International Conference on 3D Vision (3DV), 2014, pp. 657–664.

Mengyuan Liu received the B.E. degree in intelligence science and technology in 2012, and is working toward the Doctor degree in the School of EE&CS, Peking University (PKU), China. His research interests include Action Recognition and Localization. He has published articles in IEEE International Conference Robotics and Biomimetics (ROBIO), IEEE International Conference on Image Processing (ICIP), IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Hong Liu received the Ph.D. degree in mechanical electronics and automation in 1996, and serves as a Full Professor in the School of EE&CS, Peking University (PKU), China. Prof. Liu has been selected as Chinese Innovation Leading Talent supported by National High-level Talents Special Support Plan since 2013. He is also the Director of Open Lab on Human Robot Interaction, PKU, his research fields include computer vision and robotics, image processing, and pattern recognition. Dr. Liu has published more than 150 papers and gained Chinese National Aero-space Award, Wu Wenjun Award on Artificial Intelligence, Excellence Teaching Award, and Candidates of Top Ten Outstanding Professors in PKU. He is an IEEE member, vice president of Chinese Association for Artificial Intelligent (CAAI), and vice chair of Intelligent Robotics Society of CAAI. He has served as keynote speakers, co-chairs, session chairs, or PC members of many important international conferences, such as IEEE/RSJ IROS, IEEE ROBIO, IEEE SMC and IIHMSP, recently also serves as reviewers for many international journals such as Pattern Recognition, IEEE Transactions on Signal Processing, and IEEE Transactions on PAMI.